HaoTong

2.2 The form of the Bernoulli distribution given by (2.2) is not symmetric between the two values of x. In some situations, it will be more convenient to use an equivalent formulation for which $x \in \{-1, 1\}$, in which case the distribution can be written

$$p(x|\mu) = \left(\frac{1-\mu}{2}\right)^{(1-x)/2} \left(\frac{1+\mu}{2}\right)^{(1+x)/2}$$

where $\mu \in [-1,1]$. Show that the distribution (2.261) is normalized, and evaluate its mean, variance, and entropy.

Sol. Because Bernoulli only has two situations. $x=1$, and $x=-1$.

So. $p(x=1|\mu) + p(x=-1|\mu) = \frac{1+\mu}{2} + \frac{1-\mu}{2} = 1$

mean: $E(x) = \sum x p(x) = 1 \cdot p(x=1|\mu) + (-1) \cdot p(x=-1|\mu) = \frac{1+\mu}{2} - \frac{1-\mu}{2} = \mu$

variance $\cancel{E(x)}$ $Var(x) = E[(x-\mu)^2] = \cancel{E} E(x^2) - E^2(x) = 1^2 p(x=1|\mu) + (-1)^2 p(x=-1|\mu) - \mu^2 = 1-\mu^2$

entropy: $H(x) = -\sum p(x) \log p(x) = -[p(x=1|\mu) \cdot \log p(x=1|\mu) + p(x=-1|\mu) \cdot \log p(x=-1|\mu)]$

$$= -\left[\frac{1+\mu}{2} \cdot \log \frac{1+\mu}{2} + \frac{1-\mu}{2} \cdot \log \frac{1-\mu}{2}\right]$$

$$= -\frac{1+\mu}{2} \log \frac{1+\mu}{2} - \frac{1-\mu}{2} \cdot \log \frac{1-\mu}{2}$$

2.3. In this exercise, we prove that binomial distribution (2.9) is normalised. First use the definion (2.10) of the ~~nut~~ number of combinations of m identical objects choosen from a total of N to show that.

$$\binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$$

use this result to prove by induction the following result

$$(1+x)^N = \sum_{m=0}^{N} \binom{N}{m} x^m$$

which is known as the ~~binomial~~ binomial theorem, and which is valid for all real values of x. Finally, show that the binomial distrubition is normalized, so that

$$\sum_{m=0}^{N} \binom{N}{m} \mu^m (1-\mu)^{N-m} = 1$$

which can be done by first pulling out a factor $(1-\mu)^N$ out of the summation and then making use of the binomial theorem.

Sol. ① $\binom{N}{m} = \frac{N!}{(N-m)!m!}$

$$\binom{N}{m} + \binom{N}{m-1} = \frac{N!}{(N-m)!m!} + \frac{N!}{(N-m+1)!(m-1)!} = \frac{N!(N+1-m)}{(N-m+1)!\,m!} + \frac{N!\,m}{(N-m+1)!\,m!}$$

$$= \frac{N!(N+1)}{(N-m+1)!\,m!} = \frac{(N+1)!}{(N+1-m)!\,m!} = \binom{N+1}{m}$$

$$\therefore \binom{N}{m} + \binom{N}{m-1} = \binom{N+1}{m}$$

② $$(1+x)^{N+1} = (1+x)\sum_{m=0}^{N}\binom{N}{m}x^m = \sum_{m=0}^{N}\binom{N}{m}x^m + \sum_{m=0}^{N}\binom{N}{m}x^{m+1}$$

$$\sum_{m=0}^{N}\binom{N}{m}x^m = \binom{N}{0}x^0 + \sum_{m=1}^{N}\binom{N}{m}x^m = 1 + \sum_{m=1}^{N}\binom{N}{m}x^m$$

$$\sum_{m=0}^{N}\binom{N}{m}x^{m+1} = \sum_{m=1}^{N+1}\binom{N}{m-1}x^m = \sum_{m=1}^{N}\binom{N}{m-1}x^m + \binom{N}{N}x^{N+1} = \sum_{m=1}^{N}\binom{N}{m-1}x^m + x^{N-1}$$

$$\therefore (1+x)^{N+1} = 1 + \sum_{m=1}^{N}\binom{N}{m}x^m + \sum_{m=1}^{N}\binom{N}{m-1}x^m + x^{N+1} = 1 + x^{N+1} + \sum_{m=1}^{N}\binom{N+1}{m}x^m$$

