Dear Reader,

In the following pages, we present the proof of a number of equations in more detail compared to the book. Some of these equations were pointed out to us by readers as well as tutors. Lengthy and technical steps toward the proof of an equation cannot be included in the book, due to length limitations as well as to pedagogical reasons, since they destruct the attention from the main points.

Mathematics should be kept to a minimum and used only to the extend that contribute to the better understanding of the method. Of course, what a minimum is in such cases is not very clear and it is subjective to a large extend. On the other hand, there are readers who would like to be exposed to more technical details. This file attempts to address the needs of such readers.

In case you feel that there are more equations that need further elaboration, we are happy to look at them and include them in a future update of this file. Please do contact us via email.

Sergios Theodoridis
Kostas Koutroumbas

Derivation of equation (2.34)

The densities $p(\boldsymbol{x}|\omega_i)$ are multivariate normal, hence

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{l/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right).$$

If we replace the above expression in equation 2.33 , we have

$$\begin{aligned} g_i(\boldsymbol{x}) &= \ln p(\boldsymbol{x}|\omega_i) + \ln P(\omega_i) \\ &= \ln\left[\frac{1}{(2\pi)^{l/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)\right] + \ln P(\omega_i). \end{aligned}$$

Using the properties of logarithms, we get

$$\begin{aligned} g_i(\boldsymbol{x}) &= \ln\left[\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right)\right] + \ln[(2\pi)^{-l/2}] + \ln(|\boldsymbol{\Sigma}_i|^{-1/2}) + \ln P(\omega_i) \\ &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + (-l/2)\ln(2\pi) + (-1/2)\ln(|\boldsymbol{\Sigma}_i|) + \ln P(\omega_i). \end{aligned}$$

Derivation of equation (2.43)

The discriminant function is

$$\begin{aligned} g_i(\boldsymbol{x}) &= \frac{1}{\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{x} + w_{i0} \\ w_{i0} &= \ln P(\omega_i) - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i = \ln P(\omega_i) - \frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i. \end{aligned}$$

The hyperplane between two classes is defined by

$$\begin{aligned} & g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) = 0 \\ \Leftrightarrow & \frac{1}{\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{x} + \ln P(\omega_i) - \frac{1}{2\sigma^2}\boldsymbol{\mu}_i^T\boldsymbol{\mu}_i - \frac{1}{\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{x} - \ln P(\omega_j) + \frac{1}{2\sigma^2}\boldsymbol{\mu}_j^T\boldsymbol{\mu}_j = 0 \\ \Leftrightarrow & \frac{1}{\sigma^2}\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T\boldsymbol{x} - \frac{1}{2\sigma^2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) + \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) = 0 \\ \Leftrightarrow & \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) + \sigma^2\ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) = 0 \\ \Leftrightarrow & \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T\left(\boldsymbol{x} - \left[\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \sigma^2\ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\right]\right) = 0 \\ \Leftrightarrow & \left(\underbrace{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}_{\mathbf{w}}\right)^T\left(\boldsymbol{x} - \left[\underbrace{\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \sigma^2\ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}}_{\boldsymbol{x}_0}\right]\right) = 0 \end{aligned}$$

Derivation of equation (2.46)

The discriminant function is

$$\begin{aligned} g_i(\boldsymbol{x}) &= \boldsymbol{\mu}_i^T \boldsymbol{\Sigma^{-1}} \boldsymbol{x} + w_{i0} \\ w_{i0} &= \ln P(\omega_i) - \frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i. \end{aligned}$$

The hyperplane between two classes is defined by

$$\begin{aligned} & g_i(\boldsymbol{x}) - g_j(\boldsymbol{x}) = 0 \\ \Leftrightarrow & \boldsymbol{\mu}_i^T \boldsymbol{\Sigma^{-1}} \boldsymbol{x} + \ln P(\omega_i) - \frac{1}{2}\boldsymbol{\mu}_i^T \boldsymbol{\Sigma^{-1}} \boldsymbol{\mu}_i - \boldsymbol{\mu}_j^T \boldsymbol{\Sigma^{-1}} \boldsymbol{x} - \ln P(\omega_j) + \frac{1}{2}\boldsymbol{\mu}_j^T \boldsymbol{\Sigma^{-1}} \boldsymbol{\mu}_j = 0 \\ \Leftrightarrow & \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) + \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) = 0 \\ \Leftrightarrow & \left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x} - \left[\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma^{-1}} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}\right]\right) = 0 \\ \Leftrightarrow & \underbrace{\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\right)^T \boldsymbol{\Sigma}^{-1}}_{\mathbf{w}} \left(\boldsymbol{x} - \left[\underbrace{\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2_{\boldsymbol{\Sigma^{-1}}}}}_{\boldsymbol{x}_0}\right]\right) = 0 \end{aligned}$$

