

# Notes on Bayesian Logistic Regression

Max Leung

May 7, 2023

## 1 Laplace Approximation

$$\begin{aligned} \ln f(\mathbf{z}) &\approx \ln f(\mathbf{z}_0) + \overbrace{\nabla \ln f(\mathbf{z}_0)'}^{\mathbf{o}'} (\mathbf{z} - \mathbf{z}_0) + \frac{1}{2} (\mathbf{z} - \mathbf{z}_0)' \overbrace{\nabla \nabla \ln f(\mathbf{z}_0)}^{-\mathbf{A}} (\mathbf{z} - \mathbf{z}_0) \\ &= \ln f(\mathbf{z}_0) - \frac{1}{2} (\mathbf{z} - \mathbf{z}_0)' \mathbf{A} (\mathbf{z} - \mathbf{z}_0) \end{aligned}$$

$$\begin{aligned} f(\mathbf{z}) &\approx \exp(\ln f(\mathbf{z}_0) - \frac{1}{2} (\mathbf{z} - \mathbf{z}_0)' \mathbf{A} (\mathbf{z} - \mathbf{z}_0)) \\ &= \exp(\ln f(\mathbf{z}_0)) \exp(-\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)' \mathbf{A} (\mathbf{z} - \mathbf{z}_0)) \\ &= f(\mathbf{z}_0) \exp(-\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)' \mathbf{A} (\mathbf{z} - \mathbf{z}_0)) \end{aligned}$$

If we approximate  $f(\cdot)$  by  $N(\mathbf{z}_0, \mathbf{A}^{-1})$ , we have

$$\begin{aligned} &\approx \frac{1}{(2\pi)^{M/2} |\mathbf{A}|^{-1/2}} \underbrace{\exp\{-\frac{1}{2} (\mathbf{z}_0 - \mathbf{z}_0)' \mathbf{A} (\mathbf{z}_0 - \mathbf{z}_0)\}}_1 \exp(-\frac{1}{2} (\mathbf{z} - \mathbf{z}_0)' \mathbf{A} (\mathbf{z} - \mathbf{z}_0)) \\ &= N(\mathbf{z} | \mathbf{z}_0, \mathbf{A}^{-1}) = q(\mathbf{z}) \end{aligned}$$

where  $\mathbf{z}_0 = \arg \max_{\mathbf{z}} \ln f(\mathbf{z})$  and  $\mathbf{A} = -\nabla \nabla \ln f(\mathbf{z}_0)$

## 2 Bayesian Logistic Regression (Laplace Approximation Approach)

### 2.1 Posterior Distribution of Parameters

Assume we have prior density  $p(\mathbf{w} | \mathbf{X}) = p(\mathbf{w}) = N(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)$ . Likelihood function is  $p(\mathbf{t} | \mathbf{w}, \mathbf{X}) = \prod_{n=1}^N p(C_1 | \mathbf{x}_n; \mathbf{w})^{t_n} (1 - p(C_1 | \mathbf{x}_n; \mathbf{w}))^{1-t_n}$  where  $p(C_1 | \mathbf{x}_n; \mathbf{w}) = \sigma(\mathbf{w}' \mathbf{x}_n)$ . Posterior density is

$$\begin{aligned} p(\mathbf{w} | \mathbf{t}, \mathbf{X}) &\propto p(\mathbf{t} | \mathbf{w}, \mathbf{X}) p(\mathbf{w} | \mathbf{X}) \\ &= \prod_{n=1}^N \sigma(\mathbf{w}' \mathbf{x}_n)^{t_n} (1 - \sigma(\mathbf{w}' \mathbf{x}_n))^{1-t_n} N(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0) \end{aligned}$$

