Final solutions
ECE 271
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos Fall 2016

1. a) The log-likelihood is

$$l(\theta) = \sum_i \log h(x_i) + \nu(\theta) \sum_i T(x_i) - nA(\theta).$$

Setting its derivative to zero, we have

$$\frac{\partial \nu}{\partial \theta} \sum_i T(x_i) = n \frac{\partial A}{\partial \theta},$$

and the equation follows. For this solution to be a maximum we need that

$$\frac{\partial^2 l}{\partial \theta^2} = \frac{\partial^2 \nu}{\partial \theta^2} \sum_i T(x_i) - n \frac{\partial^2 A}{\partial \theta^2} < 0$$

i.e. the additional constraint that

$$\frac{\partial^2 A}{\partial \theta^2} > \frac{\partial^2 \nu}{\partial \theta^2} \frac{1}{n} \sum_i T(x_i)$$

b) According to bayes decision rule, the optimal decision function, under the "0/1" loss is "pick $i = 0$" if

$$\begin{aligned} P_{Y|X}(0|\mathcal{D}) &> P_{Y|X}(1|\mathcal{D}) \\ P_{X|Y}(\mathcal{D}|0)P_Y(0) &> P_{X|Y}(\mathcal{D}|1)P_Y(1) \\ P_{X|Y}(x_1, \ldots, x_n|0)P_Y(0) &> P_{X|Y}(x_1, \ldots, x_n|1)P_Y(1) \end{aligned}$$

Taking the log on both sides, and using the independence assumption,

$$\begin{aligned} \log \frac{P_{X|Y}(x_1, \ldots, x_n|0)}{P_{X|Y}(x_1, \ldots, x_n|1)} + \log \frac{P_Y(0)}{P_Y(1)} &> 0 \\ [\nu(\theta_0) - \nu(\theta_1)] \sum_i T(x_i) - n[A(\theta_0) - A(\theta_1)] + \log \frac{\pi_0}{\pi_1} &> 0 \\ s_n &> \frac{1}{[\nu(\theta_0) - \nu(\theta_1)]} \left([A(\theta_0) - A(\theta_1)] + \frac{1}{n} \log \frac{\pi_1}{\pi_0} \right), \end{aligned}$$

where we have assumed that $\nu(\theta_0) \geq \nu(\theta_1)$. Therefore, the threshold is, $T = \frac{1}{[\nu(\theta_0)-\nu(\theta_1)]} \left([A(\theta_0) - A(\theta_1)] + \frac{1}{n} \log \frac{\pi_1}{\pi_0} \right)$.

c)

$$
\begin{aligned}
KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)] &= E_{X|Y}\left[\log \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)}\middle| Y=1\right] \\
&= [\nu(\theta_1)-\nu(\theta_0)]E_{X|Y}\left[T(x)|Y=1\right] - [A(\theta_1)-A(\theta_0)] \\
&= -\left\{[\nu(\theta_0)-\nu(\theta_1)]E_{X|Y}\left[T(x)|Y=1\right] - [A(\theta_0)-A(\theta_1)]\right\} \\
KL[P_{X|Y}(x|0)||P_{X|Y}(x|1)] &= E_{X|Y}\left[\log \frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)}\middle| Y=0\right] \\
&= [\nu(\theta_0)-\nu(\theta_1)]E_{X|Y}\left[T(x)|Y=0\right] - [A(\theta_0)-A(\theta_1)]
\end{aligned}
$$

d) Note that the Bayes decision rule is to pick class $i = 0$

$$[\nu(\theta_0)-\nu(\theta_1)]\,\frac{1}{n}\sum_i T(x_i) - [A(\theta_0)-A(\theta_1)] > \frac{1}{n}\log\frac{\pi_1}{\pi_0}$$

and class 1 otherwise. As n goes to infinity this becomes

$$[\nu(\theta_0)-\nu(\theta_1)]\,E_X\left[T(x)\right] - [A(\theta_0)-A(\theta_1)] > 0$$

where X is the random variable from which the sample $\mathcal{D}$ to classify was drawn. If the sample comes from class 0, then $E_X\left[T(x)\right] = E_{X|Y}\left[T(x)|Y=0\right]$. If it came from class 1, then $E_X\left[T(x)\right] = E_{X|Y}\left[T(x)|Y=1\right]$. Hence, the left-hand side of the BDR becomes

$$KL[P_{X|Y}(x|0)||P_{X|Y}(x|1)]$$

for a sample from class 0 and

$$-KL[P_{X|Y}(x|1)||P_{X|Y}(x|0)]$$

for a sample from class 1. Since the KL divergence is never negative, this implies that class 0 will always be chosen when the sample comes from class 0 and class 1 will always be chosen for when the sample comes from class 0. This shows that the classifier has zero error.