Derivation of equation (3.28)

The MSE between the desired and true output is

$$J(\boldsymbol{w}) = E[|y - \boldsymbol{x}^T\boldsymbol{w}|^2] = \int (y - \boldsymbol{x}^T\boldsymbol{w})^2 p(\boldsymbol{x}) d\boldsymbol{x}$$

where $p(\boldsymbol{x})$ is the pdf of $\boldsymbol{x}$ for which we have

$$p(\boldsymbol{x}) = \sum_{i=1}^{2} p(\boldsymbol{x}|\omega_i)P(\omega_i)$$

hence

$$\begin{aligned} J(\boldsymbol{w}) &= \int (y - \boldsymbol{x}^T\boldsymbol{w})^2 p(\boldsymbol{x}|\omega_1)P(\omega_1)d\boldsymbol{x} + \int (y - \boldsymbol{x}^T\boldsymbol{w})^2 p(\boldsymbol{x}|\omega_2)P(\omega_2)d\boldsymbol{x} \\ &= P(\omega_1)\int (y - \boldsymbol{x}^T\boldsymbol{w})^2 p(\boldsymbol{x}|\omega_1)d\boldsymbol{x} + P(\omega_2)\int (y - \boldsymbol{x}^T\boldsymbol{w})^2 p(\boldsymbol{x}|\omega_2)d\boldsymbol{x}. \end{aligned}$$

However for class ω_1, $y = 1$ and for class ω_2, $y = -1$. Hence,

$$J(\boldsymbol{w}) = P(\omega_1)\int (1 - \boldsymbol{x}^T\boldsymbol{w})^2 p(\boldsymbol{x}|\omega_1)d\boldsymbol{x} + P(\omega_2)\int (1 + \boldsymbol{x}^T\boldsymbol{w})^2 p(\boldsymbol{x}|\omega_2)d\boldsymbol{x}.$$

Derivation of equations (3.65)-(3.66)

The logarithm of the likelihood ratios is modeled via the linear equations

$$\ln\frac{P(\omega_i)}{P(\omega_M)} = w_{i0} + \boldsymbol{w}_i^T\boldsymbol{x}, \quad i = 1, 2, \dots, M-1$$

The natural logarithm function is defined as the inverse function of the exponential function $\exp[\ln(x)] = x$, leading to

$$\frac{P(\omega_i)}{P(\omega_M)} = \exp\left(w_{i0} + \boldsymbol{w}_i^T\boldsymbol{x}\right).$$

Note that the unknown parameters $w_{i0}, \boldsymbol{w}_i$, must be chosen to ensure that the probabilities add to one, thus

$$\begin{aligned} \frac{\sum_{i=1}^{M-1} P(\omega_i)}{P(\omega_M)} &= \sum_{i=1}^{M-1} \exp\left(w_{i0} + \boldsymbol{w}_i^T\boldsymbol{x}\right) \\ \Leftrightarrow \frac{1 - P(\omega_M)}{P(\omega_M)} &= \sum_{i=1}^{M-1} \exp\left(w_{i0} + \boldsymbol{w}_i^T\boldsymbol{x}\right) \\ \Leftrightarrow P(\omega_M) &= \frac{1}{1 + \sum_{i=1}^{M-1} \exp\left(w_{i0} + \boldsymbol{w}_i^T\boldsymbol{x}\right)}. \end{aligned}$$