which is not a well known joint density function

$$\begin{aligned} \ln p(\mathbf{w} | \mathbf{t}, \mathbf{X}) &= \ln \left[ \prod_{n=1}^N \sigma(\mathbf{w}' \mathbf{x}_n)^{t_n} (1 - \sigma(\mathbf{w}' \mathbf{x}_n))^{1-t_n} N(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0) \right] \\ &= \sum_{n=1}^N [t_n \ln \sigma(\mathbf{w}' \mathbf{x}_n) + (1 - t_n) \ln (1 - \sigma(\mathbf{w}' \mathbf{x}_n))] + \ln [N(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)] \\ &= \sum_{n=1}^N [t_n \ln \sigma(\mathbf{w}' \mathbf{x}_n) + (1 - t_n) \ln (1 - \sigma(\mathbf{w}' \mathbf{x}_n))] + \ln \left[ \frac{1}{(2\pi)^{D/2} |\mathbf{S}_0|^{1/2}} \exp\left\{-\frac{1}{2} (\mathbf{w} - \mathbf{m}_0)' \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0)\right\} \right] \\ &= \sum_{n=1}^N [t_n \ln \sigma(\mathbf{w}' \mathbf{x}_n) + (1 - t_n) \ln (1 - \sigma(\mathbf{w}' \mathbf{x}_n))] + \ln \left[ \frac{1}{(2\pi)^{D/2} |\mathbf{S}_0|^{1/2}} \right] - \frac{1}{2} (\mathbf{w} - \mathbf{m}_0)' \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \end{aligned}$$

We can approximate  $p(\mathbf{w}|\mathbf{t}, \mathbf{X})$  by Laplace Approximation. As a result, our posterior follows multivariate normal distribution.

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) \approx q(\mathbf{w}) = N(\mathbf{w}|\mathbf{w}_{MAP}, \mathbf{S}^{-1})$$

where  $\mathbf{w}_{MAP} = \arg \max_{\mathbf{w}} \ln p(\mathbf{w}|\mathbf{t}, \mathbf{X})$

$$\begin{aligned} \frac{\partial \ln p(\mathbf{w}|\mathbf{t}, \mathbf{X})}{\partial \mathbf{w}} \Big|_{\mathbf{w}_{MAP}} &= \mathbf{0} \\ \frac{\partial \sum_{n=1}^N [t_n \ln \sigma(\mathbf{w}' \mathbf{x}_n) + (1 - t_n) \ln (1 - \sigma(\mathbf{w}' \mathbf{x}_n))] + \ln \left[ \frac{1}{(2\pi)^{D/2} |\mathbf{S}_0|^{1/2}} \right] - \frac{1}{2} (\mathbf{w} - \mathbf{m}_0)' \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0)}{\partial \mathbf{w}} \Big|_{\mathbf{w}_{MAP}} &= \mathbf{0} \\ \mathbf{X}'(\mathbf{t} - \mathbf{p}) - \mathbf{S}_0^{-1}(\mathbf{w}_{MAP} - \mathbf{m}_0) &= \mathbf{0} \end{aligned}$$

where  $\mathbf{p} = (\sigma(\mathbf{w}'_{MAP} \mathbf{x}_1), \dots, \sigma(\mathbf{w}'_{MAP} \mathbf{x}_D))'$

There is no closed form solution for  $\mathbf{w}_{MAP}$

$$\begin{aligned} \mathbf{S} &= -\nabla \nabla \ln p(\mathbf{w}_{MAP}|\mathbf{t}) \\ &= -\nabla \nabla \left\{ \sum_{n=1}^N [t_n \ln \sigma(\mathbf{w}' \mathbf{x}_n) + (1 - t_n) \ln (1 - \sigma(\mathbf{w}' \mathbf{x}_n))] + \ln \left[ \frac{1}{(2\pi)^{D/2} |\mathbf{S}_0|^{1/2}} \right] - \frac{1}{2} (\mathbf{w} - \mathbf{m}_0)' \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0) \right\} \\ &= -(-\mathbf{X}' \mathbf{W} \mathbf{X} - \mathbf{S}_0^{-1}) \\ &= \mathbf{X}' \mathbf{W} \mathbf{X} + \mathbf{S}_0^{-1} \end{aligned}$$

where  $(\mathbf{W})_{ii} = \sigma(\mathbf{w}'_{MAP} \mathbf{x}_i)(1 - \sigma(\mathbf{w}'_{MAP} \mathbf{x}_i))$  and  $(\mathbf{W})_{ij} = 0$  for  $i \neq j$

## 2.2 Special Case

if  $\mathbf{m}_0$  is chosen to be  $\mathbf{w}_{MAP}$  then

$$\begin{aligned} \mathbf{X}'(\mathbf{t} - \mathbf{p}) - \mathbf{S}_0^{-1}(\mathbf{w}_{MAP} - \mathbf{w}_{MAP}) &= \mathbf{0} \\ \mathbf{X}'(\mathbf{t} - \mathbf{p}) &= \mathbf{0} \end{aligned} \quad \text{same as FOC of MLE}$$

Thus,  $\mathbf{w}_{MAP} = \mathbf{w}_{MLE}$  in such case, which can be found by Iterated Reweighted Least Squares (IRLS) algorithm.

