Bayesian Inference

Shantanu Kodgirwar

January 27, 2021

1 Maximum likelihood estimation

Input vectors are given as $\boldsymbol{x} = (x_1, \dots, x_N)^T$ and the output/target variables as $\boldsymbol{t} = (t_1, \dots, t_N)^T$ and the polynomial coefficients as $\boldsymbol{w} = (w_1, \dots, w_M)^T$.

$$\begin{aligned} t_i &= \sum_{i=1}^{N} y(x_i, \boldsymbol{w}) && (1) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{M} y_k(x_i) w_k && (2) \\ &= (\boldsymbol{X}\boldsymbol{w})_i + n_i && (3) \end{aligned}$$

Assumed distribution of n_i:

$$n_i \sim \mathcal{N}(0, \sigma^2) \tag{4}$$

The values of t, given the values of x follows a Gaussian distribution.

$$t_i \sim \mathcal{N}(y(x_i, \boldsymbol{w}), \sigma^2) \tag{5}$$

The likelihood function is defined as follows.

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \right] \tag{6}$$

A precision parameter β is defined which is given as $\beta^{-1} = \sigma^2$. Thus, we have the following modified likelihood function as follows.

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\left[-\frac{\beta}{2} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \right] \tag{7}$$

Assuming the data is drawn independently, the likelihood is the joint probability given as the product of individual marginal probabilities. It is also assumed the value of β is known or assumed.

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \tag{8}$$

The log likelihood of equation 8 is given as

$$\begin{aligned}
\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &= \sum_{i=1}^{N} \ln p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &(9)\\
&= \sum_{i=1}^{N} \ln \left\{ \sqrt{\frac{\beta}{2\pi}} \exp\left[\frac{-\beta}{2}\left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)^2\right]\right\} &(10)\\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\sum_{i=1}^{N}\left(t_i - (\boldsymbol{X}\boldsymbol{w})_i\right)^2 &(11)\\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\left\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right\|^2 &(12)
\end{aligned}$$

The posterior probability to determine the parameters is given as the product of likelihood function and prior.

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w}) \tag{13}$$

The value of the prior $p(\boldsymbol{w}) = 1$.

By maximizing the negative likelihood (or posterior distribution with prior as one) with respect to $\boldsymbol{w}$,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}}\left\{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\right\} &\overset{!}{=} 0 &(14)\\
\frac{\partial}{\partial \boldsymbol{w}}\left\{-\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta)\right\} &\overset{!}{=} 0 &(15)\\
\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}\left\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right\|^2\right\} &\overset{!}{=} 0 &(16)\\
\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}\left(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right)^T\left(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\right)\right\} &\overset{!}{=} 0 &(17)\\
\beta\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{t}\right) &\overset{!}{=} 0 &(18)
\end{aligned}$$

Therefore, $\boldsymbol{w}_{\text{ML}}$ is evaluated.

$$\boldsymbol{w}_{\text{ML}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{19}$$

It can be seen that the maximum likelihood results into least square estimator. We can similarly estimate β_{ML} by maximizing the posterior with respect to β. The known value of $\boldsymbol{w}_{\text{ML}}$ can now be utilized here.

Taking the log likelihood in equation 9, the following could be shown.

$$\frac{\partial}{\partial \beta}\left\{-\ln p(\boldsymbol{w}_{\text{ML}} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\right\} \overset{!}{=} 0 \tag{20}$$

$$\frac{\partial}{\partial \beta}\left\{-\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}_{\text{ML}}, \beta)\right\} \overset{!}{=} 0 \tag{21}$$

$$\frac{\partial}{\partial \beta}\left\{-\frac{N}{2}\ln \beta + \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{ML}}\|^2\right\} \overset{!}{=} 0 \tag{22}$$

$$\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{ML}}\|^2 - \frac{N}{\beta} \overset{!}{=} 0 \tag{23}$$

Therefore, the value β_{ML} is determined to be

$$\beta_{\text{ML}} = \frac{N}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{ML}}\|^2} \tag{24}$$

2 Maximum a posteriori estimation

In the case of maximum a posteriori (MAP) estimation, the distribution of prior over parameters is known.