$m=0 \Rightarrow \binom{N+1}{0}x^0 = 1$. $\quad \therefore (1+x)^{N+1} = \sum_{m=0}^{N+1}\binom{N+1}{m}x^m \quad \therefore (1+x)^N = \sum_{m=0}^{N}\binom{N}{m}x^m$

$m=N+1 \Rightarrow \binom{N+1}{N+1}x^{N+1} = x^{N+1}$

③ $$\sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = \sum_{m=0}^{N}\binom{N}{m}\cdot\mu^m\cdot\frac{(1-\mu)^N}{(1-\mu)^m} = (1-\mu)^N\cdot\sum_{m=0}^{N}\binom{N}{m}\left(\frac{\mu}{1-\mu}\right)^m$$

from above. $\sum_{m=0}^{N}\binom{N}{m}\left(\frac{\mu}{1-\mu}\right)^m = \left(1+\frac{\mu}{1-\mu}\right)^N = \left(\frac{1}{1-\mu}\right)^N$

$$\therefore \sum_{m=0}^{N}\binom{N}{m}\mu^m(1-\mu)^{N-m} = (1-\mu)^N\left(\frac{1}{1-\mu}\right)^N = 1^N = 1$$

2.6. Make use of the result (2.265) to show that the mean, variance, and mode of the beta distribution (2.13) are given respectively by

$$E[\mu] = \frac{a}{a+b}, \quad var[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

$$mode[\mu] = \frac{a-1}{a+b-2}$$

Sol. $$E[\mu] = \int_0^1 \mu\cdot\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\cdot\int_0^1\mu^a(1-\mu)^{b-1}d\mu = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}$$

$$= \frac{(a+b-1)!}{(a-1)!\,(b-1)!}\cdot\frac{a!\,(b-1)!}{(a+b)!} = \frac{a}{a+b}$$

$$var[\mu] = E[\mu^2] - E[\mu]^2 = \frac{(a+1)a}{(a+b+1)(a+b)} - \left(\frac{a}{a+b}\right)^2 = \frac{a^3+a^2+ab+a^2b}{(a+b+1)(a+b)^2} - \frac{a^2+a^2b+a^3}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)^2}$$

the mode ~~o~~ occurs where the distribution reaches a maximum, where the derivate is 0.

$$\frac{d\,Beta}{d\mu} = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left[(a-1)\mu^{a-2}\cdot(1-\mu)^{b-1} + \mu^{a-1}(1-\mu)^{b-2}\cdot(b-1)\cdot(-1)\right] = 0$$

$$\therefore\ (a-1)\mu^{a-2}(1-\mu)^{b-1} = \mu^{a-1}(1-\mu)^{b-2}(b-1)$$

$$(a-1)(1-\mu) = \mu(b-1) \qquad \Rightarrow \mu = \frac{a-1}{a+b-2}$$

$$(a-1) - \mu(a-1) = \mu(b-1)$$

$$\therefore\ mode[\mu] = \frac{a-1}{a+b-2}$$

2.10. Using the property $\Gamma(x+1) = x\Gamma(x)$ of the gamma function, derive the following results for the mean, ~~va~~ variance, and covariance of the Dirichlet distribution given by (2.38)

$$E[\mu_j] = \frac{\alpha_j}{\alpha_0} \qquad Var[\mu_j] = \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}$$

$$cov[\mu_j, \mu_l] = -\frac{\alpha_j\alpha_l}{\alpha_0^2(\alpha_0+1)}$$

where α_0 is defined by (2.39).

Sol. $Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \cdot \prod_{k=1}^{K} \mu_k^{\alpha_k-1}$

from $\int Dir(\mu|\alpha) = 1$

$$\Rightarrow \int \prod_{k=1}^{K} \mu_k^{\alpha_k-1} = \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0)}$$

$$E[\mu_j] = \int_0^1 \mu_j Dir(\mu|\alpha)\,d\mu = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \cdot \int \mu_j \cdot \prod_{k=1}^{K} \mu_k^{\alpha_k-1}\,d\mu$$

$$\therefore E[\mu_j] = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \cdot \frac{\Gamma(\alpha_1)\cdots\Gamma(\alpha_j+1)\cdots\Gamma(\alpha_K)}{\Gamma(\alpha_0+1)} = \frac{\Gamma(\alpha_j+1)}{\Gamma(\alpha_j)} \times \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+1)} = \frac{\alpha_j}{\alpha_0}$$