Additionally, if  $\mathbf{S}_0$  is chosen to be close to  $\mathbf{O}^{-1}$ . We have

$$\begin{aligned} \mathbf{S} &\approx \mathbf{X}' \mathbf{W} \mathbf{X} + (\mathbf{O}^{-1})^{-1} \\ &= \mathbf{X}' \mathbf{W} \mathbf{X} \end{aligned}$$

Thus,

$$\mathbf{S}^{-1} = (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} = -(-\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} = -(\nabla \nabla p(\mathbf{t}|\mathbf{w}_{MLE}, \mathbf{X}))^{-1} = \underbrace{(-\nabla \nabla p(\mathbf{t}|\mathbf{w}_{MLE}, \mathbf{X}))^{-1}}_{\mathbf{I}(\mathbf{w}_{MLE})}$$

$\mathbf{I}(\mathbf{w}_{MLE})^{-1}$  is the asymptotic variance of  $\sqrt{D}(\mathbf{w}_{MLE} - \mathbf{w}_{TRUE})$

Thus, we have  $p(\mathbf{w}|\mathbf{t}, \mathbf{X}) \approx N(\mathbf{w}|\mathbf{w}_{MLE}, \mathbf{I}(\mathbf{w}_{MLE})^{-1})$

## 2.3 Predictive Distribution

Let  $\tilde{T}$  be the predicted target / dependent variable. Predictive distribution is

$$\begin{aligned} p(\tilde{T} = 1|\mathbf{t}, \mathbf{X}) &= \int p(\tilde{T} = 1, \mathbf{w}|\mathbf{t}, \mathbf{X}) \partial \mathbf{w} \\ &= \int p(\tilde{T} = 1|\mathbf{w}, \mathbf{t}, \mathbf{X}) p(\mathbf{w}|\mathbf{t}, \mathbf{X}) \partial \mathbf{w} \\ &\approx \int \sigma(\mathbf{w}' \mathbf{x}) q(\mathbf{w}) \partial \mathbf{w} \end{aligned}$$

Note that  $\int \delta(a - \mathbf{w}'\mathbf{x})\sigma(a)da = \delta(\mathbf{w}'\mathbf{x} - \mathbf{w}'\mathbf{x})\sigma(\mathbf{w}'\mathbf{x}) = 1 \cdot \sigma(\mathbf{w}'\mathbf{x})$  as  $\delta(0) = 1$  and  $\delta(a) = 0$  for  $\forall a \neq 0$

$$\begin{aligned}
&= \int \int \delta(a - \mathbf{w}'\mathbf{x})\sigma(a)da \, q(\mathbf{w})\partial\mathbf{w} \\
&= \int \int \delta(a - \mathbf{w}'\mathbf{x})\sigma(a)q(\mathbf{w})\partial\mathbf{w}da \\
&= \int \sigma(a) \underbrace{\int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})\partial\mathbf{w}}_{p(a)} da
\end{aligned}$$

$p(a)$ 's moments can be evaluated as

$$\begin{aligned}
\mathbb{E}_p(a) &= \int p(a)a \, da \\
&= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})\partial\mathbf{w}a \, da \\
&= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})a \, da\partial\mathbf{w} \\
&= \int \int \delta(a - \mathbf{w}'\mathbf{x})a \, da \, q(\mathbf{w})\partial\mathbf{w} \\
&= \int \delta(\mathbf{w}'\mathbf{x} - \mathbf{w}'\mathbf{x})\mathbf{w}'\mathbf{x}q(\mathbf{w})\partial\mathbf{w} \\
&= \int \mathbf{w}'\mathbf{x}q(\mathbf{w})\partial\mathbf{w} \\
&= \mathbb{E}_q(\mathbf{w}'\mathbf{x}) \\
&= \mathbb{E}_q(\mathbf{w})'\mathbf{x}
\end{aligned}$$