Equation 3.66 is derived from 3.65, as follows

$$\begin{aligned}
\frac{P(\omega_i)}{P(\omega_M)} &= \exp(w_{i0} + \boldsymbol{w}_i^T \boldsymbol{x}) \\
\Leftrightarrow P(\omega_i) &= \exp(w_{i0} + \boldsymbol{w}_i^T \boldsymbol{x}) P(\omega_M) \\
\Leftrightarrow P(\omega_i) &= \frac{\exp(w_{i0} + \boldsymbol{w}_i^T \boldsymbol{x})}{1 + \sum_{i=1}^{M-1} \exp\left(w_{i0} + \boldsymbol{w}_i^T \boldsymbol{x}\right)}.
\end{aligned}$$

Derivation of equation (4.74)

$$\begin{aligned}
&\|\phi(\boldsymbol{x}) - \boldsymbol{\mu}_1\|^2 > \|\phi(\boldsymbol{x}) - \boldsymbol{\mu}_1\|^2 \\
\Leftrightarrow &\|\phi(\boldsymbol{x})\|^2 - 2\langle \phi(\boldsymbol{x}), \boldsymbol{\mu}_1 \rangle + \|\boldsymbol{\mu}_1\|^2 > \|\phi(\boldsymbol{x})\|^2 - 2\langle \phi(\boldsymbol{x}), \boldsymbol{\mu}_2 \rangle + \|\boldsymbol{\mu}_2\|^2 \\
\Leftrightarrow &2\langle \phi(\boldsymbol{x}), \boldsymbol{\mu}_2 \rangle - 2\langle \phi(\boldsymbol{x}), \boldsymbol{\mu}_1 \rangle > (\|\boldsymbol{\mu}_2\|^2 - \|\boldsymbol{\mu}_1\|^2) \\
\Leftrightarrow &\langle \phi(\boldsymbol{x}), (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \rangle > 1/2(\|\boldsymbol{\mu}_2\|^2 - \|\boldsymbol{\mu}_1\|^2)
\end{aligned}$$

Derivation of equation (5.22)

We know that $\int p(\boldsymbol{x})d\boldsymbol{x} = 1, \int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x} = \boldsymbol{\mu}$ and $E\{\boldsymbol{x}\boldsymbol{x}^T\} = \boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}$. Hence,

$$\begin{aligned}
D_{ij} &= \int p(\boldsymbol{x}|\omega_i) \ln \frac{p(\boldsymbol{x}|\omega_i)}{p(\boldsymbol{x}|\omega_j)} d\boldsymbol{x} \\
&= \int p(\boldsymbol{x}|\omega_i)\Big\{ \ln p(\boldsymbol{x}|\omega_i) - \ln p(\boldsymbol{x}|\omega_j)\Big\} d\boldsymbol{x} \\
&= -\left(- \int p(\boldsymbol{x}|\omega_i) \ln p(\boldsymbol{x}|\omega_i) d\boldsymbol{x}\right) - \int p(\boldsymbol{x}|\omega_i) \ln p(\boldsymbol{x}|\omega_j) d\boldsymbol{x} \\
&= -H(p(\boldsymbol{x}|\omega_i)) - \int p(\boldsymbol{x}|\omega_i) \ln p(\boldsymbol{x}|\omega_j) d\boldsymbol{x}
\end{aligned}$$

The entropy of the multivariate Gaussian distribution is

$$\begin{aligned}
H(p(\boldsymbol{x}|\omega_i)) &= -\int p(\boldsymbol{x}|\omega_i) \ln p(\boldsymbol{x}|\omega_i) d\boldsymbol{x} = -E\left[\ln p(\boldsymbol{x}|\omega_i\right] \\
&= \frac{l}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + E\left[\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right] \\
&= \frac{l}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \frac{1}{2}E[\boldsymbol{x}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}] + \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \ldots \\
&- \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i
\end{aligned}$$

However,

$$\begin{aligned}
\frac{1}{2}E[\boldsymbol{x}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}] &= \frac{1}{2}E[\text{Trace}\{\boldsymbol{x}^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}\}] \\
&= \frac{1}{2}E[\text{Trace}\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}\boldsymbol{x}^T\}] \\
&= \frac{1}{2}\text{Trace}\left\{E[\boldsymbol{\Sigma}_i^{-1}\boldsymbol{x}\boldsymbol{x}^T]\right\} \\
&= \frac{1}{2}\text{Trace}\left\{\boldsymbol{\Sigma}_i^{-1}[\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T]\right\} \\
&= \frac{1}{2}\text{Trace}\boldsymbol{I} + \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i
\end{aligned}$$