$$Var[\mu_j] = E[\mu_j^2] - E^2[\mu_j] = \frac{\Gamma(\alpha_j+2)}{\Gamma(\alpha_j)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+2)} - \frac{\alpha_j^2}{\alpha_0^2} = \frac{\alpha_j(\alpha_j+1)}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j^2}{\alpha_0^2} = \frac{\alpha_j(\alpha_0-\alpha_j)}{\alpha_0^2(\alpha_0+1)}$$

$$Cov[\mu_j,\mu_l] = E[\mu_j \mu_l] - E[\mu_j]E[\mu_l] = \frac{\Gamma(\alpha_j+1)\Gamma(\alpha_l+1)}{\Gamma(\alpha_j)\Gamma(\alpha_l)} \cdot \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+2)} - \frac{\alpha_j}{\alpha_0}\cdot\frac{\alpha_l}{\alpha_0}$$

$$= \frac{\alpha_j\alpha_l}{\alpha_0(\alpha_0+1)} - \frac{\alpha_j\alpha_l}{\alpha_0^2} = \frac{-\alpha_j\alpha_l}{\alpha_0^2(\alpha_0+1)}$$

2.42. Evaluate the mean, variance, and mode of the Gamma ~~distri~~ distribution (2.146).

Sol. $Gam(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} exp(-b\lambda)$

$\frac{1}{\Gamma(a)}$ is the coefficient of normalization. So. $\int b^a \lambda^{a-1} exp(-b\lambda)\,d\lambda = \Gamma(a)$

$$\therefore E[\lambda] = \frac{1}{\Gamma(a)} \int \lambda \cdot b^a \cdot \lambda^{a-1} exp(-b\lambda)\,d\lambda = \frac{1}{\Gamma(a)} \cdot \frac{1}{b} \cdot \int b^{a+1} \lambda^a exp(-b\lambda)\,d\lambda = \frac{\Gamma(a+1)}{\Gamma(a)\,b} = \frac{a}{b}$$

$$Var[\lambda] = E[\lambda^2] - E^2[\lambda] = \frac{\Gamma(a+2)}{\Gamma(a)\cdot b^2} - \frac{a^2}{b^2} = \frac{(a+1)a}{b^2} - \frac{a^2}{b^2} = \frac{a^2}{b^2}$$

$$\frac{d\,Gam}{d\lambda} = \frac{1}{\Gamma(a)} \cdot b^a \cdot (a-1) \cdot \lambda^{a-2} e^{-b\lambda} + \frac{1}{\Gamma(a)} \cdot b^a \lambda^{a-1} e^{-b\lambda} \cdot (-b) = 0$$

$$(a-1)\lambda^{a-2} \cdot e^{-b\lambda} = \lambda^{a-1} e^{-b\lambda} b \quad \Rightarrow \lambda = \frac{a-1}{b} \qquad \therefore\ mode[\lambda] = \frac{a-1}{b}$$

3.2. Show that the matrix $\Phi(\Phi^T\Phi)^{-1}\Phi^T$ takes any vector $\bar{v}$ and projects it onto the space spanned by the columns of Φ. Use this result to show that the least-squares solution (3.15) corresponds to an orthogonal projection of the vector t onto the mainfold S as shown in Figure 3.2.

Sol. from (3.15), we know $W_{ML} = (\Phi^T\Phi)^{-1}\Phi^T t$.

we can get. $\bar{y} = \Phi \cdot W_{ML} = \Phi(\Phi^T\Phi)^{-1}\Phi^T t$

and. $\bar{y}$ is a vector in the space which spanned by the columns of Φ.

So. matrix $\Phi(\Phi^T\Phi)^{-1}\Phi^{-1}$ projects a vector onto the space spanned by the columns of Φ.