$$\begin{aligned}
Var_p(a) &= \mathbb{E}_p(a^2) - \mathbb{E}_p(a)^2 \\
&= \mathbb{E}_p(a^2 - \mathbb{E}_p(a)^2) \\
&= \int p(a)(a^2 - \mathbb{E}_p(a)^2)da \\
&= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})\partial\mathbf{w}(a^2 - \mathbb{E}_p(a)^2)da \\
&= \int \int \delta(a - \mathbf{w}'\mathbf{x})q(\mathbf{w})(a^2 - \mathbb{E}_p(a)^2)da\partial\mathbf{w} \\
&= \int q(\mathbf{w}) \int \delta(a - \mathbf{w}'\mathbf{x})(a^2 - \mathbb{E}_p(a)^2)da\partial\mathbf{w} \\
&= \int q(\mathbf{w})\delta(\mathbf{w}'\mathbf{x} - \mathbf{w}'\mathbf{x})((\mathbf{w}'\mathbf{x})^2 - \mathbb{E}_p(\mathbf{w}'\mathbf{x})^2)\partial\mathbf{w} \\
&= \int q(\mathbf{w})((\mathbf{w}'\mathbf{x})^2 - \mathbb{E}_p(\mathbf{w}'\mathbf{x})^2)\partial\mathbf{w} \\
&\approx \int q(\mathbf{w})((\mathbf{w}'\mathbf{x})^2 - \mathbb{E}_q(\mathbf{w}'\mathbf{x})^2)\partial\mathbf{w} \\
&= \mathbb{E}_q((\mathbf{w}'\mathbf{x})^2 - \mathbb{E}_q(\mathbf{w}'\mathbf{x})^2) \\
&= Var_q(\mathbf{x}'\mathbf{w}) \\
&= \mathbf{x}'Var_q(\mathbf{w})\mathbf{x}
\end{aligned}$$

### 3 EM Algorithm

$$\ln[L(\boldsymbol{\theta})] = \ln[p(\mathbf{X}|\boldsymbol{\theta})] = \ln\left[\sum_{\mathbf{z}} p(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})\right] = \ln\left[\sum_{\mathbf{z}} p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})\right]$$

$$\begin{aligned} \ln[L(\boldsymbol{\theta})] - \ln[L(\boldsymbol{\theta}_n)] &= \ln\left[\sum_{\mathbf{z}} p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})\right] - \ln[p(\mathbf{X}|\boldsymbol{\theta}_n)] \\ &= \ln\left[\sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)}\right] - \ln[p(\mathbf{X}|\boldsymbol{\theta}_n)] \end{aligned}$$

Jensen's Inequality says  $f(\sum_i \lambda_i x_i) \geq \sum_i \lambda_i f(x_i)$  if  $f$  is concave,  $\sum_i \lambda_i = 1$  and  $\lambda_i \geq 0$  for  $\forall i$ . As  $\ln(\cdot)$  is concave and  $p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)$  is density function, we have

$$\begin{aligned} &\geq \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)}\right] - \ln[p(\mathbf{X}|\boldsymbol{\theta}_n)] \\ &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)}\right] - \ln[p(\mathbf{X}|\boldsymbol{\theta}_n)] \cdot 1 \\ &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)}\right] - \ln[p(\mathbf{X}|\boldsymbol{\theta}_n)] \cdot \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \\ &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)}\right] - \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln[p(\mathbf{X}|\boldsymbol{\theta}_n)] \\ &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \left\{ \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)}\right] - \ln[p(\mathbf{X}|\boldsymbol{\theta}_n)] \right\} \\ &= \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] \end{aligned}$$

Thus,

$$\ln[L(\boldsymbol{\theta})] \geq \ln[L(\boldsymbol{\theta}_n)] + \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] =: l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$$

It can be shown that  $\ln[L(\boldsymbol{\theta})] = l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  when  $\boldsymbol{\theta} = \boldsymbol{\theta}_n$

$$\begin{aligned} l(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n) &= \ln[L(\boldsymbol{\theta}_n)] + \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_n)p(\mathbf{z}|\boldsymbol{\theta}_n)}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] \\ &= \ln[L(\boldsymbol{\theta}_n)] + \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln[1] \\ &= \ln[L(\boldsymbol{\theta}_n)] \end{aligned}$$

Thus, we know that log likelihood function wraps  $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  i.e.,  $\ln[L(\boldsymbol{\theta})] \geq l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  and  $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  touches log likelihood function at  $\boldsymbol{\theta}_n$  i.e.,  $l(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n) = \ln[L(\boldsymbol{\theta}_n)]$ .