Hence

$$H(p(\boldsymbol{x}|\omega_i)) = \frac{l}{2}\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \frac{1}{2}\text{Trace}\boldsymbol{I}.$$

$$
\begin{aligned}
D_{ij} &= -\frac{1}{2}\ln\left\{(2\pi)^l|\boldsymbol{\Sigma}_i|\right\} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I} - \int p(\boldsymbol{x}|\omega_i)\ln p(\boldsymbol{x}|\omega_j)d\boldsymbol{x}\\
&= -\frac{1}{2}\ln\left\{(2\pi)^l|\boldsymbol{\Sigma}_i|\right\} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I} - E_{\omega_i}\Big[\ln p(\boldsymbol{x}|\omega_j)\Big]\\
&= -\frac{1}{2}\ln\left\{(2\pi)^l|\boldsymbol{\Sigma}_i|\right\} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I}\ldots\\
&- E_{\omega_i}\left[\ln\left\{\frac{1}{(2\pi)^{l/2}|\boldsymbol{\Sigma}_j|^{1/2}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right)\right\}\right]\\
&= -\frac{1}{2}\ln\left\{(2\pi)^l|\boldsymbol{\Sigma}_i|\right\} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I} + \frac{1}{2}\ln\left\{(2\pi)^l\right\} - \frac{1}{2}\ln\left\{|\boldsymbol{\Sigma}_j|^{-1}\right\}\ldots\\
&- E_{\omega_i}\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right]\\
&= -\frac{l}{2}\ln\{(2\pi)\} - \frac{1}{2}\ln\{|\boldsymbol{\Sigma}_i|\} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I} + \frac{l}{2}\ln\{(2\pi)\} + \frac{1}{2}\ln\{|\boldsymbol{\Sigma}_j|\}\ldots\\
&+ \frac{1}{2}E_{\omega_i}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{x}\right] + \frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_i\\
&= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I}\ldots\\
&+ \frac{1}{2}E_{\omega_i}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{x}\right] + \frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \frac{1}{2}\boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_i
\end{aligned}
$$

However

$$
\begin{aligned}
\frac{1}{2}E_{\omega_i}\left[\boldsymbol{x}^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{x}\right] = \frac{1}{2}E_{\omega_i}\left[\mathrm{Trace}\{\boldsymbol{x}^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{x}\}\right] &= \frac{1}{2}\mathrm{Trace}\left\{E_{\omega_i}[\boldsymbol{\Sigma}_j^{-1}\boldsymbol{x}\boldsymbol{x}^T]\right\}\\
&= \frac{1}{2}\mathrm{Trace}\left\{\boldsymbol{\Sigma}_j^{-1}\left[\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^T\right]\right\}\\
&= \frac{1}{2}\mathrm{Trace}\left\{\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\right\} + \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_i
\end{aligned}
$$

Thus

$$
\begin{aligned}
&D_{ij} = \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I} + ...\\
&\frac{1}{2}\left(\mathrm{Trace}\{\boldsymbol{\Sigma}_{\boldsymbol{j}}^{\mathbf{-1}}\boldsymbol{\Sigma}_{i}\} + \boldsymbol{\mu}_i\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_i^T - \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_i + \boldsymbol{\mu}_j^T\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\right)\\
&= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I} + \frac{1}{2}\mathrm{Trace}\{\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\} + \frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)
\end{aligned}
$$

Similarly, D_{ji} is derived

$$
D_{ji} = \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_i|}{|\boldsymbol{\Sigma}_j|} - \frac{1}{2}\mathrm{Trace}\boldsymbol{I} + \frac{1}{2}\mathrm{Trace}\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j\} + \frac{1}{2}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)
$$

The divergence d_{ij} is

$$\begin{aligned}
d_{ij} &= D_{ij} + D_{ji} \\
&= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|} + \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_i|}{|\boldsymbol{\Sigma}_j|} + \frac{1}{2}\mathrm{Trace}\{\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\} + \frac{1}{2}\mathrm{Trace}\{\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j\} - \frac{1}{2}\mathrm{Trace}2\boldsymbol{I} + ... \\
&\frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\
&= \frac{1}{2}\mathrm{Trace}\{\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j - 2\boldsymbol{I}\} + \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\left(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)
\end{aligned}$$