2. We know that when the classes have equal probability and identity covariance, the boundary between class i and j is the hyperplane with normal

$$\mathbf{w}_{ij} = (\mu_i - \mu_j)$$

and bias

$$b_{ij} = \frac{\mu_i + \mu_j}{2}.$$

Let's consider the decision "say class 1", which is the set of points on the side of μ_1 of the hyperplanes $\mathbf{W}_j = (\mathbf{w}_{1j} = \mu_1 - \mu_j, b_{1j} = \frac{\mu_1+\mu_j}{2}), j = 2, 3, 4$. In particular

$$\begin{aligned} \mathbf{W}_2 &= ((0,2)^T, (1,0)^T) \\ \mathbf{W}_3 &= ((2,2)^T, (0,0)^T) \\ \mathbf{W}_4 &= ((2,0)^T, (0,1)^T) \end{aligned}$$

which are, respectively, the lines

$$\begin{aligned} x_2 &= 0 \\ x_2 &= -x_1 \\ x_1 &= 0 \end{aligned}$$

leading to the region $x_1 \geq 0, x_2 \geq 0, x_2 \geq -x_1$ which can be simplified to

$$\mathcal{R}_1 = \{\mathbf{x} | x_1 \geq 0, x_2 \geq 0\}.$$

Similarly, it can be shown that

$$\begin{aligned} \mathcal{R}_2 &= \{\mathbf{x} | x_1 \geq 0, x_2 \leq 0\} \\ \mathcal{R}_3 &= \{\mathbf{x} | x_1 \leq 0, x_2 \leq 0\} \\ \mathcal{R}_4 &= \{\mathbf{x} | x_1 \leq 0, x_2 \geq 0\}. \end{aligned}$$

b)

i) Note that the boundaries between classes $1, 2$, and 4 are the same, indicating that these classes maintain their parameters. The boundaries between classes 3 and 1,2 appear to have been shifted along the lines that unite the means. This implies that the priors have been changed, namely the prior $P_Y(3)$ is now smaller. The diagonal line, is part of the boundary between classes 2 and 4. It was previously redundant, but now becomes active. Once again it is consistent with the parameters of 1, 2, and 4 not changing.

A final note is that if $P_Y(3)$ was multiplied by $\alpha < 1$, the remaining probability $(1-\alpha)P_Y(3)$ must be divided by the three other classes. Note that the BDR only depends on the ratios between priors and distributing the probability equally by all classes is identical to "not changing the parameters". Hence, we must add $\frac{(1-\alpha)}{3}P_Y(3)$ to each of the remaining $P_Y(i)$. Overall the parameters are

$$\begin{aligned} \mu_i' &= \mu_i, \forall i \\ \Sigma_i' &= \Sigma_i, \forall i \\ P_Y(i)' &= P_Y(i) + \frac{(1-\alpha)}{3} P_Y(3), i \neq 3 \\ P_Y(i)' &= \alpha P_Y(i), i = 3 \end{aligned}$$

where the prime indicates the new values, and $\alpha < 1$.

ii) Note that the boundaries between classes $1, 2$, and 4 are the same, indicating that these classes maintain their parameters. We now seem to have the case where all boundaries have been pushed away from class 3. This could be accomplished by simply increasing the prior for that class. We therefore have a situation similar to **i)**. The new parameters are

$$\begin{aligned}
\mu_i' &= \mu_i, \forall i \\
\Sigma_i' &= \Sigma_i, \forall i \\
P_Y(i)' &= P_Y(i) - \frac{(\alpha-1)}{3} P_Y(3), i \neq 3 \\
P_Y(i)' &= \alpha P_Y(i), i = 3
\end{aligned}$$

with $\alpha > 1$.

