Bayesian Inference

Shantanu Kodgirwar

February 13, 2021

1 Gaussian Distribution

In the case of a single variable x, the Gaussian/normal distribution can be written in the form

$$\mathcal{N}(x \mid \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1}$$

where, μ is the mean and σ^2 is the variance.

1.1 Maximum likelihood estimation

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

$$\begin{aligned} t_i &= \sum_{i=1}^{N} y(x_i, \boldsymbol{w}) && (2)\\ &= \sum_{i=1}^{N}\sum_{k=1}^{M} y_k(x_i) w_k && (3)\\ &= (\boldsymbol{X}\boldsymbol{w})_i + n_i && (4) \end{aligned}$$

Assumed distribution of n_i:

$$n_i \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

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

$$t_i \sim \mathcal{N}(y(x_i, \boldsymbol{w}), \sigma^2) \tag{6}$$

The likelihood function is defined as follows.

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \Big] \tag{7}$$

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

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \sqrt{\frac{\beta}{2\pi}} \exp\Big[-\frac{\beta}{2} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \Big] \tag{8}$$

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.

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \tag{9}$$

The log likelihood of equation 9 is given as

$$\begin{aligned}
\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &= \sum_{i=1}^{N} \ln p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) && (10)\\
&= \sum_{i=1}^{N} \ln \Big\{ \sqrt{\frac{\beta}{2\pi}} \exp\Big[\frac{-\beta}{2} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \Big] \Big\} && (11)\\
&= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \sum_{i=1}^{N} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 && (12)\\
&= \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 && (13)
\end{aligned}$$

Therefore,

$$\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \propto -\frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 \tag{14}$$

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

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w}) \tag{15}$$

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

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

$$\frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{16}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta)\} \stackrel{!}{=} 0 \tag{17}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 \right\} \stackrel{!}{=} 0 \tag{18}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}) \right\} \stackrel{!}{=} 0 \tag{19}$$

$$\beta \left(\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t} \right) \stackrel{!}{=} 0 \tag{20}$$

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

$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{21}$$

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 $\boldsymbol{w}_{\mathrm{ML}}$ can now be utilized here.

Taking the log likelihood in equation 10, the following could be shown.

$$\frac{\partial}{\partial \beta} \{-\ln p(\boldsymbol{w}_{\mathrm{ML}} \mid \boldsymbol{t}, \boldsymbol{X}, \beta)\} \stackrel{!}{=} 0 \tag{22}$$

$$\frac{\partial}{\partial \beta} \{-\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}_{\mathrm{ML}}, \beta)\} \stackrel{!}{=} 0 \tag{23}$$

$$\frac{\partial}{\partial \beta} \left\{ -\frac{N}{2} \ln \beta + \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\mathrm{ML}}\|^2 \right\} \stackrel{!}{=} 0 \tag{24}$$

$$\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\mathrm{ML}}\|^2 - \frac{N}{\beta} \stackrel{!}{=} 0 \tag{25}$$

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

$$\beta_{\mathrm{ML}} = \frac{N}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\mathrm{ML}}\|^2} \tag{26}$$

1.2 Maximum a posteriori estimation

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

1.2.1 Conjugate Gaussian prior

The prior distribution is given as follows

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}/\alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2} \|\boldsymbol{w}\|^2\right\} \tag{27}$$

The posterior distribution is shown as follows.

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \alpha) \tag{28}$$

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

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) + \ln p(\boldsymbol{w} \mid \alpha) \tag{29}$$

The log likelihood is known from equation 10. Therefore,

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{M}{2} \ln\left(\frac{\alpha}{2\pi}\right) - \frac{\alpha}{2} \|\boldsymbol{w}\|^2 \tag{30}$$

Therefore,

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto -\frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 - \frac{\alpha}{2} \|\boldsymbol{w}\|^2 \tag{31}$$

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

$$\frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)\} \stackrel{!}{=} 0 \tag{32}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{\frac{\beta}{2} \left[(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})\right] + \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w}\right\} \stackrel{!}{=} 0 \tag{33}$$

$$\beta\left(\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t}\right) + \alpha \boldsymbol{w} \stackrel{!}{=} 0 \tag{34}$$

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

$$\boldsymbol{w}_{\text{MAP}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{35}$$

1.2.2 Jeffreys prior

Jeffreys prior is given as

$$p(\sigma) = \frac{1}{\sigma} \tag{36}$$

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

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

Therefore,

$$p(\beta) = \frac{1}{\beta} \tag{37}$$

Evaluation for precision parameter β

The posterior distribution is given as

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\beta) \tag{38}$$

The log likelihood of the posterior distribution is given as

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) + \ln p(\beta) \tag{39}$$

Using equation 10 to give the log likelihood, the above equation would be as follows

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln 2\pi - \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 - \ln \beta \tag{40}$$

For MAP estimation on $\boldsymbol{w}$, $\boldsymbol{w}_{\text{MAP}} = \boldsymbol{w}_{\text{ML}}$ (refer equation 21) because only a Gaussian likelihood is used for parameter distribution.

