Data Augmentation with Polya-Gamma Latent Variables for Logistic Models

Shen Qu [*]
Portland State University
and
Supervisor Prof. Daniel Taylor-Rodriguez
Portland State University

January 24, 2021

Abstract

In Bayesian Statistics, Data Augmentation (DA) methods can generate the auxiliary or latent variables to solve some questions that cannot sample directly or the probability density/mass functions doesn't have a closed form. This paper introduces the basic DA approaches and several examples. For binary response, both probit models and logitstic model can use Gibb's samplers to simulate the target distribution through the single-layer latent variables. For logitstic regression models, generating the posterior distribution from the Pólya–Gamma family is a fast, exact simulation method.

Keywords: Data Augmentation, Logistic Models, Polya-Gamma

[*] STAT 501 Statistical Literature and Problems; The Fariborz Maseeh Department of Mathematics + Statistics

1 Introduction

Data Augmentation (DA) is a technique of newly creating synthetic data from existing data. In Statistics contest, DA algorithm includes, but not limited to Markov chain Monte Carlo (MCMC) method. There are plenty of mathematical and computational techniques such as EM algorithm which can fertilize the DA field exploration. In Deep Learning field, the DA techniques add some slightly modified copies of already existing data for increasing the amount of data (Shorten & Khoshgoftaar 2019). This paper focus on the DA applications of MCMC in Bayesian statistics. It introduces the basic ideas of MCMC algorithm through several examples and explain its main properties and conditions, that is constructing a successes DA Markov chain requires the condition of *Harris ergodic*, which includes three properties: irreducible, aperiodic, and Harris recurrent.

These approaches can be applied on the binary response regression models. As a type of Generalized Linear Models (GLMs), the response is treated as the results of a Bernoulli process. Using a link function of probit or logit, one can construct a model with the dependent binary variable $\mathbf{Y}$, the vectors of covariates $\mathbf{X}$ and parameters $\boldsymbol{\beta}$.

There are many approaches fitting these models through the Gibb's samplers. From literature, we find two simple simulation methods with some single-layer latent variables for the probit and logistic regression models. In probit version, the Bayesian methods generate the location-mixture latent variables from two truncated normal distributions. For logistic models, the latent variables is drawn from the Polya-Gamma family. Hence, its posterior distribution is a scale mixture with the linear part.

2 Literature

A main goal of Bayesian approach is to identify the posterior distribution. There are three main methods to estimate the target posterior density of β: asymptotic expansions, numerical integration, and Monte Carlo integration. When the sample size is small, asymptotic expansions may not work well. When the canonical exponential families with normal or conjugate priors are concave on the log scale, the numerical method can converge quickly to estimate the posterior joint or marginal distribution (O'Hagan & Forster 2004). In many cases, the numerical integration is difficult for the high-dimension models. Zellner & Rossi (1984) proposed an importance sampling generalization of Monte Carlo simulation. They found that Monte Carlo methods are more adequate and efficient to achieve a good approximation. Gelfand & Smith (1990) develop these methods to Gibbs sampling. When sampling from the conditional distribution is easier, Gibbs sampler is a highly useful MCMC method to sample from multivariate distributions. Metropolis-Hastings algorithm (Tierney 1994) is another important progress of MCMC. It is the generalized version of Gibb's sampler. A systematic review can be found in Agresti & Hitchcock (2005) and Agresti (2012).

For binary response, the generalized linear models use a link function such as probit or logit to connect the linear part to the response. Albert & Chib (1993) provide a computational simple Bayesian approach for probit regression modeling. They assume the relationship between the binary data and the latent variables. The question can be solved through an underlying normal regression model. The models have a hierarchical prior structure as follow:

$$Z \sim N_p(\boldsymbol{X\beta, I}), \quad \boldsymbol{\beta} \sim N_p(\boldsymbol{A\beta_0, \sigma^2 I}), \quad (\beta_0, \sigma^2) \sim \pi(\beta_0, \sigma^2)$$

where Z is the latent variable, $\boldsymbol{\beta}$ is the parameter vector of the linear predictors. $\boldsymbol{A\beta_0}$ is the linear subspace, and assuming $\boldsymbol{\beta_0, \sigma^2}$ are independent and noninformative priors. This method can also extend to ordered multinomial responses and mixture models (Hoff 2009, Gelman et al. 2020).