2.1 Gaussian distribution of prior

The prior distribution is given as follows

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/\alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2}\|\boldsymbol{w}\|^2\right\} \tag{25}$$

The posterior distribution is shown as follows.

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \alpha) \tag{26}$$

where β as defined earlier is the precision parameter of the likelihood, α is the hyperparameter (also a precision parameter of the prior distribution) which controls the distribution of model parameters. It is assumed that the value of α and β is known.

The log of the posterior is given as follows

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) + \ln p(\boldsymbol{w} \mid \alpha) \tag{27}$$

The log likelihood is known from equation 9. Therefore,

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \left\| \boldsymbol{t} - \boldsymbol{X}\boldsymbol{w} \right\|^2 + \frac{M}{2} \ln \left(\frac{\alpha}{2\pi} \right) - \frac{\alpha}{2} \left\| \boldsymbol{w} \right\|^2 \quad (28)$$

Maximizing the negative log of posterior with respect to $\boldsymbol{w}$.

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ - \ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \right\} \stackrel{!}{=} 0 \quad (29)$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} \left[(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}) \right] + \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} \right\} \stackrel{!}{=} 0 \quad (30)$$

$$\beta \left(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t} \right) + \alpha \boldsymbol{w} \stackrel{!}{=} 0 \quad (31)$$

Let us assign a regularization parameter $\lambda = \alpha/\beta$. Therefore, the value of parameter using maximum a posteriori estimation $\boldsymbol{w}_{\text{MAP}}$ is given as

$$\boldsymbol{w}_{\text{MAP}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{X}^T \boldsymbol{t} \quad (32)$$

2.2 Jeffreys prior

2.2.1 Evaluation for precision parameter β

Jeffreys prior is given as

$$p(\sigma) = \frac{1}{\sigma} \quad (33)$$

Parameter transformation: $\sigma \rightarrow \beta = 1/\sigma^2$

$$\begin{aligned} p_\beta(\beta) &= p_\sigma(\sigma) \Big|_{\sigma = 1/\sqrt{\beta}} \left| \frac{d\sigma}{d\beta} \right| \\ &\propto \sqrt{\beta} \frac{1}{1/\sigma^3} \Big|_{\sigma = 1/\sqrt{\beta}} \\ &= \sqrt{\beta} \sigma^3 \Big|_{\sigma = 1/\sqrt{\beta}} \\ &= \sqrt{\beta} \beta^{-3/2} = 1/\beta \end{aligned}$$

Therefore,

$$p(\beta) = \frac{1}{\beta} \quad (34)$$

The posterior distribution is given as

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\beta) \quad (35)$$

The log likelihood of the posterior distribution is given as

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) + \ln p(\beta) \tag{36}$$

Using equation 9 to give the log likelihood, the above equation would be as follows

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln 2\pi - \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 - \ln\beta \tag{37}$$

Maximizing the log posterior with respect to $\boldsymbol{w}$,

$$\frac{\partial}{\partial \boldsymbol{w}}\{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{38}$$

$$\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2\right\} \stackrel{!}{=} 0 \tag{39}$$

$$\frac{\partial}{\partial \boldsymbol{w}}\left\{\frac{\beta}{2}(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})\right\} \stackrel{!}{=} 0 \tag{40}$$

$$\beta\left(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T\boldsymbol{t}\right) \stackrel{!}{=} 0 \tag{41}$$

Therefore, we find that, $\boldsymbol{w}_{\text{MAP}}$ is

$$\boldsymbol{w}_{\text{MAP}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{42}$$

We can see that $\boldsymbol{w}_{\text{MAP}} = \boldsymbol{w}_{\text{ML}}$ (refer equation 19).
Now, maximizing the log posterior with respect to β, equation 37 is used.