Maximizing the log posterior with respect to β, equation 40 is used.

$$\frac{\partial}{\partial \beta} \{ - \ln p(\boldsymbol{w}_{\text{MAP}} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) \} \stackrel{!}{=} 0 \tag{41}$$

$$\frac{\partial}{\partial \beta} \left\{ \frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2 \right\} - \frac{\partial}{\partial \beta} \left\{ \frac{N}{2} \ln \beta \right\} + \frac{\partial}{\partial \beta} \{ \ln \beta \} \stackrel{!}{=} 0 \tag{42}$$

$$\frac{1}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2 - \frac{N}{2\beta} + \frac{1}{\beta} \stackrel{!}{=} 0 \tag{43}$$

Therefore,

$$\beta_{\text{MAP}} = \frac{N-2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}}\|^2} \tag{44}$$

1.3 Jeffreys hyperprior with conjugate Gaussian prior

Jeffreys prior for the hyperparameter α is given as follows,

$$p(\alpha) = \frac{1}{\alpha} \tag{45}$$

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

$$p(\boldsymbol{w}, \alpha) = p(\boldsymbol{w} \mid \alpha)p(\alpha) \tag{46}$$

The distribution $p(\boldsymbol{w} \mid \alpha)$ is known from equation 27. Therefore, the joint probability $p(w, \alpha)$, i.e, Jeffreys hyperprior over hyperparameter α and Gaussian prior over parameter $\boldsymbol{w}$ is evaluated as follows.

$$p(\boldsymbol{w}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2}\|\boldsymbol{w}\|^2\right\} \times \frac{1}{\alpha} \tag{47}$$

$$p(\boldsymbol{w}, \alpha) = \left(\frac{\alpha}{2\pi}\right)^{(M-2)/2} \exp\left\{-\frac{\alpha}{2}\|\boldsymbol{w}\|^2\right\} \tag{48}$$

Evaluation for hyperparameter $\boldsymbol{\alpha}$, precision parameter $\boldsymbol{\beta}$, parameter $\boldsymbol{w}$

The posterior distribution (Conjugate Gaussian prior and Jeffreys prior combined) is given as follows.

$$p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta)p(\boldsymbol{w}, \alpha)p(\beta) \tag{49}$$

The distribution of $p(\beta)$ is known from equation 37 to be $p(\beta) = 1/\beta$. Therefore, the log of the posterior distribution is

$$\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) = \ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \alpha, \beta) + \ln p(\boldsymbol{w}, \alpha) + \ln p(\beta) \tag{50}$$

Evaluating the above equation results to the following.

$$\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X}) = \frac{N}{2}\ln\beta - \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 + \frac{(M-2)}{2}\ln\alpha - \frac{\alpha}{2}\|\boldsymbol{w}\|^2 - \ln\beta \tag{51}$$

Initially, maximizing the posterior (refer equation 51) with respect to $\boldsymbol{w}$, we get

$$\frac{\partial}{\partial \boldsymbol{w}} \{-\ln p(\boldsymbol{w}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X})\} \stackrel{!}{=} 0 \tag{52}$$

$$\frac{\partial}{\partial \boldsymbol{w}} \left\{ \frac{\beta}{2} \left[(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T (\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}) \right] + \frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} \right\} \stackrel{!}{=} 0 \tag{53}$$

$$\beta \left(\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^T \boldsymbol{t} \right) + \alpha \boldsymbol{w} \stackrel{!}{=} 0 \tag{54}$$

Given the regularization parameter $\lambda = \alpha/\beta$, $\boldsymbol{w}_{\text{MAP}}$ is the same as per the equation 35.

$$\boldsymbol{w}_{\text{MAP}} = \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{X}^T \boldsymbol{t} \tag{55}$$