And if the projection is an orthognal projection, the vector from orign to end must be or vertical with any vector is space S. ~~It means The formula is list at below.~~

~~It means $(y-t)^T \cdot \varphi_j$~~ It means $(y-t)^T \cdot \varphi_j$ must equal to 0.

$$(y-t)^T\varphi_j = (\phi W_{ML} - t)^T\varphi_j = (\Phi(\Phi^T\Phi)^{-1}\Phi^T t - t)^T \cdot \varphi_j = t^T \cdot (\Phi(\Phi^T\Phi)^{-1}\Phi^T - I)^T\varphi_j$$

$$= t^T([\Phi(\Phi^T\Phi)^{-1}\Phi^T]^T - I^T)\varphi_j = t(\Phi \cdot (\Phi^T\Phi)^{-1} \cdot \Phi^T - I)\varphi_j$$

$$= t \cdot (\Phi(\Phi^T\Phi)^{-1}\Phi^T\varphi_j - \varphi_j)$$

$$\Phi \cdot (\Phi^T\Phi)^{-1} \cdot \Phi^T\varphi_j = [\Phi(\Phi^T\Phi)^{-1} \cdot \Phi^T \cdot \Phi]_j = [\Phi]_j = \varphi_j$$

$$\therefore (y-t)^T\varphi_j = t(\varphi_j - \varphi_j) = 0$$

$\therefore$ the projection is an orthognol projection

3.3. Consider a data set in which each data point t_n is associated with a weighting factor $r_n > 0$, so that the sum-of-squares error function becomes

$$E_D(w) = \frac{1}{2}\sum_{n=1}^{N} r_n\{t_n - w^T\phi(x_n)\}^2$$

Find an expression for the solution w^* that minimizes this error function. Give two alternative interpretations of the weighted sum-of-squares error function in terms of (i) data dependent noise and (ii) replicated data points.

Sol. Loss Function $E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2$

It can be written as $E_D(w) = \frac{1}{2}(t - \phi w)^T \cdot (t - \phi w)$

t is a vector. w is a vector. Φ is a matrix

So, if $E_D(w)=\frac{1}{2}\sum_{n=1}^{N} r_n\{t_n-w^T\Phi(x_n)\}^2$

It can be written as $E_D(w)=\frac{1}{2}(t-\phi w)^T\cdot R\cdot(t-\phi w)$

where t, w are vectors, ϕ is ~~sample~~ the matrix. $R=diag(r_1, r_2, \cdots r_n)$

$\therefore$ from formula $\frac{\partial}{\partial s}(x-As)^T W(x-As)=-2A^T W(x-As)$

we can get $\frac{\partial E_D(w)}{\partial w}=-2\times\frac{1}{2}\times\phi^T R\cdot(t-\phi w)=0$

so. $\phi^T R\cdot t=\phi^T R\cdot\phi w \Rightarrow w=(\phi^T R\phi)^{T}\phi^T R t$

From 3.10 ~ 3.12, we can see the r_n can be regarded as β (precision), the precision of noise data.

Alternatively, r_n can be regarded as an effective number of replicated ~~data~~ observations of data point (x_n, t_n); this becomes particularly clear if we consider 3.104 with r_n taking positive values, although it is valid for any $r_n>0$

3.8. Consider the linear basis function model in Section 3.1, and suppose that we have already observed N data points, so that the posterior distribution over w is given by (3.49). This posterior can be regarded as the perior for the next obervation. By considering an additional data point (x_{N+1}, t_{N+1}), and by completing the square in the exponential, show that the resulting posterior distribution is again given by (3.49) but with S_N replaced by S_{N+1} and m_N replaced by m_{N+1}.

Sol. The posterior distribution is the product of prior distribution and likilyhood.

$$p(w|t)=N(w|m_N, S_N)$$
$$=\frac{1}{\sqrt{(2\pi)^D|S_N|}}\exp\{-\frac{1}{2}(w-m_N)^T S_N^{-1}(w-m_N)\}$$

likelihood. $p(t_{N+1}|x_{N+1}, w)=(\frac{\beta}{2\pi})^{\frac{1}{2}}e^{-\frac{\beta}{2}\{t_{N+1}-w^T\phi(x_{N+1})\}^2}$

$\therefore$ $p(w|t)\cdot p(t_{N+1}|x_{N+1}, w)\propto\exp\{-\frac{1}{2}(w-m_N)^T S_N^{-1}(w-m_N)-\frac{\beta}{2}(t_{N+1}-w^T\phi(x_{N+1}))^2\}$

we only need to concern the exponential part.