The Bayesian logistic models using approximation methods for quite some time. Piegorsch & Casella (1996) suggest the parametric empirical Bayes methods using marginal maximum likelihood estimate hyperparameters and extend the link function. Greenland (2001) discuss that the approximation may be inadequate with sparse data and they suggest to use conjugate priors to conduct exact analysis. A new step is that Polson et al. (2013) introduce an new method that assume the latent variables follow a Polya-Gamma distribution. It is as simple as Albert & Chib (1993) 's method, which don't need multiple layers. This approach assign the linear part at the scale parameter. Choi & Hobert (2013) prove that this polya-gamma Gibbs sampler for logistic model is uniformly ergodic. Some further discussions of this DA algorithm include convergence rates of the DA and the Haar PX-DA algorithm (Choi 2014), a sandwich version of DA (Choi & Hobert 2016), and the analysis of variance (Choi & Román 2017).

3 Appoaches

3.1 Basic ideas

- Expectation

Assume a random variable $g(X)$ and pdf $f_X(x) : \mathbb{R}^p \to [0, \infty)$. Denote interested $E[g(x)]$ is the expected value or mean of $g(X)$ (Casella & Berger 2002, *Def2.2.1*).

$$E_{f_X}[g(x)] = \int_{\mathbb{R}^p} g(x) f_X(x) dx$$

- Joint, conditional, and marginal density

A probability function with a bivariate random vector $f(x, z)$ is called joint pdf.

A marginal probability density function of X is the integral of joint pdf $f(x, z)$ on X: $f_X(x) = \int_{\mathbb{R}^p} f(x, z) dz$.

The relationship among joint pdf, marginal pdf, and conditional pdf of X given Z is: $f(x, z) = f(x|z)f(z)$

Example 1 Given $f_X(x) = 3x^2 I_{(0,1)}(x)$, then $E[X] = \int_0^1 x f_X(x) dx = \frac{3}{4}x^4\Big|_0^1 = 0.75$.

Suppose the joint pdf $f(x, z) = 3x I_{(0<z<x<1)}$, then we can get all the rest of density functions:

$$\begin{aligned} f_Z(z) &= \int_z^1 f(x, z) dx = \frac{3}{2}(1 - z^2) \\ f_{Z|X}(z|x) &= \frac{f(x, z)}{f_Z(z)} = \frac{1}{x} I(0 < z < x) \\ f_{X|Z}(x|z) &= \frac{f(x, z)}{f_X(x)} = \frac{2x}{1 - z^2} I(z < x < 1) \end{aligned}$$

Example 2: Given a joint pdf $f(x,z) = (\sqrt{2}\pi)^{-1}\exp\left[-(x^2 - \sqrt{2}xz + z^2)\right]$, which is a bivariate normal density with means $\mu_X = \mu_Z = 0$, variances $\sigma_X = \sigma_Z = 1$, and correlation $\rho = 2^{-1/2}$, we can get $f_X(x)$ by integral of $f(x,z)$.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Z}(x,z)\,\mathrm{d}z \\ &= \frac{\exp(-\frac{1}{2}x^2)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\frac{1}{2}}} \exp\left[-\frac{1}{2}(z - \frac{x}{\sqrt{2}})^2\right] \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}x^2) \end{aligned}$$

Another way is, suppose $f(z|x) \sim N(\frac{x}{\sqrt{2}}\frac{1}{2})$, we also can get $f_X(x) = \frac{f(x,z)}{f(z|x)} = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)$

Then, we can utilize these properties to construct MCMC algorithm to solve the more complex questions.

- Monte Carlo approximation

In some case, *constructing estimates of features of probability distributions are impossible or very difficult to compute analytically*

if we have data from an unknown probability distribution belonging to a model we can estimate that distribution in some way. if we have a feature (parameter) of the true distribution we want to estimate, use Monte Carlo to estimate the feature by random quantities generated using the estimated distribution. (Doksum 2015, ch. 10.1)

Regardless of the distribution, if we draw i.i.d. $g(X_1), g(X_2), \ldots, g(X_m)$ from $f_X(x)$, then, for large m, the empirical distribution of is arbitrarily close to the distribution under$f_X(x)$. That is $\frac{1}{m}\sum_{i=1}^{m} g(X_i) \overset{a.s}{\to} E_{f_X}[g(x)]$.