Finally, maximizing posterior(refer equation 51) with respect to β, we get the following

$$\frac{\partial}{\partial \beta} \{-\ln p(\boldsymbol{w}_{\text{MAP}}, \alpha, \beta \mid \boldsymbol{t}, \boldsymbol{X})\} \stackrel{!}{=} 0 \tag{56}$$

$$\frac{\partial}{\partial \beta} \left\{ \frac{\beta}{2} \left\| \boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}} \right\|^2 \right\} - \frac{\partial}{\partial \beta} \left\{ \frac{N}{2} \ln \beta \right\} + \frac{\partial}{\partial \beta} \{\ln \beta\} \stackrel{!}{=} 0 \tag{57}$$

$$\frac{1}{2} \left\| \boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}} \right\|^2 - \frac{N}{2\beta} + \frac{1}{\beta} \stackrel{!}{=} 0 \tag{58}$$

Therefore, β_{MAP} is as follows.

$$\beta_{\text{MAP}} = \frac{N - 2}{\left\| \boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{MAP}} \right\|^2} \tag{59}$$

It can also be seen that $\beta_{\text{MAP}} = \beta_{\text{ML}}$ (refer equation 26).

Maximizing the posterior (refer equation 51) with respect to α, we get the following.

$$\frac{\partial}{\partial \alpha} \{-\ln p(\boldsymbol{w}_{\text{MAP}}, \alpha, \beta_{\text{MAP}} \mid \boldsymbol{t}, \boldsymbol{X})\} \stackrel{!}{=} 0 \tag{60}$$

$$\left\| \boldsymbol{w}_{\text{MAP}} \right\|^2 - \frac{M - 2}{\alpha} \stackrel{!}{=} 0 \tag{61}$$

Therefore, α_{MAP} is given as follows.

$$\alpha_{\text{MAP}} = \frac{M - 2}{\left\| \boldsymbol{w}_{\text{MAP}} \right\|^2} \tag{62}$$

2 Multivariate Gaussian Distribution

For a N-dimensional vector $\boldsymbol{x} = (x_1, \ldots, x_N)^T$, the multivariate Gaussian distribution takes the following form

$$\mathcal{N}(x \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{N/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{63}$$

where $\boldsymbol{\mu}$ is an N-dimensional mean vector, $\boldsymbol{\Sigma}$ is an $N \times N$ covariance matrix and $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

2.1 Conjugate Bayesian analysis of the Gaussian distribution

In the conjugate Bayesian analysis, the prior and likelihood distribution is the same. It can be shown that the resulting posterior is a multivariate Gaussian when the prior and likelihood are a Gaussian distribution.

Rewriting the posterior distribution as proportional to likelihood and prior.

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \alpha) \tag{64}$$

Therefore, it can be shown that the $\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)$ for a Gaussian distribution as

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto -\frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 - \frac{\alpha}{2}\|\boldsymbol{w}\|^2 \tag{65}$$

Evaluating this further,

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto -\frac{\beta}{2}(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w})^T(\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}) - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{66}$$

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto -\frac{1}{2}\boldsymbol{w}^T(\underbrace{\beta\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I}}_{\boldsymbol{\Sigma}^{-1}})\boldsymbol{w} + \boldsymbol{w}^T\underbrace{\beta\boldsymbol{X}^T\boldsymbol{t}}_{\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} \tag{67}$$

Therefore, $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ is given as

$$\boldsymbol{\Sigma} = (\beta\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1} \tag{68}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\beta\boldsymbol{X}^T\boldsymbol{t} \tag{69}$$

$$\boldsymbol{\mu} = (\beta\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}\beta\boldsymbol{X}^T\boldsymbol{t} \tag{70}$$

$$\boldsymbol{\mu} = (\beta\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}\beta\boldsymbol{X}^T\boldsymbol{t} \tag{71}$$

$$\boldsymbol{\mu} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{t} \tag{72}$$

$$\boldsymbol{\mu} \equiv w_{\text{MAP}} \tag{73}$$

The posterior $\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta)$ is further given as follows

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\mu \tag{74}$$

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto -\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w} + \boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\mu + \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu - \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu \tag{75}$$

$$\ln p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto -\frac{1}{2}(\boldsymbol{w}-\mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\mu) + \frac{1}{2}\mu^T\boldsymbol{\Sigma}^{-1}\mu \tag{76}$$

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{w}-\mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\mu)\right\} \tag{77}$$

The posterior over the regression parameters $w = (w_1, w_2, ..., w_M)^T$ is a multivariate Gaussian.

$$p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \alpha, \beta) = \frac{1}{|2\pi|^{(M+1)/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{w}-\mu)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\mu)\right\} \tag{78}$$

where $\boldsymbol{\Sigma}$ is an $(M+1)\times(M+1)$ covariance matrix, given as $\boldsymbol{\Sigma} = (\beta\boldsymbol{X}^T\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}$, $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$. and μ is an $(M+1)$-dimensional mean vector given as $\mu = \boldsymbol{\Sigma}\beta\boldsymbol{X}^T\boldsymbol{t}$.