$$(w-m_N)^T S_N^{-1}(w-m_N)+\beta[t_{N+1}-w^T\phi(x_{N+1})]^2$$
$$=w^T S_N^{-1}w-m_N^T S_N^{-1}w-w^T S_N^{-1}m_N+m_N^T S_N^{-1}m_N+\beta t_{N+1}^2-2\beta t_N w^T\phi(x_{N+1})+w^T\phi(x_{N+1})\phi(x_{N+1})^T\cdot w$$

$w^T(S_N^{-1}+\beta\phi(x_{N+1})\phi_{N+1}^T)w - 2w^T(S_N^{-1}m_N+\beta\phi_{N+1}t_{N+1})+C$

So the posterior distribution also a normal distribution.

$$S_{N+1}^{-1} = S_N^{-1}+\beta\phi(x_{N+1})\phi_{N+1}^T$$

$$m_{N+1} = S_{N+1}(S_N^{-1}m_N+\beta\phi(x_{N+1})t_{N+1})$$

from the (3.50) and (3.51) we can see both two formula has the same form.

3.12. We saw in Section 2.3.6 that the conjugate prior for a Gaussian distribution with unknown mean and unknown precision (inverse variance) is a normal-gamma distribution. This property also holds for the case of the conditional Gaussian distribution $p(t|x,w,\beta)$ of the linear regression model. If we consider the likelihood function (3.10), then the conjugate prior for w and β is given by

$$p(w,\beta) = \mathcal{N}(w|m_0,\beta^{-1}S_0)\,\mathrm{Gam}(\beta|a_0,b_0)$$

Show that the corresponding posterior distribution takes the same functional form.

So that $$p(w,\beta|t) = \mathcal{N}(w|m_N,\beta^{-1}S_N)\,\mathrm{Gam}(\beta|a_N,b_N)$$

Sol. $p(w,\beta) = \mathcal{N}(w|m_0,\beta^{-1}S_0)\,\mathrm{Gam}(\beta|a_0,b_0)$

$p(t|X,w,\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|w^T\phi(x_n),\beta^{-1})$

$p(w,\beta|t) = \mathcal{N}(w|m_0,\beta^{-1}S_0)\,\mathrm{Gam}(\beta|a_0,b_0)\cdot\prod_{n=1}^{N}\mathcal{N}(t_n|w^T\phi(x_n),\beta^{-1})$

$$=\frac{1}{|2\pi|^{\frac{D}{2}}|\beta^{-1}S_0|^{\frac12}}\exp\{-\tfrac12(w-m_0)^T\beta S_0^{-1}(w-m_0)\}\cdot\frac{1}{\Gamma(a)}b_0^{a_0}\beta^{a_0-1}\exp(-b_0\beta)\cdot\prod_{n=1}^{N}\frac{1}{(2\pi)^{\frac{D}{2}}|\beta^{-1}|^{\frac12}}$$

$$\cdot\exp\{-\tfrac12\sum_{n=1}^{N}\frac{(t_n-w^T\phi x_n)^2}{\beta^{-1}}\}$$

$$= C_1\cdot b_0^{a_0}\beta^{a_0-1}\cdot\exp\{-\tfrac{\beta}{2}(w^TS_0^{-1}w-2w^TS_0^{-1}m_0+m_0^TS_0^{-1}m_0)-b_0\beta-\tfrac{\beta}{2}\sum_{n=1}^{N}(t_n-w^T\phi(x_n))^2\}$$

$$= C_1 b_0^{a_0}\beta^{a_0-1}\exp\{-\tfrac{\beta}{2}w^T(S_0^{-1}+\phi^T\phi)w+w^T(\beta S_0^{-1}m_0+\beta\phi^Tt)$$
$$-\beta(\tfrac12 m_0^TS_0^{-1}m_0+b_0+\tfrac12\sum_{n=1}^{N}t_n^2)+C_2\}$$

$p(w,\beta|t) = \mathcal{N}(w|m_N,\beta^{-1}S_N)\,\mathrm{Gam}(\beta|a_N,b_N)$

$$=\frac{1}{|2\pi|^{\frac{D}{2}}|\beta^{-1}S_N|^{\frac12}}\exp\{-\tfrac{\beta}{2}(w^TS_N^{-1}w-2w^TS_N^{-1}m_N+m_N^TS_N^{-1}m_N\}\cdot\frac{1}{\Gamma(a_N)}b_N^{a_N}\beta^{a_N-1}\cdot\exp\{-b_N\beta\}$$

$$= C\cdot b_N^{a_N}\cdot\beta^{a_N-1}\exp\{-\tfrac{\beta}{2}(w^TS_N^{-1}w-2w^TS_N^{-1}m_N+m_N^TS_N^{-1}m_N)-b_N\beta\}$$

$$= C\cdot b_N^{a_N}\cdot\beta^{a_N-1}\exp\{-\tfrac{\beta}{2}(w^TS_N^{-1}w)+\beta\cdot w^TS_N^{-1}m_N-\beta(\tfrac12 m_N^TS_N^{-1}m_N+b_N)\}$$