Derivation of equations (10.18)-(10.20)

$$
\begin{aligned}
Q(\boldsymbol{\Theta};\boldsymbol{\Theta}(t)) &= \sum_{i=1}^{N_u}\sum_{y=1}^{M} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t)) \ln\left(p(\boldsymbol{x}_i|y;\boldsymbol{\mu}_y,\sigma_y^2)P_y\right) \\
&+ \sum_{y=1}^{M}\sum_{i=1}^{N_y} \ln\left(p(\boldsymbol{z}_{iy}|y;\boldsymbol{\mu}_y,\sigma_y^2)P_y\right) \\
&= \sum_{i=1}^{N_u}\sum_{y=1}^{M} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t)) \ln p(\boldsymbol{x}_i|y;\boldsymbol{\mu}_y,\sigma_y^2) + \sum_{i=1}^{N_u}\sum_{y=1}^{M} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t)) \ln P_y \\
&+ \sum_{y=1}^{M}\sum_{i=1}^{N_y} \ln p(\boldsymbol{z}_{iy}|y;\boldsymbol{\mu}_y,\sigma_y^2) + \sum_{y=1}^{M}\sum_{i=1}^{N_y} \ln P_y
\end{aligned}
$$

We first maximize the previous equation subject to $\sum_{y=1}^{M} P_y = 1$, hence the Lagrangian becomes

$$
L(\boldsymbol{\Theta},\lambda) = Q(\boldsymbol{\Theta};\boldsymbol{\Theta}(t)) - \lambda(\sum_{y=1}^{M} P_y - 1)
$$

where λ is the Lagrange multiplier due to equality constraint. The partial derivative with respect to P_y, for a specific y each time, is equal to

$$
\begin{aligned}
\frac{\partial L}{\partial P_y} =& \frac{1}{P_y}N_y + \frac{1}{P_y}\sum_{i=1}^{N_u}\sum_{y=1}^{M} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t)) - \lambda = 0 \\
&\sum_{i=1}^{N_u}\sum_{y=1}^{M} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t)) + N_y = \lambda P_y \\
&P_y = \frac{1}{\lambda}\left(\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t)) + N_y\right)
\end{aligned}
$$

The value of λ is recovered by summing the constraint over y

$$
\sum_{y=1}^{M} P_y = 1 = \frac{1}{\lambda}\left(N_u + N_l\right) \rightarrow \lambda = \frac{1}{N_u + N_l}
$$

Hence,

$$
P_y = \frac{1}{N_u + N_l}\left(\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t)) + N_y\right).
$$

Maximization with respect to $(\boldsymbol{\mu}_y, \sigma_y^2)$ depends on the specific form of the involved pdfs. For the case of Gaussian mixtures with diagonal covariance matrices $(\boldsymbol{\Sigma}_y = \sigma_y^2 \boldsymbol{I})$

$$p(\boldsymbol{x}_i|y; \boldsymbol{\mu}_y, \sigma_y^2) = \frac{1}{(2\pi\sigma_y^2)^{l/2}} \exp\left(-\frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_y)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_y)}{2\sigma_y^2}\right)$$

or

$$\ln p(\boldsymbol{x}_i|y; \boldsymbol{\mu}_y, \sigma_y^2) = -\frac{l}{2}\ln(2\pi) - l\ln\sigma_y - \frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_y)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_y)}{2\sigma_y^2}$$

Maximization with respect to $\boldsymbol{\mu}_y$ is achieved by requiring the derivative with respect to $\boldsymbol{\mu}_y$ to be equal to zero:

$$\frac{\partial}{\partial \boldsymbol{\mu}_y}\left[\sum_{i=1}^{N_u}\sum_{y=1}^{M} P(y|\boldsymbol{x}_i; \boldsymbol{\Theta}(t)) \ln p(\boldsymbol{x}_i|y; \boldsymbol{\mu}_y, \sigma_y^2) + \sum_{y=1}^{M}\sum_{i=1}^{N_y} \ln p(\boldsymbol{z}_{iy}|y; \boldsymbol{\mu}_y, \sigma_y^2)\right] = 0$$

