Notes on Bayesian Logistic Regression

Max Leung

April 23, 2023

1 Laplace Approximation

$$lnf(\boldsymbol{z}) \approx lnf(\boldsymbol{z}_0) + \overbrace{\nabla lnf(\boldsymbol{z}_0)'}^{\boldsymbol{0}'}(\boldsymbol{z} - \boldsymbol{z}_0) + \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\overbrace{\nabla\nabla lnf(\boldsymbol{z}_0)}^{-\boldsymbol{A}}(\boldsymbol{z} - \boldsymbol{z}_0)$$

$$= lnf(\boldsymbol{z}_0) - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0)$$

$$f(\boldsymbol{z}) \approx exp(lnf(\boldsymbol{z}_0) - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

$$= exp(lnf(\boldsymbol{z}_0))exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

$$= f(\boldsymbol{z}_0)exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

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

$$\approx \frac{1}{(2\pi)^{M/2}|\boldsymbol{A}|^{-1/2}} \underbrace{exp\{-\frac{1}{2}(\boldsymbol{z}_0 - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z}_0 - \boldsymbol{z}_0)\}}_{1} exp(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)'\boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0))$$

$$= N(\boldsymbol{z}|\boldsymbol{z}_0, \boldsymbol{A}^{-1}) = q(\boldsymbol{z})$$

where $\boldsymbol{z}_0 = arg\ max_{\boldsymbol{z}} lnf(\boldsymbol{z})$ and $\boldsymbol{A} = -\nabla\nabla lnf(\boldsymbol{z}_0)$

2 Bayesian Logistic Regression (Laplace Approximation Approach)

2.1 Posterior Distribution of Parameters

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

$$p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}) \propto p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{X})p(\boldsymbol{w}|\boldsymbol{X})$$

$$= \prod_{n=1}^{N} \sigma(\boldsymbol{w}'\boldsymbol{x}_n)^{t_n}(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))^{1-t_n} N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)$$

which is not a well known joint density function

$$lnp(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X}) = ln[\prod_{n=1}^{N} \sigma(\boldsymbol{w}'\boldsymbol{x}_n)^{t_n}(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))^{1-t_n} N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)]$$

$$= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[N(\boldsymbol{w}|\boldsymbol{m}_0, \boldsymbol{S}_0)]$$

$$= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}exp\{-\frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0)\}]$$

$$= \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1 - t_n)ln(1 - \sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0)$$

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

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

where $\boldsymbol{w}_{MAP} = arg\ max_{\boldsymbol{w}} lnp(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X})$

$$\frac{\partial lnp(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X})}{\partial \boldsymbol{w}}|_{\boldsymbol{w}_{MAP}} = \boldsymbol{0}$$

$$\frac{\partial \sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1-t_n)ln(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)}{\partial \boldsymbol{w}}|_{\boldsymbol{w}_{MAP}} = \boldsymbol{0}$$

$$\boldsymbol{X}'(\boldsymbol{t}-\boldsymbol{p}) - \boldsymbol{S}_0^{-1}(\boldsymbol{w}_{MAP}-\boldsymbol{m}_0) = \boldsymbol{0}$$

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

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

$$\boldsymbol{S} = -\nabla\nabla lnp(\boldsymbol{w}_{MAP}|\boldsymbol{t})$$
$$= -\nabla\nabla\{\sum_{n=1}^{N}[t_n ln\sigma(\boldsymbol{w}'\boldsymbol{x}_n) + (1-t_n)ln(1-\sigma(\boldsymbol{w}'\boldsymbol{x}_n))] + ln[\frac{1}{(2\pi)^{D/2}|\boldsymbol{S}_0|^{1/2}}] - \frac{1}{2}(\boldsymbol{w}-\boldsymbol{m}_0)'\boldsymbol{S}_0^{-1}(\boldsymbol{w}-\boldsymbol{m}_0)\}$$
$$= -(-\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} - \boldsymbol{S}_0^{-1})$$
$$= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + \boldsymbol{S}_0^{-1}$$

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

2.2 Special Case

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

$$\boldsymbol{X}'(\boldsymbol{t}-\boldsymbol{p}) - \boldsymbol{S}_0^{-1}(\boldsymbol{w}_{MAP}-\boldsymbol{w}_{MAP}) = \boldsymbol{0}$$
$$\boldsymbol{X}'(\boldsymbol{t}-\boldsymbol{p}) = \boldsymbol{0} \qquad \text{same as FOC of MLE}$$

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

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

$$\boldsymbol{S} \approx \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} + (\boldsymbol{O}^{-1})^{-1}$$
$$= \boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}$$

Thus,

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

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

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

2.3 Predictive Distribution

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

$$p(\widetilde{T} = 1|\boldsymbol{t}, \boldsymbol{X}) = \int p(\widetilde{T} = 1, \boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X})\partial \boldsymbol{w}$$
$$= \int p(\widetilde{T} = 1|\boldsymbol{w}, \boldsymbol{t}, \boldsymbol{X})p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{X})\partial \boldsymbol{w}$$
$$\approx \int \sigma(\boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial \boldsymbol{w}$$

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

$$= \int \int \delta(a - \boldsymbol{w}'\boldsymbol{x})\sigma(a)da \ q(\boldsymbol{w})\partial\boldsymbol{w}$$

$$= \int \int \delta(a - \boldsymbol{w}'\boldsymbol{x})\sigma(a)q(\boldsymbol{w})\partial\boldsymbol{w}da$$

$$= \int \sigma(a) \underbrace{\int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial\boldsymbol{w}}_{p(a)} da$$