3 Gamma Distribution

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

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N}\mathcal{N}(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta^{-1}) \tag{79}$$

$$\propto \beta^{\frac{N}{2}}\exp\left\{-\frac{\beta}{2}\sum_{i=1}^{N}(t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2\right\} \tag{80}$$

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.

$$\text{Gam}(\beta \mid a, b) = \frac{b^a}{\Gamma(a)}\beta^{a-1}\exp(-b\beta) \tag{81}$$

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

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.

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

$$\mathbb{E}[\beta] = \frac{a}{b} \tag{82}$$

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

$$\text{mode}[\beta] = \frac{a-1}{b} \tag{83}$$

3.1 Jeffreys Priors

Let us consider the posterior distribution considering the Jeffreys prior for $p(\beta)$ as given in equation 37.

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{t}, \boldsymbol{X}, \beta) &\propto p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) p(\beta) && (84) \\ &\propto \beta^{\frac{N}{2}} \exp\left\{ -\frac{\beta}{2} \sum_{i=1}^{N} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \right\} \times \frac{1}{\beta} && (85) \\ &\propto \beta^{\frac{N}{2}-1} \exp\left\{ -\frac{\beta}{2} \sum_{i=1}^{N} (t_i - (\boldsymbol{X}\boldsymbol{w})_i)^2 \right\} && (86) \end{aligned}$$

By comparing this to the gamma distribution in equation 81, mode$[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}]$ is as follows

$$\text{mode}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}] = \frac{N-2}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2} \tag{87}$$

This is equivalent to β_{MAP} from the result in equation 44. Similarly, $\mathbb{E}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}]$ is given as follows.

$$\mathbb{E}[\beta \mid \boldsymbol{t}, \boldsymbol{X}, \boldsymbol{w}] = \frac{N}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2} \tag{88}$$

This can also be shown to hold true for the joint distribution $p(\boldsymbol{w}, \alpha)$ in equation 46.

$$p(\boldsymbol{w}, \alpha) \propto \alpha^{\frac{M}{2}-1} \exp\left\{ -\frac{\alpha}{2} \|\boldsymbol{w}\|^2 \right\} \tag{89}$$

$$\text{mode}[\alpha \mid \boldsymbol{w}] = \frac{M-2}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{90}$$

This is same as α_{MAP} in equation 62. Similarly $\mathbb{E}[\alpha \mid \boldsymbol{w}]$ is

$$\mathbb{E}[\alpha \mid \boldsymbol{w}] = \frac{M}{\|\boldsymbol{w}_{\text{MAP}}\|^2} \tag{91}$$

3.2 Gamma Priors with Jeffreys Priors

Jeffreys priors are given as $p(\alpha) = \alpha^{-1}$ and $p(\beta) = \beta^{-1}$ for hyperparameter α (refer equation 45) and precision parameter β (refer equation 37) respectively.

From the gamma distribution in equation 81, a prior distribution $p(\alpha \mid a_\alpha, b_\alpha)$ can be given as follows.

$$p(\alpha \mid a_\alpha, b_\alpha) = \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha) \tag{92}$$

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

$$p(\beta \mid a_\beta, b_\beta) = \frac{b_\beta^{a_\beta}}{\Gamma(a_\beta)} \beta^{a_\beta - 1} \exp(-b_\beta \beta) \tag{93}$$

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,

$$\begin{aligned} p(\alpha \mid a_\alpha, b_\alpha, \boldsymbol{w}) &= p(\boldsymbol{w} \mid \alpha) p(\alpha \mid a_\alpha, b_\alpha) && (94) \\ &= \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left\{-\frac{\alpha}{2} \|\boldsymbol{w}\|^2\right\} \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha - 1} \exp(-b_\alpha \alpha) && (95) \end{aligned}$$

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

$$\begin{aligned} p(\beta \mid a_\beta, b_\beta, \boldsymbol{w}) &= p(\boldsymbol{w} \mid \beta) p(\beta \mid a_\beta, b_\beta) && (96) \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left\{-\frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2\right\} \frac{b_\beta^{a_\beta}}{\Gamma(a_\beta)} \beta^{a_\beta - 1} \exp(-b_\beta \beta) && (97) \end{aligned}$$

Initially, maximizing log posterior from equation 94 with respect to alpha.

$$\frac{\partial}{\partial \alpha} \{-\ln p(\alpha \mid a_\alpha, b_\alpha, \boldsymbol{w})\} \stackrel{!}{=} 0 \tag{98}$$

$$\frac{\alpha}{2} \|\boldsymbol{w}\|^2 - \frac{M}{2} \ln \alpha - (a_\alpha - 1) \ln \alpha + b_\alpha \alpha \stackrel{!}{=} 0 \tag{99}$$

Therefore, α_{MAP} is given as

$$\alpha_{MAP} = \frac{M - 2 + 2a_\alpha}{\|\boldsymbol{w}\|^2 + 2b_\alpha} \tag{100}$$

In the special case, when $a_\alpha = 0$ and $b_\alpha = 0$, the above equation is equal to 62.