iii) In this case the boundaries are obtained by shifting the original boundaries. This can be accomplished by subtracting a constant vector to all means and keeping everything else the same (i.e. by changing the origin). The new parameters are

$$\begin{aligned}
\mu_i' &= \mu_i - \mathbf{a}, \forall i \\
\Sigma_i' &= \Sigma_i, \forall i \\
P_Y(i)' &= P_Y(i), \forall i
\end{aligned}$$

with $a_1 > 0$ and $a_2 > 0$.

iv) This is similar to **i)**, but now both the priors of classes 1 and 3 have been decreased. From the symmetry it appears that they have been decreased by the same amount and the new parameters are

$$\begin{aligned}
\mu_i' &= \mu_i, \forall i \\
\Sigma_i' &= \Sigma_i, \forall i \\
P_Y(i)' &= P_Y(i) + \left(1 - \frac{\alpha}{2}\right) P_Y(3), i = 2, 4 \\
P_Y(i)' &= \frac{\alpha}{2} P_Y(i), i = 1, 3
\end{aligned}$$

where $\alpha < 1$.

v) In this case the boundaries all contain the mid-points between pairs of the original class means. This suggests that the means and the priors have not changed. The boundaries are however tilted by 45^o. This can be obtained by changing all the covariances so that the iso-contours of the Gaussians are elongated along the direction of 45^O. Hence, the new parameters are

$$\begin{aligned}
\mu_i' &= \mu_i, \forall i \\
\Sigma_i' &= \mathbf{S}, \forall i \\
P_Y(i)' &= P_Y(i), \forall i
\end{aligned}$$

where $\mathbf{S}$ is the matrix necessary to elongate the covariance. This can be shown to be the matrix with eigenvectors $1/\sqrt{2}(1, 1)$ and $1/\sqrt{2}(-1, 1)$ of eigenvalues λ and 1, respectively, where $\lambda >> 1$. Note that the boundary between classes 1 and 3 was previously redundant but is now exposed. This is the line of orientation 135^o through the origin.

vi) This is a combination of **v)** and **ii)**, but now the boundaries are pushed away from class 2. Hence, in addition to what we discussed for **v)**, the prior probability of class 2 was increased. The new parameters are

$$\begin{aligned}
\mu_i' &= \mu_i, \forall i \\
\Sigma_i' &= \mathbf{S}, \forall i \\
P_Y(i)' &= P_Y(i) - \frac{(\alpha - 1)}{3} P_Y(2), i \neq 2 \\
P_Y(i)' &= \alpha P_Y(i), i = 2
\end{aligned}$$

where $\mathbf{S}$ is as in **v**) and $\alpha > 1$.

vii) In this case, the only boundaries that have changed are those with class 1. Hence, only the parameters of this class are different. The boundaries with class 1 were pushed away and tilted. Note that this cannot be due to a covariance change because the boundaries are linear. So, all the covariances have to remain the same. Because the boundaries are tilted, the mean of class 1 must have changed. It could be that both the mean and probability have changed, but the simpler explanation is that the mean alone is different. In this case, the class probabilities would remain equal and the boundaries would go through the mid-points between pairs of means. Starting from each mean of class 2-4, you can sketch a line orthogonal to the boundary. These lines will intersect at a common point, which is the new mean of class 1. If you do this, you will see that the boundaries indeed contain the mid-points between mean pairs. This is consistent with the probabilities not changing. In summary, the only change was the multiplication of the mean of class 1 by some scaling factor larger than 1 (that pushes it away from the origin). The new parameters are

$$\begin{aligned}
\mu_i' &= \mu_i, \forall i \neq 1 \\
\mu_i' &= \alpha \mu_i, \forall i = 1 \\
\Sigma_i' &= \Sigma_i, \forall i \\
P_Y(i)' &= P_Y(i), \forall i
\end{aligned}$$

with $\alpha > 1$.