As  $\ln[L(\boldsymbol{\theta})] \geq l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ , an increase in  $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  implies an increase in  $\ln[L(\boldsymbol{\theta})]$ .  $l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$  stops increasing at its maximum. We pick  $\boldsymbol{\theta}_{n+1} = \arg \max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)$ . This guarantees that  $\ln[L(\boldsymbol{\theta})]$  increases at every step. More formally,

$$\begin{aligned} \boldsymbol{\theta}_{n+1} &= \arg \max_{\boldsymbol{\theta}} \{l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \ln[L(\boldsymbol{\theta}_n)] + \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] \right\} \end{aligned}$$

Thus,  $\sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_{n+1})p(\mathbf{z}|\boldsymbol{\theta}_{n+1})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] \geq \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right]$  for  $\forall \boldsymbol{\theta}$  including  $\boldsymbol{\theta}_n$

$$\sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_{n+1})p(\mathbf{z}|\boldsymbol{\theta}_{n+1})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] \geq \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_n)p(\mathbf{z}|\boldsymbol{\theta}_n)}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right]$$

Since  $\ln[L(\boldsymbol{\theta}_{n+1})] - \ln[L(\boldsymbol{\theta}_n)] \geq \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_{n+1})p(\mathbf{z}|\boldsymbol{\theta}_{n+1})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right]$  and  $\sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta}_n)p(\mathbf{z}|\boldsymbol{\theta}_n)}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] = 0$ , we have

$$\ln[L(\boldsymbol{\theta}_{n+1})] - \ln[L(\boldsymbol{\theta}_n)] \geq 0$$

E step and M step can be shown as

$$\begin{aligned} \boldsymbol{\theta}_{n+1} &= \arg \max_{\boldsymbol{\theta}} \{l(\boldsymbol{\theta}|\boldsymbol{\theta}_n)\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \ln[L(\boldsymbol{\theta}_n)] + \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln\left[\frac{p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})}{p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)p(\mathbf{X}|\boldsymbol{\theta}_n)}\right] \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \ln[L(\boldsymbol{\theta}_n)] + \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln[p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})] - \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln[p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n)] \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln[p(\mathbf{X}|\mathbf{z}, \boldsymbol{\theta})p(\mathbf{z}|\boldsymbol{\theta})] \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \left\{ \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n) \ln[p(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})] \right\} \\ &= \arg \max_{\boldsymbol{\theta}} \{ \mathbb{E}_{\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n} [\ln[p(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})]] \} \end{aligned}$$

$\mathbb{E}_{\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}_n}$  is the E step.  $\arg \max_{\boldsymbol{\theta}}$  is the M step.

## 4 Bayesian Logistic Regression (Approximate EM Approach)

### 4.1 Uniform Prior

Posterior density function is

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) &\propto L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X})U(\boldsymbol{\beta}|\mathbf{X}) \\ &= L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) \cdot \text{constant} \\ &\propto L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) \end{aligned} \quad \text{Uniform prior is not a function of beta.}$$

Thus, posterior mode is the same as maximum likelihood estimate. Log posterior density is

$$\ln[p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})] \propto \ln[L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X})]$$

Log likelihood function is approximated as negative weighted least squares function locally (see IRLS).

$$\begin{aligned} &\approx -\sum_{i=1}^N w_i (z_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \\ &\propto -\frac{1}{2} \sum_{i=1}^N w_i (z_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \end{aligned}$$

where

$$\begin{aligned} z_i &= \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \frac{y_i - p_i}{w_i} \\ w_i &= p_i(1 - p_i) \\ p_i &= \text{logistic}(\mathbf{x}'_i \hat{\boldsymbol{\beta}}) \end{aligned} \quad \hat{\boldsymbol{\beta}} \text{ is from last step}$$