$$\frac{\partial}{\partial \beta}\{-\ln p(\boldsymbol{w}_{\text{MAP}} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{43}$$

$$\frac{\partial}{\partial \beta}\left\{\frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2\right\} - \frac{\partial}{\partial \beta}\left\{\frac{N}{2}\ln\beta\right\} + \frac{\partial}{\partial \beta}\{\ln\beta\} \stackrel{!}{=} 0 \tag{44}$$

$$\frac{1}{2}\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2 - \frac{N}{2\beta} + \frac{1}{\beta} \stackrel{!}{=} 0 \tag{45}$$

Therefore,

$$\beta_{\text{MAP}} = \frac{N - 2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2} \tag{46}$$

2.2.2 Evaluation for hyperparameter α

Jeffreys prior for the hyperparameter α is given as follows,

$$p(\alpha) = \frac{1}{\alpha} \tag{47}$$

The joint distribution $p(w, \alpha)$ is given as follows

$$p(\boldsymbol{w}, \alpha) = p(\boldsymbol{w} \mid \alpha) p(\alpha) \tag{48}$$

The distribution $p(\boldsymbol{w} \mid \alpha)$ is known from equation 25. Therefore, the joint probability $p(w, \alpha)$ is evaluated as follows.

$$\begin{aligned} p(\boldsymbol{w}, \alpha) &= \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2} \|\boldsymbol{w}\|^2\right\} \times \frac{1}{\alpha} & (49) \\ p(\boldsymbol{w}, \alpha) &= \left(\frac{\alpha}{2\pi}\right)^{(M-2)/2} \exp\left\{-\frac{\alpha}{2} \|\boldsymbol{w}\|^2\right\} & (50) \end{aligned}$$

The posterior distribution is given as follows.

$$p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta) p(\boldsymbol{w}, \alpha) p(\beta) \tag{51}$$

The distribution of $p(\beta)$ is known from equation 34 to be $p(\beta) = 1/\beta$. Therefore, the log of the posterior distribution is

$$\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta) + \ln p(\boldsymbol{w}, \alpha) + \ln p(\beta) \tag{52}$$

Evaluating the above equation results to the following.

$$\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) = \frac{N}{2} \ln \beta - \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{(M-2)}{2} \ln \alpha - \frac{\alpha}{2} \|\boldsymbol{w}\|^2 - \ln \beta \tag{53}$$

Initially, maximizing the posterior (refer equation 53) with respect to $\boldsymbol{w}$, we get

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X})\} &\stackrel{!}{=} 0 & (54) \\ \frac{\partial}{\partial \boldsymbol{w}} \left\{\frac{\beta}{2} \left[(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})\right] + \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w}\right\} &\stackrel{!}{=} 0 & (55) \\ \beta \left(\boldsymbol{X}^T \boldsymbol{X} \boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t}\right) + \alpha \boldsymbol{w} &\stackrel{!}{=} 0 & (56) \end{aligned}$$

Given the regularization parameter $\lambda = \alpha/\beta$, $\boldsymbol{w}_{\text{MAP}}$ is the same as per the equation 32.

$$\boldsymbol{w}_{\text{MAP}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{57}$$

Finally, maximizing posterior(refer equation 53) with respect to β, we get the following

$$\begin{aligned} \frac{\partial}{\partial \beta} \{-\ln p(\boldsymbol{w}_{\text{MAP}}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X})\} &\stackrel{!}{=} 0 & (58) \\ \frac{\partial}{\partial \beta} \left\{\frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2\right\} - \frac{\partial}{\partial \beta} \left\{\frac{N}{2} \ln \beta\right\} + \frac{\partial}{\partial \beta} \{\ln \beta\} &\stackrel{!}{=} 0 & (59) \\ \frac{1}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2 - \frac{N}{2\beta} + \frac{1}{\beta} &\stackrel{!}{=} 0 & (60) \end{aligned}$$

Therefore, β_{MAP} is as follows.

$$\beta_{\text{MAP}} = \frac{N-2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2} \tag{61}$$

It can also be seen that $\beta_{\text{MAP}} = \beta_{\text{ML}}$ (refer equation 24).