$\therefore$ $S_N^{-1} = S_0^{-1} + \phi^T\phi$

$S_N^{-1} m_N = S_0^{-1} m_0 + \phi^T t$

$m_N = S_N(S_0^{-1} m_0 + \phi^T t)$

$\frac{1}{2} m_N^T S_N^{-1} m_N + b_N = \frac{1}{2} m_0^T S_0^{-1} m_0 + b_0 + \frac{1}{2}\sum_{n=1}^{N} t_n^2$

$b_N = b_0 + \frac{1}{2}(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + \sum_{n=1}^{N} t_n^2)$

$\beta^{\frac{1}{2}} \cdot \beta^{a_0 - 1} \cdot \prod_{n=1}^{N} \beta^{\frac{1}{2}} = \beta^{\frac{1}{2}} \cdot \beta^{a_N - 1}$

$\beta^{a_0 - 1 + \frac{1}{2} + \frac{N}{2}} = \beta^{a_N - 1 + \frac{1}{2}}$

$\therefore a_N = a_0 + \frac{N}{2}$

3.15. Consider a linear basis function model for regression in which the parameters α and β are set using the evidence framework. Show that the function $E(m_N)$ defined by 3.82 satisfies the relation $2E(m_N) = N$.

Sol. we can use the formula ~~3.29~~ 3.92 and 3.95.

$\alpha = \frac{r}{m_N^T m_N}$ $\qquad \frac{1}{\beta} = \frac{1}{N-r}\sum_{n=1}^{N}\{t_n - m_N^T \phi(x_n)\}^2 = \frac{1}{N-r}\|t_n - \phi m_N\|^2$

$\therefore$ $m_N^T m_N = \frac{r}{\alpha}$ $\qquad \|t_n - \Phi m_N\|^2 = \frac{N-r}{\beta}$

$\therefore$ $E(m_N) = \frac{\beta}{2}\|t - \Phi m_N\|^2 + \frac{\alpha}{2} m_N^T m_N = \frac{\beta}{2}\cdot\frac{N-r}{\beta} + \frac{\alpha}{2}\cdot\frac{r}{\alpha} = \frac{N-r}{2} + \frac{r}{2} = \frac{N}{2}$

$\therefore$ $2E(m_N) = N$

3.19. Show that the integration over w in the Bayesian linear regression model gives the result (3.85). Hence show that the log marginal likelihood is given by (3.86).

Sol. From the standard Gaussian Distribution, we can get

$\int N(x|\mu, \Sigma) = \int \frac{1}{(2\pi)^{\frac{D}{2}}} \cdot \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\} = 1$

$\int \exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\} = (2\pi)^{\frac{D}{2}} \cdot |\Sigma|^{\frac{1}{2}}$

So. in the (3.85). $\int \exp\{-\frac{1}{2}(w - m_N)^T A (w - m_N)\} dw = (2\pi)^{\frac{M}{2}} |A^{-1}|^{\frac{1}{2}}$

$\therefore$ $\int \exp\{-E(w)\} dw = \exp\{-E(m_N)\}(2\pi)^{\frac{M}{2}} |A|^{-\frac{1}{2}}$

$\therefore$ $\ln p(t|\alpha, \beta) = \ln (\frac{\beta}{2\pi})^{\frac{N}{2}} \cdot (\frac{\alpha}{2\pi})^{\frac{M}{2}} \cdot \exp\{-E(m_N)\}(2\pi)^{\frac{M}{2}} |A|^{-\frac{1}{2}}$

$= \frac{N}{2}\ln\beta + \frac{M}{2}\ln\alpha - E(m_N) - \frac{1}{2}\ln|A| - \frac{N}{2}\ln 2\pi$

3.23 Show that the marginal probability of the data, in other words the model evidence, for the model described in Exercise 3.12 is given by

$$p(t) = \frac{1}{(2\lambda)^{N/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a_0)} \frac{|S_N|^{1/2}}{|S_0|^{1/2}}$$

by first marginalizing ~~th~~ with respect to w and then with respect to β.