### 4.2 Independent Normal / Gaussian Prior

$$\begin{aligned} \ln[p(\boldsymbol{\beta}|\mathbf{y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X})] &\propto \ln[L(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X})N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X})] \\ &= \ln[L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X})] + \ln[N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X})] \\ &\approx -\frac{1}{2} \sum_{i=1}^N w_i (z_i - \mathbf{x}'_i \boldsymbol{\beta})^2 - \frac{1}{2} \sum_{j=1}^J \left( \frac{(\beta_j - \mu_j)^2}{\sigma_j^2} + \ln(\sigma_j^2) \right) + \text{constant} \\ &= -\frac{1}{2} \left\{ \sum_{i=1}^N \frac{1}{w_i} (z_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \sum_{j=1}^J \frac{1}{\sigma_j^2} (\mu_j - 1 \cdot \beta_j)^2 + \sum_{j=1}^J \ln(\sigma_j^2) \right\} + \text{constant} \\ &\approx -\frac{1}{2} \left\{ \sum_{i=1}^N \frac{1}{w_i} (z_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \sum_{j=1}^J \frac{1}{\sigma_j^2} (\mu_j - \mathbf{e}'_j \boldsymbol{\beta})^2 \right\} \quad \mathbf{e}'_j \text{ is the } j \text{ row of identity matrix} \\ &= -\frac{1}{2} \{ (\mathbf{z} - \mathbf{X}\boldsymbol{\beta})' \mathbf{W} (\mathbf{z} - \mathbf{X}\boldsymbol{\beta}) + (\boldsymbol{\mu} - \mathbf{I}_J \boldsymbol{\beta})' \boldsymbol{\Sigma} (\boldsymbol{\mu} - \mathbf{I}_J \boldsymbol{\beta}) \} \\ &= -\frac{1}{2} (\mathbf{z}_* - \mathbf{X}_* \boldsymbol{\beta})' \mathbf{W}_* (\mathbf{z}_* - \mathbf{X}_* \boldsymbol{\beta}) \end{aligned}$$

where

$$\begin{aligned} \mathbf{z}_* &= \begin{pmatrix} \mathbf{z} \\ \boldsymbol{\mu} \end{pmatrix} \\ \mathbf{X}_* &= \begin{pmatrix} \mathbf{X} \\ \mathbf{I}_J \end{pmatrix} \\ \mathbf{W}_* &= \begin{pmatrix} \mathbf{W} & \mathbf{O} \\ \mathbf{O} & \boldsymbol{\Sigma} \end{pmatrix} \end{aligned}$$

Thus, original IRLS algorithm can be applied with augmented data ( $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$  from prior normal distribution) to estimate  $\boldsymbol{\beta}$

$$\begin{aligned}
\ln[N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X})] &= \ln\left[\prod_{j=1}^J \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2} \frac{(\beta_j - \mu_j)^2}{\sigma_j^2}\right)\right] \\
&= \sum_{j=1}^J \ln\left[\frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2} \frac{(\beta_j - \mu_j)^2}{\sigma_j^2}\right)\right] \\
&= \sum_{j=1}^J \left\{ \ln[1] - \ln[(2\pi\sigma_j^2)^{1/2}] + \ln\left[\exp\left(-\frac{1}{2} \frac{(\beta_j - \mu_j)^2}{\sigma_j^2}\right)\right] \right\} \\
&= \sum_{j=1}^J \left\{ -\frac{1}{2} \ln[2\pi\sigma_j^2] - \frac{1}{2} \frac{(\beta_j - \mu_j)^2}{\sigma_j^2} \right\} \\
&= \sum_{j=1}^J \left\{ -\frac{1}{2} \ln[2\pi] \right\} + \sum_{j=1}^J \left\{ -\frac{1}{2} \ln[\sigma_j^2] - \frac{1}{2} \frac{(\beta_j - \mu_j)^2}{\sigma_j^2} \right\} \\
&= \text{constant} + -\frac{1}{2} \sum_{j=1}^J \left\{ \ln[\sigma_j^2] + \frac{(\beta_j - \mu_j)^2}{\sigma_j^2} \right\}
\end{aligned}$$

### 4.3 Independent Student-t Prior

$$\ln[p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})] \propto \ln[L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X})t(\boldsymbol{\beta}|\mathbf{X})]$$

Multivariate Student-t density is a product of Multivariate Normal's and Inverse Chi Squares' density after "integrate out"

$$\begin{aligned}
&= \ln[L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) \int N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X}) \text{Inv} - \chi^2(\boldsymbol{\Sigma}|\mathbf{X}) d\boldsymbol{\Sigma}] \\
&\approx \ln[L(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X})] + \ln[N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{X})] + \ln[\text{Inv} - \chi^2(\boldsymbol{\Sigma}|\mathbf{X})] \\
&\approx -\frac{1}{2} \left\{ \sum_{i=1}^N \frac{1}{w_i^{-1}} (z_i - \mathbf{x}_i' \boldsymbol{\beta})^2 + \sum_{j=1}^J \frac{1}{\sigma_j^2} (\mu_j - 1 \cdot \beta_j)^2 + \sum_{j=1}^J \ln(\sigma_j^2) \right\} + \text{constant} - p(\sigma_j|v_j, s_j) \\
&= -\frac{1}{2} (\mathbf{z}_* - \mathbf{X}_* \boldsymbol{\beta})' \mathbf{W}_* (\mathbf{z}_* - \mathbf{X}_* \boldsymbol{\beta}) - \frac{1}{2} \sum_{j=1}^J \ln(\sigma_j^2) + \text{constant} - p(\sigma_j|v_j, s_j)
\end{aligned}$$