Maximizing the posterior (refer equation 53) with respect to α, we get the following.

$$\begin{aligned} \frac{\partial}{\partial \alpha} \{-\ln p(\boldsymbol{w}_{\text{MAP}}, \alpha, \beta_{\text{MAP}} \mid \boldsymbol{t}, \boldsymbol{X})\} &\stackrel{!}{=} 0 &(62)\\ \|\boldsymbol{w}_{\text{MAP}}\|^2 - \frac{M-2}{\alpha} &\stackrel{!}{=} 0 &(63) \end{aligned}$$

Therefore, α_{MAP} is given as follows.

$$\alpha_{\text{MAP}} = \frac{M-2}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{64}$$

3 Gamma Distribution

The likelihood function for the target values given β (precision parameter) can be written as

$$\begin{aligned} p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &= \prod_{i=1}^{N} \mathcal{N}(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta^{-1}) &(65)\\ &\propto \beta^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2\right\} &(66) \end{aligned}$$

The corresponding conjugate prior[1] should be proportional to the product of the power of β and the exponential of the linear function of β.

Gamma distribution is a conjugate prior to several likelihood distributions (Gaussian, Poisson, exponential, etc). For example, gamma distribution over precision parameter β can be given as follows.

$$\text{Gam}(\beta \mid a, b) = \frac{b^a}{\Gamma(a)} \beta^{a-1} \exp(-b\beta) \tag{67}$$

where, $\Gamma(a)$ is a gamma function that ensures the gamma distribution is properly normalized, a is called as the shape parameter and b is called as the rate parameter.

[1] posterior distribution is in the same probability distribution family as the prior distribution

The expectation $\mathbb{E}[\beta]$ is given as follows.

$$\mathbb{E}[\beta] = \frac{a}{b} \tag{68}$$

The mode, mode $[\beta]$ which is equivalent to maximizing the posterior with respect to β is given as follows.

$$\text{mode}[\beta] = \frac{a-1}{b} \tag{69}$$

Let us consider the posterior distribution considering the Jeffreys prior for $p(\beta)$ as given in equation 34.

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) &\propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\beta) && (70) \\ &\propto \beta^{\frac{N}{2}} \exp\left\{-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2\right\} \times \frac{1}{\beta} && (71) \\ &\propto \beta^{\frac{N}{2}-1} \exp\left\{-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2\right\} && (72) \end{aligned}$$

By comparing this to the gamma distribution in equation 67, mode $[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}]$ is as follows

$$\text{mode}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}] = \frac{N-2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2} \tag{73}$$

This is equivalent to β_{MAP} from the result in equation 46. Similarly, $\mathbb{E}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}]$ is given as follows.

$$\mathbb{E}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}] = \frac{N}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2} \tag{74}$$

This can also be shown to hold true for the joint distribution $p(\boldsymbol{w}, \alpha)$ in equation 48.

$$p(\boldsymbol{w}, \alpha) \propto \alpha^{\frac{M}{2}-1} \exp\left\{-\frac{\alpha}{2}\|\boldsymbol{w}\|^2\right\} \tag{75}$$

$$\text{mode}[\alpha \mid \boldsymbol{w}] = \frac{M-2}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{76}$$

This is same as α_{MAP} in equation 64. Similarly $\mathbb{E}[\alpha \mid \boldsymbol{w}]$ is

$$\mathbb{E}[\alpha \mid \boldsymbol{w}] = \frac{M}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{77}$$

3.1 Gamma Priors with Jeffreys Priors

Jeffreys priors are given as $p(\alpha) = \alpha^{-1}$ and $p(\beta) = \beta^{-1}$ for hyperparameter α (refer equation 47) and precision parameter β (refer equation 34) respectively.