viii) This is the same as **vii)** but the reasoning applies to the means of class 1 and 3. Note that the symmetry of the figure suggests that the relationship $\mu_3 = -\mu_1$ continues to hold. Hence the new parameters are

$$\begin{aligned}
\mu_i' &= \mu_i, i \in \{2, 4\} \\
\mu_i' &= \alpha \mu_i, i \in \{1, 3\} \\
\\
\Sigma_i' &= \Sigma_i, \forall i \\
P_Y(i)' &= P_Y(i), \forall i
\end{aligned}$$

c) We have seen that the boundary between two Gaussian classes of equal variance is the hyperplane of parameters

$$\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j) \tag{1}$$

$$b = \frac{\mu_i + \mu_j}{2} + \alpha \log \frac{P_Y(i)}{P_Y(j)} (u_i - \mu_j). \tag{2}$$

where α is a constant that depends on the distance between the means. b only affects the position of the hyperplane along the line that connect the means. This holds for all plots and, in all the cases, the position is the mid-point between the means, which indicates equal class probabilities. Hence, all plots are in principle consistent with a BDR. To confirm this, we have to check if $\mathbf{w}$ is also consistent.

i) Since the Gaussian contours are circles, the Gaussians have covariance $\sigma^2 I$. Hence

$$\mathbf{w} \propto (\mu_i - \mu_j)$$

i.e. the plane should be orthogonal to the line that connects the means. Since this is indeed the case, the boundary is that of the BDR.

ii) Since the Gaussian contours are tilted ellipses, the Gaussians have non-diagonal covariance Σ. Hence

$$\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$$

is not along the direction of the line between the means. It follows that the plane cannot be orthogonal to the line that connects the means. Since this is the case, the boundary cannot be that of the BDR.

iii) This is exactly like **i)** and the boundary is that of the BDR.

iv) This is the trickiest case. One could say that this is like **ii)** and hence not a BDR. On the other hand, this can be obtained from **iii)** by simple application of a transformation

$$x' = RSx$$

where R is a rotation and S a scaling that stretches along the direction x_1. Since there is a one to one mapping between this case and **iii)**, which is a BDR, this must be a BDR. How can, however, this be different than **ii)**?. Well, note that it is not always the case that

$$\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$$

is not along the direction of $(\mu_i - \mu_j)$. This will happen if and only if $(\mu_i - \mu_j)$ is an eigenvector of Σ. The eigenvectors are the directions that define the orientation of the ellipse (the axes where the stretching occurs). In this plot $(\mu_i - \mu_j)$ is indeed aligned with one of these axes, which was not the case in **ii)**. Hence, unlike **ii)**, the boundary is that of the BDR.

3.a.i) As usual, we maximize the log of the cost. This is the log-likelihood of $\mathcal{D}$ under the predictive distribution, i.e.

$$\mathcal{L} = -\sum_i \frac{(x_i - \mu_n)^2}{2(\sigma^2 + \sigma_n^2)} - \frac{n}{2}\log(2\pi(\sigma^2 + \sigma_n^2)).$$

Setting the derivative wrt μ_0 to zero

$$\begin{aligned}
\frac{d\mathcal{L}}{d\mu_0} &= \frac{d\mathcal{L}}{d\mu_n}\frac{d\mu_n}{d\mu_0} \\
&= \sum_i \frac{2(x_i - \mu_n)}{2(\sigma^2 + \sigma_n^2)}(1 - \alpha_n) = 0
\end{aligned}$$

we obtain

$$\mu_n^* = \frac{1}{n}\sum_i x_i = \mu_{ML}$$

from which it follows that

$$\mu_{ML} = \alpha_n \mu_{ML} + (1 - \alpha_n)\mu_0^*$$

and $\mu_0^* = \mu_{ML}$.