Gelman et al. (2008, 2013) propose an approximate EM algorithm, in which IRLS and EM algorithm are combined and altered, to estimate  $\boldsymbol{\beta}$

1. Use last  $\boldsymbol{\beta}$  or initial value to get  $\mathbf{z}$  and  $\mathbf{W}$ . Use last  $\sigma_j^2$  or initial value  $s_j^2$  to get  $\boldsymbol{\Sigma}$
2. Now we have  $\mathbf{z}_*$ ,  $\mathbf{W}_*$  and  $\mathbf{X}_*$ . We minimize  $(\mathbf{z}_* - \mathbf{X}_* \boldsymbol{\beta})' \mathbf{W}_* (\mathbf{z}_* - \mathbf{X}_* \boldsymbol{\beta})$  to get new  $\boldsymbol{\beta}^+$  i.e.,

$$\boldsymbol{\beta}^+ = (\mathbf{X}_*' \mathbf{W}_* \mathbf{X}_*)^{-1} (\mathbf{X}_*' \mathbf{W}_* \mathbf{z}_*)$$

We also get  $\widehat{\text{Var}}(\boldsymbol{\beta}^+)$

3. Approximate E-step: First note that

$$\mathbb{E}[(\beta_j - \mu_j)^2] = \text{Var}(\beta_j - \mu_j) + (\mathbb{E}[\beta_j - \mu_j])^2 \approx \widehat{\text{Var}}(\beta_j^+) + (\beta_j^+ - \mu_j)^2$$

Expected value of log posterior density is

$$\mathbb{E}(\ln[p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})]) \approx \sum_{j=1}^J \frac{1}{\sigma_j^2} \mathbb{E}[(\mu_j - \beta_j)^2] + \mathbb{E}\left[-\frac{1}{2} \left\{ \sum_{i=1}^N \frac{1}{w_i^{-1}} (z_i - \mathbf{x}_i' \boldsymbol{\beta})^2 + \sum_{j=1}^J \ln(\sigma_j^2) \right\} + \text{constant} - p(\sigma_j|v_j, s_j)\right]$$

Substitute  $\mathbb{E}[(\mu_j - \beta_j)^2]$  with  $\widehat{\text{Var}}(\beta_j^+) + (\beta_j^+ - \mu_j)^2$

4. M-step: Maximize approximate expected log posterior density  $\mathbb{E}(\ln[p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X})])$  with respect to  $\sigma_j^2$  and get

$$\sigma_j^2 = \frac{\widehat{\text{Var}}(\beta_j^+) + (\beta_j^+ - \mu_j)^2 + v_j s_j^2}{1 + v_j}$$

5. Repeat step 1 to 4 until convergence of  $\boldsymbol{\beta}$

### 4.3.1 Observation

Without step 3 and 4, the algorithm is the same as that for independent normal prior.

Without step 3 and 4 and data augmentation, the algorithm is the same as IRLS.

Normal and uniform distribution is special cases of Student-t distribution by setting degree of freedom  $v_j$  and scale  $s_j$  to infinity, respectively. When  $v_j = 1$ , Student-t distribution is the same as Cauchy distribution.

Gelman et al. (2008) suggest default values of hyper-parameters  $\mu_j = 0$ ,  $s_j = 2.5$  and  $v_j = 1$  (So, it is Cauchy default prior) for non-intercept.

## 5 Reference

Bishop, C. M. (2006). Pattern Recognition and Machine Learning. New York :Springer.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y.-S. (2008). A Weakly Informative Default Prior Distribution For Logistic And Other Regression Models. *Annals of Applied Statistics*, 2(4), 1360-1383.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/b16018>

Borman, S. (2006). The Expectation Maximization Algorithm. A Short Tutorial