From the gamma distribution in equation 67, a prior distribution $p(\alpha \mid a_\alpha, b_\alpha)$ can be given as follows.

$$p(\alpha \mid a_\alpha, b_\alpha) = \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp\left(-b_\alpha \alpha\right) \tag{78}$$

Similarly for $p(\beta \mid a_\beta, b_\beta)$ can be given as follows.

$$p(\beta \mid a_\beta, b_\beta) = \frac{b_\beta^{a_\beta}}{\Gamma(a_\beta)} \beta^{a_\beta - 1} \exp\left(-b_\beta \beta\right) \tag{79}$$

In the special case when $a_\alpha = 0, b_\alpha = 0, a_\beta = 0, b_\beta = 0$. The gamma distribution above results to Jeffreys priors $p(\alpha)$ and $p(\beta)$.

The posterior distribution $p(\alpha \mid a_\alpha, b_\alpha, \boldsymbol{w})$ is therefore,

$$\begin{aligned} p(\alpha \mid a_\alpha, b_\alpha, \boldsymbol{w}) &= p(\boldsymbol{w} \mid \alpha) p(\alpha \mid a_\alpha, b_\alpha) && (80) \\ &= \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2} \|\boldsymbol{w}\|^2\right\} \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp\left(-b_\alpha \alpha\right) && (81) \end{aligned}$$

Similarly, the posterior distribution $p(\beta \mid a_\beta, b_\beta, \boldsymbol{w})$

$$\begin{aligned} p(\beta \mid a_\beta, b_\beta, \boldsymbol{w}) &= p(\boldsymbol{w} \mid \beta) p(\beta \mid a_\beta, b_\beta) && (82) \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2\right\} \frac{b_\beta^{a_\beta}}{\Gamma(a_\beta)} \beta^{a_\beta - 1} \exp\left(-b_\beta \beta\right) && (83) \end{aligned}$$

Initially, maximizing log posterior from equation 80 with respect to alpha.

$$\begin{aligned} \frac{\partial}{\partial \alpha} \left\{-\ln p(\alpha \mid a_\alpha, b_\alpha, \boldsymbol{w})\right\} &\overset{!}{=} 0 && (84) \\ \frac{\alpha}{2} \|\boldsymbol{w}\|^2 - \frac{M}{2} \ln \alpha - (a_\alpha - 1) \ln \alpha + b_\alpha \alpha &\overset{!}{=} 0 && (85) \end{aligned}$$

Therefore, α_{MAP} is given as

$$\alpha_{MAP} = \frac{M - 2 + 2a_\alpha}{\|\boldsymbol{w}\|^2 + 2b_\alpha} \tag{86}$$

In the special case, when $a_\alpha = 0$ and $b_\alpha = 0$, the above equation is equal to 64.

Similarly, maximizing log posterior from equation 82 with respect to beta,

$$\begin{aligned} \frac{\partial}{\partial \beta} \left\{-\ln p(\beta \mid a_\beta, b_\beta, \boldsymbol{w})\right\} &\overset{!}{=} 0 && (87) \\ \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 - \frac{N}{2} \ln \beta - \left(a_\beta - 1\right) \ln \beta + b_\beta \beta &\overset{!}{=} 0 && (88) \end{aligned}$$

Therefore, β_{MAP} is given as follows

$$\beta_{MAP} = \frac{N - 2 + 2a_\beta}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 + 2b_\beta} \tag{89}$$

In the special case, $a_\beta = 0$ and $b_\beta = 0$, the above equation is equal to 46.