Sol. $p(t) = \iint p(t|w,\beta)\, p(w|\beta)\, dw\, p(\beta)\, d\beta$, from $p(t) = \int p(w,\beta)\, p(t|w,\beta)\, dw\, d\beta$

$$= \iint \left(\frac{\beta}{2\lambda}\right)^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}(t-\Phi w)^T(t-\phi w)\right\} \left(\frac{\beta}{2\lambda}\right)^{\frac{M}{2}} |S_0|^{-\frac{1}{2}} \exp\left\{-\frac{\beta}{2}(w-m_0)^T S_0^{-1}(w-m_0)\right\} dw$$
$$\Gamma(a_0)^{-1} b_0^{a_0} \beta^{a_0-1} \exp(-b_0\beta)\, d\beta$$

$$= \frac{b_0^{a_0}}{((2\lambda)^{M+N}|S_0|)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(t-\Phi w)^T(t-\Phi w)\right\} \exp\left\{-\frac{\beta}{2}(w-m_0)^T S_0^{-1}(w-m_0)\right\} dw$$
$$\beta^{a_0-1} \beta^{N/2} \beta^{M/2} \exp\{-b_0\beta\}\, d\beta$$

$$= \frac{b_0^{a_0}}{((2\lambda)^{M+N}|S_0|)^{1/2}} \iint \exp\left\{-\frac{\beta}{2}(w-m_N)^T S_N^{-1}(w-m_N)\right\} dw \exp\left\{-\frac{\beta}{2}(t^T t + m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N)\right\}$$
$$\beta^{a_N-1} \beta^{M/2} \exp\{-b_0\beta\}\, d\beta$$

where we have completed the square for the quadratic form in w, using

$$m_N = S_N[S_0^{-1} m_0 + \phi^T t]$$
$$\beta S_N^{-1} = \beta(S_0^{-1} + \phi^T\phi)$$
$$a_N = a_0 + \frac{N}{2}$$
$$b_N = b_0 + \frac{1}{2}\left(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + \sum_{n=1}^{N} t_n^2\right)$$

Now, we are ready to do the intergration, first over w and then β. and re-arrange the terms to obtain the desired result

$$p(t) = \frac{b_0^{a_0}}{((2\lambda)^{M+N}|S_0|)^{1/2}} (2\lambda)^{M/2} |S_N|^{1/2} \int \beta^{a_N-1} \exp\{-b_N\beta\}\, d\beta = \frac{1}{(2\lambda)^{N/2}} \frac{|S_N|^{1/2}}{|S_0|^{1/2}} \frac{b_0^{a_0}}{b_N^{a_N}} \frac{\Gamma(a_N)}{\Gamma(a_0)}$$

3.24. Repeat the previous exercise but now use Bayes' theorem in the form

$$p(t) = \frac{p(t|w,\beta)\, p(w,\beta)}{p(w,\beta|t)}$$

and then substitute for the prior and posterior distributions and the likelihood function in order to derive the result (3.118)

Sol. $p(t) = \dfrac{p(t|w,\beta)\,p(w,\beta)}{p(w,\beta|t)}$

$$p(t|w,\beta) = \cancel{(2\pi)^{N}} \prod_{n=1}^{N} N(t_n|w^T\phi,\beta^{-1}) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}(t-\phi w)^T(t-\phi w)\right\}$$

$$p(w,\beta) = N(w|m_0,\beta^{-1}S_0)\cdot Gam(\beta|a_0 b_0)$$
$$= \frac{\beta^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}}|S_0|^{-\frac{1}{2}}\exp\left\{-\frac{\beta}{2}(w-m_0)^T S_0^{-1}(w-m_0)\right\}\frac{1}{\Gamma(a_0)}b_0^{a_0}\beta^{a_0-1}\exp(-b_0\beta)$$

$$p(w,\beta|t) = N(w|m_N,\beta^{-1}S_N)\,Gam(\beta|a_N,b_N)$$
$$= \frac{\beta^{\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}}|S_N|^{-\frac{1}{2}}\exp\left\{-\frac{\beta}{2}(w-m_N)^T S_N^{-1}(w-m_N)\right\}\frac{1}{\Gamma(a_N)}b_N^{a_N}\beta^{a_N-1}\exp(-b_N\beta)$$