3.a.ii) Plugging this value in the log-likelihood we have

$$\begin{aligned}
\mathcal{L}^*(\sigma_0^2) &= -\frac{n}{2(\sigma^2 + \sigma_n^2)}\frac{1}{n}\sum_i (x_i - \mu_{ML})^2 - \frac{n}{2}\log(2\pi(\sigma^2 + \sigma_n^2)) \\
&\propto -\frac{1}{(\sigma^2 + \sigma_n^2)}\nu - \log(\sigma^2 + \sigma_n^2)
\end{aligned}$$

where $\nu = \frac{1}{n}\sum_i (x_i - \mu_{ML})^2$ is the sample variance of $\mathcal{D}$ and we dropped all constants. Setting the derivative wrt σ_0^2 to zero we obtain

$$\begin{aligned}
\frac{d\mathcal{L}}{d\sigma_0^2} &= \frac{d\mathcal{L}}{d\sigma_n^2}\frac{d\sigma_n^2}{d\sigma_0^2} \\
&= \left(\frac{1}{(\sigma^2 + \sigma_n^2)^2}\nu - \frac{1}{\sigma^2 + \sigma_n^2}\right)\frac{\sigma^2}{n}\frac{n(\sigma^2 + n\sigma_0^2) - n\sigma_0^2 n}{(\sigma^2 + n\sigma_0^2)^2} \\
&= \frac{\sigma^2}{(\sigma^2 + \sigma_n^2)}\left(\frac{\nu}{(\sigma^2 + \sigma_n^2)} - 1\right)\frac{\sigma^2}{(\sigma^2 + n\sigma_0^2)^2} = 0.
\end{aligned}$$

For finite n this holds if $\sigma_0^2 = \infty$ or

$$\nu = \sigma^2 + \sigma_n^2 = \sigma^2\left(1 + \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2}\right) \Leftrightarrow \frac{\sigma_0^2}{\sigma^2 + n\sigma_0^2} = \frac{\nu}{\sigma^2} - 1 \Leftrightarrow \sigma_0^2 = \frac{\sigma^2\left(\frac{\nu}{\sigma^2} - 1\right)}{1 - n\left(\frac{\nu}{\sigma^2} - 1\right)}.$$

Hence

$$(\sigma_0^*)^2 = \frac{\sigma^2\left(\frac{\nu}{\sigma^2} - 1\right)}{1 - n\left(\frac{\nu}{\sigma^2} - 1\right)}.$$

3.b.i) The complete data log-likelihood is

$$\log P_{x,\mu;\mu_0,\sigma_0^2}(\mathcal{D}, \mu; \mu_0, \sigma_0^2) = \log P_{x|\mu}(\mathcal{D}|\mu) + \log P_{\mu;\mu_0,\sigma_0^2}(\mu; \mu_0, \sigma_0^2)$$

$$
\begin{aligned}
&= -\sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{1}{2}\log(2\pi\sigma_0^2) \\
&\propto -\frac{1}{2\sigma^2}\left(-2\mu\sum_i x_i + n\mu^2\right) - \frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) - \frac{1}{2}\log(\sigma_0^2) \\
&\propto -\frac{n}{\sigma^2}\left(-2\mu\mu_{ML} + \mu^2\right) - \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2) - \log(\sigma_0^2) \\
&= -\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 + 2\left(\frac{n\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right)\mu - \frac{\mu_0^2}{\sigma_0^2} - \log(\sigma_0^2)
\end{aligned}
$$

where we have dropped constants that do not depend on the prior parameters.

3.b.ii) The Q function is

$$
\begin{aligned}
Q[(\mu_0,\sigma_0^2)|(\mu_0,\sigma_0^2)^{(i)}] &= E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\log P_{x,\mu;\mu_0,\sigma_0^2}(\mathcal{D},\mu;\mu_0,\sigma_0^2)|\mathcal{D}] \\
&= -\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu^2|\mathcal{D}] + 2\left(\frac{n\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu|\mathcal{D}] - \frac{\mu_0^2}{\sigma_0^2} - \log(\sigma_0^2) \\
&= -\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right) < \mu^2 > +2\left(\frac{n\mu_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) < \mu > -\frac{\mu_0^2}{\sigma_0^2} - \log(\sigma_0^2) \\
&\propto -\frac{1}{\sigma_0^2}(< \mu^2 > -2\mu_0 < \mu > +\mu_0^2) - \log(\sigma_0^2),
\end{aligned}
$$

with

$$
\begin{aligned}
< \mu^2 > &= E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu^2|\mathcal{D}] \qquad (3) \\
< \mu > &= E_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}[\mu|\mathcal{D}]. \qquad (4)
\end{aligned}
$$