Under the assumption that $a_\alpha = b_\alpha = a_\beta = b_\beta = \epsilon$, expectation, variance and mode can be given as follows.

$$\mathbb{E}[\alpha] = \frac{a_\alpha}{b_\alpha} = 1 \tag{90}$$

$$\mathrm{var}[\alpha] = \frac{a_\alpha}{b_\alpha^2} = \frac{1}{\epsilon} \tag{91}$$

$$\mathrm{mode}[\alpha] = \frac{a_\alpha - 1}{b_\alpha} = \frac{\epsilon - 1}{\epsilon} \tag{92}$$

The results are similar to $\mathbb{E}[\beta]$, $\mathrm{var}[\beta]$ and $\mathrm{mode}[\beta]$.

If the value of σ is known, a_β and b_β can be estimated.

$$\mathbb{E}[\beta] = \frac{a_\beta}{b_\beta} = \frac{1}{\sigma^2} \tag{93}$$

The $\mathrm{var}[\beta]$ is also known. Let us assume $\mathrm{var}[\beta] = \varepsilon$.

$$\begin{aligned}\mathrm{var}[\beta] &= \mathbb{E}\left[\beta^2\right] - \mathbb{E}\left[\beta\right]^2 &(94)\\ &= \frac{a_\beta}{b_\beta} = \varepsilon &(95)\end{aligned}$$

Therefore this results to the following.

$$a_\beta = \frac{1}{\varepsilon\sigma^4} \tag{96}$$

$$b_\beta = \frac{1}{\varepsilon\sigma^2} \tag{97}$$

4 Laplace distribution

The Laplace distribution is also called as the double exponential distribution because it can be thought of as 2 exponential distributions spliced together back-to-back. A random variable has a Laplace(μ, σ) distribution if its probability density function is given as follows.

$$f(x \mid \mu, \sigma) = \frac{1}{2\sigma} \exp - \frac{|x - \mu|}{\sigma} \tag{98}$$

where, μ is the location parameter and σ as scale parameter.

Under the assumption that every output/target values follows a Laplace distribution. The likelihood $p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \sigma)$ following the Laplace distribution is given as follows.

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \sigma) = \frac{1}{2\sigma} \exp \left\{ - \frac{|t_i - (\boldsymbol{X}\boldsymbol{w})_i|}{\sigma} \right\} \tag{99}$$

Let us define a precision parameter β which is given as follows.

$$\beta = \frac{1}{\sigma} \tag{100}$$

According to the product rule, the total likelihood is given as the product of the individual marginal probabilities as follows.

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \tag{101}$$

The log likelihood of the above equation 101 can be given as follows.

$$\begin{aligned}
\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &= \sum_{i=1}^{N} \ln p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) && (102)\\
&= \sum_{i=1}^{N} \ln \left\{ \frac{\beta}{2} \exp \left(-\beta \left| t_i - (\boldsymbol{X}\boldsymbol{w})_i \right| \right) \right\} && (103)\\
&= \frac{N}{2} \ln \beta - \beta \sum_{i=1}^{N} |t_i - (\boldsymbol{X}\boldsymbol{w})_i| && (104)\\
&= \frac{N}{2} \ln \beta - \beta \, \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\| && (105)
\end{aligned}$$

By maximizing the log likelihood with respect to $\boldsymbol{w}$, the value of $\boldsymbol{w}_{\text{ML}}$ can be found. However, it cannot be treated analytically. Therefore, a suitable optimizer such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm could be used.

After evaluating $\boldsymbol{w}_{\text{ML}}$, β_{ML} could be found out by maximizing the log likelihood with respect to β.

$$\frac{\partial}{\partial \beta} \{ -\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}_{\text{ML}}, \beta) \} \overset{!}{=} 0 \tag{106}$$

$$\frac{\partial}{\partial \beta} \left\{ \beta \, \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{ML}}\| - \frac{N}{2} \ln \beta \right\} \overset{!}{=} 0 \tag{107}$$

Therefore, β_{ML} is given as follows

$$\beta_{\text{ML}} = \frac{N}{2\left\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{ML}}\right\|} \tag{108}$$

If priors are considered, such as Gamma priors or Jeffreys priors, the above equation would be modified.