- Markov Chain

When it is impossible to simulate from $f_X(x)$, such as $f_X(x)$ doesn't have closed form or $f(x,z)$ is hard to integral, we can construct a Markov chain. Suppose $X = \{X_i\}_{i=0}^{\infty}$, with state space $\mathbf{X}$. If the current state of the chain is $X_i = x$, then the density of the next state, X_{n+1}, is $k(\cdot|x)$. **???**

- Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo is a homogeneous positive recurrent Markov chain (Tanner & Wong 1987, Swendsen & Wang 1987). $K(x, x')$ is called the transition kernel which has an unique stationary density $k(x'|x)$. The conditional density $k(x'|x)$ is also called the Markov transition density (Mtd).

Our goal is to find $f_X(x)$. The conditions are that a joint pdf $f : \mathbb{R}^p \times \mathbb{R}^q \to [0, \infty)$ exists, which includes a auxiliary real variable Z. If the invariant x-marginal pdf $f_X(x) = \int_{\mathbb{R}^q} f(x, z)dz$ holds. And the associated two conditional pdfs $f_{X|Z}(x'|z)$ and $f_{Z|X}(z|x)$ are applicable for simulation.

We can shows that $k(x'|x)$ is a pdf of a random variable X'

$$\begin{aligned}
\int_X k(x'|x)dx' &= \int_X \left[\int_Z f_{X|Z}(x'|z) f_{Z|X}(z|x)dz \right] dx' \\
&= \int_Z f_{Z|X}(z|x) \left[\int_X f_{X|Z}(x'|z)dx' \right] dz \\
&= \int_Z f_{Z|X}(z|x)dz = 1
\end{aligned}$$

Even without the $k(x'|x)$ in closed form, we still can simulate the Markov chain X as follows.

Given an initial value x_0

1. Draw $z^{(1)} \sim f_{Z|X}(\cdot|x^{(0)})$.

2. Draw $x^{(1)} \sim f_{X|Z}(\cdot|z^{(1)})$.

Repeat the two steps ...

i. Draw $z^{(i)} \sim f_{Z|X}(\cdot|x^{(i-1)})$.

i+1. Draw $x^{(i)} \sim f_{X|Z}(\cdot|z^{(i)})$.

At the end, we can calculate $E_{f_X}[g(x)] \approx \frac{1}{m} \sum_{i=1}^{m} g(X_i)$

Example 3 The Student's t_4 with four degree of freedom.

Assume we don't know the target $f_X(x) = \frac{3}{8}(1+\frac{x^2}{4})^{-\frac{5}{2}}$. We can take $f(x,z) = \frac{4}{\sqrt{2\pi}} z^{\frac{3}{2}} \exp\{-z(\frac{x^2}{2}+2)\} I_{(0,\infty)}(z)$, which satisfied $f_X(x) = \int_{\mathbb{R}^p} f(x,z)dz$. Then, it is easy to draw from two conditional pdf of $(Z|X=x) \sim Gamma(\frac{5}{2}, \frac{x^2}{2}+2)$ and $(X|Z=z) \sim N(0, z^{-1})$.

- Simple slice sample

The **simple slice sampler** is a DA technique through a *stepping-out and shrinkage* procedures (Neal 2003). If we can factorize $f_X(x)$ to $q(x)l(x)$, supposing an auxiliary variable Z satisfies $(Z|X=x) \sim Uniform(0, l(x))$. The joint distribution over x and z define a region $\{(x,z) : 0 \le z \le l(x)\}$ under the curve of $l(x)$. A random state of $x^{(0)}$ can give a vertical "slice" for all possible z bounded from zero to $l(x^{(0)})$. Then an uniformly selected $z^{(1)}$ can give a horizontal "slice" for all x inside the curve. That slice is from zero to $\int_{z^{(1)}}^{l(x)} f_{X|Z}(x|z)dx$, the union of intervals that all x satisfy $l(x) \ge z^{(1)}$. The next state $x^{(1)}$ around $x^{(0)}$ is accepted only if it doesn't step out the horizontal "slice". Each updating will shrink (x,z) until the whole region is explored.

Example 1 Cont. Rewrite $f_X(x) = 3x^2 I_{(0,1)}(x) = [3xI_{(0,1)}(x)][x] = q(x)l(x)$.