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

$$\mathbb{E}_p(a) = \int p(a)a \ da$$

$$= \int \int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial\boldsymbol{w}a \ da$$

$$= \int \int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})a \ da\partial\boldsymbol{w}$$

$$= \int \int \delta(a - \boldsymbol{w}'\boldsymbol{x})a \ da \ q(\boldsymbol{w})\partial\boldsymbol{w}$$

$$= \int \delta(\boldsymbol{w}'\boldsymbol{x} - \boldsymbol{w}'\boldsymbol{x})\boldsymbol{w}'\boldsymbol{x}q(\boldsymbol{w})\partial\boldsymbol{w}$$

$$= \int \boldsymbol{w}'\boldsymbol{x}q(\boldsymbol{w})\partial\boldsymbol{w}$$

$$= \mathbb{E}_q(\boldsymbol{w}'\boldsymbol{x})$$

$$= \mathbb{E}_q(\boldsymbol{w})'\boldsymbol{x}$$

$$Var_p(a) = \mathbb{E}_p(a^2) - \mathbb{E}_p(a)^2$$

$$= \mathbb{E}_p(a^2 - \mathbb{E}_p(a)^2)$$

$$= \int p(a)(a^2 - \mathbb{E}_p(a)^2)da$$

$$= \int \int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})\partial\boldsymbol{w}(a^2 - \mathbb{E}_p(a)^2)da$$

$$= \int \int \delta(a - \boldsymbol{w}'\boldsymbol{x})q(\boldsymbol{w})(a^2 - \mathbb{E}_p(a)^2)da\partial\boldsymbol{w}$$

$$= \int q(\boldsymbol{w}) \int \delta(a - \boldsymbol{w}'\boldsymbol{x})(a^2 - \mathbb{E}_p(a)^2)da\partial\boldsymbol{w}$$

$$= \int q(\boldsymbol{w})\delta(\boldsymbol{w}'\boldsymbol{x} - \boldsymbol{w}'\boldsymbol{x})((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_p(\boldsymbol{w}'\boldsymbol{x})^2)\partial\boldsymbol{w}$$

$$= \int q(\boldsymbol{w})((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_p(\boldsymbol{w}'\boldsymbol{x})^2)\partial\boldsymbol{w}$$

$$\approx \int q(\boldsymbol{w})((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_q(\boldsymbol{w}'\boldsymbol{x})^2)\partial\boldsymbol{w}$$

$$= \mathbb{E}_q((\boldsymbol{w}'\boldsymbol{x})^2 - \mathbb{E}_q(\boldsymbol{w}'\boldsymbol{x})^2)$$

$$= Var_q(\boldsymbol{x}'\boldsymbol{w})$$

$$= \boldsymbol{x}'Var_q(\boldsymbol{w})\boldsymbol{x}$$

3 Bayesian Logistic Regression (Approximate EM Approach)

3.1 Uniform Prior

Posterior density function is

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

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

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

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

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

where

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

3.2 Independent Normal / Gaussian Prior

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

where

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

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

$$ln[N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{X})] = ln[\prod_{j=1}^{J} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp(-\frac{1}{2}\frac{(\beta_j - \mu_j)^2}{\sigma_j^2})]$$

$$= \sum_{j=1}^{J} ln[\frac{1}{\sqrt{2\pi\sigma_j^2}} exp(-\frac{1}{2}\frac{(\beta_j - \mu_j)^2}{\sigma_j^2})]$$

$$= \sum_{j=1}^{J} \{ln[1] - ln[(2\pi\sigma_j^2)^{1/2}] + ln[exp(-\frac{1}{2}\frac{(\beta_j - \mu_j)^2}{\sigma_j^2})]\}$$

$$= \sum_{j=1}^{J} \{-\frac{1}{2}ln[2\pi\sigma_j^2] - \frac{1}{2}\frac{(\beta_j - \mu_j)^2}{\sigma_j^2}\}$$

$$= \sum_{j=1}^{J} \{-\frac{1}{2}ln[2\pi]\} + \sum_{j=1}^{J} \{-\frac{1}{2}ln[\sigma_j^2] - \frac{1}{2}\frac{(\beta_j - \mu_j)^2}{\sigma_j^2}\}$$

$$= constant + -\frac{1}{2}\sum_{j=1}^{J} \{ln[\sigma_j^2] + \frac{(\beta_j - \mu_j)^2}{\sigma_j^2}\}$$

3.3 Independent Student-t Prior

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

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

$$= ln[L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}) \int N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{X})Inv - \chi^2(\boldsymbol{\Sigma}|\boldsymbol{X})\partial\boldsymbol{\Sigma}]$$

$$\approx ln[L(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X})] + ln[N(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{X})] + ln[Inv - \chi^2(\boldsymbol{\Sigma}|\boldsymbol{X})]$$

$$\approx -\frac{1}{2}\{\sum_{i=1}^{N} \frac{1}{w_i^{-1}}(z_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 + \sum_{j=1}^{J} \frac{1}{\sigma_j^2}(\mu_j - 1 \cdot \beta_j)^2 + \sum_{j=1}^{J} ln(\sigma_j^2)\} + constant - p(\sigma_j|v_j, s_j)$$

$$= -\frac{1}{2}(\boldsymbol{z}_* - \boldsymbol{X}_*\boldsymbol{\beta})'\boldsymbol{W}_*(\boldsymbol{z}_* - \boldsymbol{X}_*\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{J} ln(\sigma_j^2) + constant - p(\sigma_j|v_j, s_j)$$

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

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

We also get $\widehat{Var}(\boldsymbol{\beta}^+)$
3. Approximate E-step: First note that

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

Expected value of log posterior density is

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

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

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

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

3.3.1 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.

4 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