Similarly, maximizing log posterior from equation 96 with respect to beta,

$$\frac{\partial}{\partial \beta} \{-\ln p(\beta \mid a_\beta, b_\beta, \boldsymbol{w})\} \stackrel{!}{=} 0 \tag{101}$$

$$\frac{\beta}{2} \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 - \frac{N}{2} \ln \beta - \left(a_\beta - 1\right) \ln \beta + b_\beta \beta \stackrel{!}{=} 0 \tag{102}$$

Therefore, β_{MAP} is given as follows

$$\beta_{MAP} = \frac{N - 2 + 2a_\beta}{\|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\|^2 + 2b_\beta} \tag{103}$$

In the special case, $a_\beta = 0$ and $b_\beta = 0$, the above equation is equal to 44.

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

$$\mathbb{E}[\alpha] = \frac{a_\alpha}{b_\alpha} = 1 \tag{104}$$

$$\text{var}[\alpha] = \frac{a_\alpha}{b_\alpha^2} = \frac{1}{\epsilon} \tag{105}$$

$$\text{mode}[\alpha] = \frac{a_\alpha - 1}{b_\alpha} = \frac{\epsilon - 1}{\epsilon} \tag{106}$$

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

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

$$\mathbb{E}[\beta] = \frac{a_\beta}{b_\beta} = \frac{1}{\sigma^2} \tag{107}$$

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

$$\text{var}[\beta] = \mathbb{E}\left[\beta^2\right] - \mathbb{E}\left[\beta\right]^2 \tag{108}$$

$$= \frac{a_\beta}{b_\beta} = \varepsilon \tag{109}$$

Therefore this results to the following.

$$a_\beta = \frac{1}{\varepsilon\sigma^4} \tag{110}$$

$$b_\beta = \frac{1}{\varepsilon\sigma^2} \tag{111}$$

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.

$$f(x \mid \mu, \sigma) = \frac{1}{2\sigma} \exp -\frac{|x - \mu|}{\sigma} \tag{112}$$

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.

$$p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \sigma) = \frac{1}{2\sigma} \exp \left\{ -\frac{|t_i - (\boldsymbol{X}\boldsymbol{w})_i|}{\sigma} \right\} \tag{113}$$

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

$$\beta = \frac{1}{\sigma} \tag{114}$$

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

$$p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{i=1}^{N} p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) \tag{115}$$

The log likelihood of the above equation 115 can be given as follows.

$$\begin{aligned}
\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}, \beta) &= \sum_{i=1}^{N} \ln p(t_i \mid \boldsymbol{X}, \boldsymbol{w}, \beta) && (116)\\
&= \sum_{i=1}^{N} \ln \left\{ \frac{\beta}{2} \exp \left(-\beta \, |t_i - (\boldsymbol{X}\boldsymbol{w})_i| \right) \right\} && (117)\\
&= \frac{N}{2} \ln \beta - \beta \sum_{i=1}^{N} |t_i - (\boldsymbol{X}\boldsymbol{w})_i| && (118)\\
&= \frac{N}{2} \ln \beta - \beta \, \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}\| && (119)
\end{aligned}$$

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.

$$\begin{aligned}
\frac{\partial}{\partial \beta} \{ -\ln p(\boldsymbol{t} \mid \boldsymbol{X}, \boldsymbol{w}_{\text{ML}}, \beta) \} &\stackrel{!}{=} 0 && (120)\\
\frac{\partial}{\partial \beta} \left\{ \beta \, \|\boldsymbol{t} - \boldsymbol{X}\boldsymbol{w}_{\text{ML}}\| - \frac{N}{2} \ln \beta \right\} &\stackrel{!}{=} 0 && (121)
\end{aligned}$$

Therefore, β_{ML} is given as follows

$$\beta_{\text{ML}} = \frac{N}{2 \left\| \boldsymbol{t} - \boldsymbol{X} \boldsymbol{w}_{\text{ML}} \right\|} \tag{122}$$

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