Note that these are just the first and second order moments of the posterior for μ under a prior of parameters $(\mu_0,\sigma_0^2)^{(i)}$. We know, from the table given with the problem, that

$$
P_{\mu|x;(\mu_0,\sigma_0^2)^{(i)}}(\mu|\mathcal{D}) = G(\mu, \mu_n^{(i)}, (\sigma_n^2)^{(i)}).
$$

It follows that

$$
\begin{aligned}
< \mu^2 > &= (\sigma_n^2)^{(i)} + (\mu_n^{(i)})^2 \\
< \mu > &= \mu_n^{(i)}
\end{aligned}
$$

and

$$
\begin{aligned}
< \mu > &= \alpha_n^{(i)}\mu_{ML} + (1-\alpha_n^{(i)})\mu_0^{(i)} \\
< \mu^2 > &= \frac{\sigma^2}{n}\alpha_n^{(i)} + < \mu >^2 \\
\alpha_n^{(i)} &= \frac{n(\sigma_0^2)^{(i)}}{\sigma^2 + n(\sigma_0^2)^{(i)}}
\end{aligned}
$$

These are the computations of the E-step.

3.b.iii) To compute the parameter updates we set the derivatives of the Q function to zero. This leads to

$$
\frac{dQ}{d\mu_0} = \frac{2}{\sigma_0^2}(< \mu > -\mu_0) = 0
$$

$$\begin{aligned}
\mu_0^{(i+1)} &= <\mu> \\
\frac{dQ}{d\sigma_0^2} &= \frac{1}{(\sigma_0^2)^2}(<\mu^2> -2\mu_0 <\mu> +\mu_0^2) - \frac{1}{\sigma_0^2} \\
&= \frac{1}{(\sigma_0^2)^2}(<\mu^2> - <\mu>^2) - \frac{1}{\sigma_0^2} = 0 \\
(\sigma_0^2)^{(i+1)} &= <\mu^2> - <\mu>^2 .
\end{aligned}$$

c) From EM equations

$$\begin{aligned}
\mu_0^{(i+1)} &= \mu_n^{(i)} = \alpha_n^{(i)}\mu_{ML} + (1-\alpha_n^{(i)})\mu_0^{(i)} \\
(\sigma_0^2)^{(i+1)} &= (\sigma_n^2)^{(i)} = \frac{\sigma^2}{n}\alpha_n^{(i)}
\end{aligned}$$

At convergence we have

$$\begin{aligned}
\mu_0^* &= \alpha_n^*\mu_{ML} + (1-\alpha_n^*)\mu_0^* \\
(\sigma_0^2)^* &= \frac{\sigma^2}{n}\alpha_n * \\
&= \frac{\sigma^2(\sigma_0^2)^*}{\sigma^2 + n(\sigma_0^2)^*}
\end{aligned}$$

from which

$$\begin{aligned}
\mu_0^* &= \mu_{ML} \\
\sigma^2 + n(\sigma_0^2)^* &= \sigma^2 \\
(\sigma_0^2)^* &= 0
\end{aligned}$$

i.e. the prior converges to a delta function centered at μ_{ML}. Hence

$$\alpha_n^* = 0$$

and the predictive distribution is a Gaussian of mean μ_{ML} and variance σ^2, i.e. the ML model. For predictive-ML, we have

$$\begin{aligned}
\mu_n^* &= \mu_{ML} \\
\alpha_n^* &= \frac{\frac{n\sigma^2\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}}{\sigma^2 + \frac{n\sigma^2\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}} = \frac{\frac{n\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}}{1 + \frac{n\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}} = \frac{n\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)} \\
(\sigma_n^2)^* &= \frac{\sigma^2}{n}\alpha_n^* = \frac{\sigma^2\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}
\end{aligned}$$

and the predictive distribution is a Gaussian of mean μ_{ML} and variance

$$\sigma^2\left[1 + \frac{\left(\frac{\nu}{\sigma^2}-1\right)}{1-n\left(\frac{\nu}{\sigma^2}-1\right)}\right].$$

For a large sample ν converges to σ^2 and this is well approximated by σ^2. Hence the two empirical Bayesian procedures produce the same predictive distribution, which is the ML model, for large samples. This is not surprising because Bayes and ML tend to be equivalent for large samples.