$$\therefore \frac{p(t|w,\beta)p(w,\beta)}{p(w,\beta|t)} = \left(\frac{1}{2\pi}\right)^{\frac{N}{2}}\cdot\frac{\cancel{\frac{\beta^{1/2}}{(2\pi)^{M/2}}}}{\cancel{\frac{\beta^{1/2}}{(2\pi)^{M/2}}}}\cdot\frac{|S_0|^{-\frac{1}{2}}}{|S_N|^{-\frac{1}{2}}}\cdot\frac{\frac{1}{\Gamma(a_0)}\cdot b_0^{a_0}\,\cancel{\beta^{a_0+\frac{N}{2}-1}}\exp\{f_1\}}{\frac{1}{\Gamma(a_N)}b_N^{a_N}\cdot\cancel{\beta^{a_N-1}}\exp\{f_2\}}$$
$$= \left(\frac{1}{2\pi}\right)^{\frac{N}{2}}\cdot\frac{|S_N|^{\frac{1}{2}}}{|S_0|^{\frac{1}{2}}}\cdot\frac{\Gamma(a_N)}{\Gamma(a_0)}\cdot\frac{b_0^{a_0}}{b_N^{a_N}}\cdot\frac{\exp\{f_1\}}{\exp\{f_2\}}$$

$$\frac{\exp\{f_1\}}{\exp\{f_2\}} = \frac{\exp\left\{-\frac{\beta}{2}(t-\phi w)^T(t-\phi w)-\frac{\beta}{2}(w-m_0)^TS_0^{-1}(w-m_0)-b_0\beta\right\}}{\exp\left\{-\frac{\beta}{2}(w-m_N)^T S_N^{-1}(w-m_N)-b_N\right\}}$$

$$-\frac{1}{2}(t-\phi w)^T(t-\phi w)-\frac{1}{2}(w-m_0)^TS_0^{-1}(w-m_0)-b_0$$
$$= -\frac{1}{2}(t^Tt-2w^T\phi^Tt+w^T\phi^T\phi w)-\frac{1}{2}(w^TS_0^{-1}w-2w^TS_0^{-1}m_0+m_0^TS_0^{-1}m_0)-b_0$$
$$= -\frac{1}{2}\left[w^T(\phi^T\phi+S_0^{-1})w-2w^T(\phi^Tt+S_0^{-1}m_0)+t^Tt+m_0^TS_0^{-1}m_0\right]-b_0$$
$$= -\frac{1}{2}\left[w^T(\phi^T\phi+S_0^{-1})w-2w^T(\phi^Tt+S_0^{-1}m_0)\right]-b_0-\frac{1}{2}(t^Tt+m_0^TS_0^{-1}m_0)$$

$\cancel{-\frac{1}{2}(w-m_N)^TS_N^{-1}(w-m_N)}$ because of $S_N^{-1}=\phi^T\phi+S_0^{-1}$, $m_N=S_N[S_0^{-1}m_0+\phi^Tt]$

$b_N=b_0+\frac{1}{2}(m_0^TS_0^{-1}m_0-m_N^TS_N^{-1}m_N+t^Tt)$

So $= -\frac{1}{2}[w^TS_N^{-1}w-2w^TS_N^{-1}m_N]-b_N-\frac{1}{2}m_N^TS_N^{-1}m_N$

$$\because -\frac{1}{2}(w-m_N)^TS_N^{-1}(w-m_N)-b_N$$
$$= -\frac{1}{2}[w^TS_N^{-1}w-2w^TS_N^{-1}m_N+m_N^TS_N^{-1}m_N]-b_N$$

$\therefore \dfrac{\exp\{f_1\}}{\exp\{f_2\}}=1$ $\quad\therefore p(t)=\dfrac{p(t|w,\beta)p(w,\beta)}{p(w,\beta|t)}=\left(\dfrac{1}{2\pi}\right)^{\frac{N}{2}}\dfrac{|S_N|^{\frac{1}{2}}}{|S_0|^{\frac{1}{2}}}\dfrac{\Gamma(a_N)}{\Gamma(a_0)}\cdot\dfrac{b_0^{a_0}}{b_N^{a_N}}$