$$\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i; \boldsymbol{\Theta}(t)) \frac{\partial}{\partial \boldsymbol{\mu}_y}\left[-\frac{l}{2}\ln(2\pi) - l\ln\sigma_y - \frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_y)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_y)}{2\sigma_y^2}\right]$$

$$+\sum_{i=1}^{N_y} \frac{\partial}{\partial \boldsymbol{\mu}_y}\left[-\frac{l}{2}\ln(2\pi) - l\ln\sigma_y - \frac{(\boldsymbol{z}_{iy} - \boldsymbol{\mu}_y)^T(\boldsymbol{z}_{iy} - \boldsymbol{\mu}_y)}{2\sigma_y^2}\right] = 0$$

$$\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i; \boldsymbol{\Theta}(t)) \frac{(\boldsymbol{\mu}_y - \boldsymbol{x}_i)}{\sigma_y^2} + \sum_{i=1}^{N_y} \frac{(\boldsymbol{\mu}_y - \boldsymbol{z}_{iy})}{\sigma_y^2} = 0$$

$$\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i; \boldsymbol{\Theta}(t))\boldsymbol{x}_i + \sum_{i=1}^{N_y} \boldsymbol{z}_{iy} = \boldsymbol{\mu}_y\left(N_y + \sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i; \boldsymbol{\Theta}(t))\right)$$

$$\boldsymbol{\mu}_y = \frac{\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i; \boldsymbol{\Theta}(t))\boldsymbol{x}_i + \sum_{i=1}^{N_y} \boldsymbol{z}_{iy}}{\left(N_y + \sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i; \boldsymbol{\Theta}(t))\right)}$$

Maximization with respect to σ_y is achieved by taking the derivative with respect to σ_y equal to zero:

$$\frac{\partial}{\partial \sigma_y}\left[\sum_{i=1}^{N_u}\sum_{y=1}^{M} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\ln p(\boldsymbol{x}_i|y;\boldsymbol{\mu}_y,\sigma_y^2)+\sum_{y=1}^{M}\sum_{i=1}^{N_y}\ln p(\boldsymbol{z}_{iy}|y;\boldsymbol{\mu}_y,\sigma_y^2)\right]=0$$

$$\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\frac{\partial}{\partial \sigma_y}\left[-\frac{l}{2}\ln(2\pi)-l\ln\sigma_y-\frac{(\boldsymbol{x}_i-\boldsymbol{\mu}_y)^T(\boldsymbol{x}_i-\boldsymbol{\mu}_y)}{2\sigma_y^2}\right]$$

$$+\sum_{i=1}^{N_y}\frac{\partial}{\partial \sigma_y}\left[-\frac{l}{2}\ln(2\pi)-l\ln\sigma_y-\frac{(\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y)^T(\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y)}{2\sigma_y^2}\right]=0$$

$$\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\left[-\frac{l}{\sigma_y}+\frac{(\boldsymbol{x}_i-\boldsymbol{\mu}_y)^T(\boldsymbol{x}_i-\boldsymbol{\mu}_y)}{\sigma_y^3}\right]$$

$$+\sum_{i=1}^{N_y}\left[-\frac{l}{\sigma_y}+\frac{(\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y)^T(\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y)}{\sigma_y^3}\right]=0$$

$$\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\frac{(\boldsymbol{x}_i-\boldsymbol{\mu}_y)^T(\boldsymbol{x}_i-\boldsymbol{\mu}_y)}{\sigma_y^3}+\sum_{i=1}^{N_y}\frac{(\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y)^T(\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y)}{\sigma_y^3}$$

$$=\frac{l}{\sigma_y}\left(N_y+\sum_{i=1}^{N_u}P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\right)$$

$$\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\|\boldsymbol{x}_i-\boldsymbol{\mu}_y\|^2+\sum_{i=1}^{N_y}\|\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y\|^2=l\sigma_y^2\left(N_y+\sum_{i=1}^{N_u}P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\right)$$

$$\sigma_y^2=\frac{\sum_{i=1}^{N_u} P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\|\boldsymbol{x}_i-\boldsymbol{\mu}_y\|^2+\sum_{i=1}^{N_y}\|\boldsymbol{z}_{iy}-\boldsymbol{\mu}_y\|^2}{l\left(N_y+\sum_{i=1}^{N_u}P(y|\boldsymbol{x}_i;\boldsymbol{\Theta}(t))\right)}$$