For $l(x) = x$, the vertical slice is $\{z|x : 0 \le z \le x\}$. $(Z|X=x) \sim Unif(0,x)$.

The horizontal slice is $\{x|z : x \ge z\}$. Given $f_{X|Z}(x|z) = \frac{2x}{1-z^2} I(z < x < 1)$, The union of intervals is $\int_z^x f_{X|Z}(x|z)dx = \frac{x^2-z^2}{1-z^2}$. Then $(X|Z=z) \sim Unif(0, \frac{x^2-z^2}{1-z^2})$

Suppose $U \sim Unif(0,1)$, we can get $(Z|X) = xU$ and $(X|Z) = \sqrt{U(1-z^2)+z^2}$ by transformation.

Given an initial value $x^{(0)}$

1. Draw $u^{(1)} \sim Unif(0,1)$
2. Calculate $z^{(1)} = x^{(0)}u^{(1)}$
3. Calculate $x^{(1)} = \sqrt{u^{(1)}(1-(z^{(1)})^2)+(z^{(1)})^2}$.

Update $(x^{(i)}, z^{(i)})$ by repeating step 1, 2, 3, ...

Calculate $E_{f_X}[g(x)] \approx \frac{1}{m}\sum_{i=1}^m g(X_i)$

3.2 Properties

3.2.1 Conditions

A well-behaved Markov Chain should be Harris ergodic, which satisfies three properties: irreducible, aperiodic, and recurrent. A sufficient condition for Harris ergodicity is

$$\mathcal{K} : k(x'|x) > 0 \quad \forall x', x \in \mathbf{X}$$

- **Detailed balance**

For all $x, x' \in \mathbf{X}$, let

$$\delta(x', x) = k(x'|x) f_X(x) = \int_Z \frac{f(x', z)}{f_Z(z)} \cdot \frac{f(z, x)}{f_X(x)} f_X(x) dz$$
$$\delta(x, x') = k(x|x') f_X(x') = \int_Z \frac{f(x, z)}{f_Z(z)} \cdot \frac{f(z, x')}{f_X(x')} f_X(x') dz$$

So $\delta(x', x) = \delta(x, x')$ is a symmetric function satisfying

$$\delta(x, x') = f_X(x) k(x'|x) = f_X(x') k(x|x'), \forall x, x'$$

which is called the *detailed balance* condition.

A type of MCMC, Metropolis Sampling Algorithm defines a symmetric Markov kernel $K_0(\cdot, \cdot)$ in the original Metropolis algorithm and

$$r(x, x') = \min\left(1, \frac{f_X(x')}{f_X(x)}\right)$$

If and only if $r(x, x') = \frac{\delta(x,x')}{f_X(x) K_0(x,x')}$

$$K(x, x') = r(x, x') K_0(x, x'), \quad x \neq x'$$

Then K is a Markov kernel satisfying detailed balance. The Markov chain is reversible with respect to $f_X(x)$. When the MCMC always satisfying detailed balance $r \equiv 1$, this algorithm is called Gibbs' Sampler (Doksum 2015, ch. 10.4.3).

- **Invariant (stationarity)**

f_X is an invariant density for K when

$$\int_X k(x'|x)f_X(x)dx = f_X(x')$$

which means that, if $X_n \sim f_X$, then $X_{n+1} \sim f_X$. The Markov chain is time homogeneous.

For the case of the multivariate distribution $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, the Gibbs sampling procedure requires the stationary density is an unique distribution defined by the full conditionals, which mean each single component of the random vector given the values of the rest.

3.2.2 Convergence

For a reversible Markov chain X satisfies Harris ergodicity, the **Strong Law of Large Numbers** holds. The marginal density of X_i will converge to the invariant density f_X no matter how the chain is started. Denote $P^n(x, A) = Pr(X_n \in A|X_0 = x)$ where A is a measurable set. $\phi(A) = \int_A f_X(x)dx$. That is

$$\lim_{n\to\infty} ||P^n(x, \cdot) - \phi(\cdot)|| \to 0$$

where $||P^n(x, \cdot) - \phi(\cdot)|| = \sup_A |P^n(x, A) - \phi(A)|$

Let $L^1(f_X)$ denote the set of functions $h : \mathbf{X} \to \mathbb{R}$ $\int_X |h(x)|f_X(x)dx < \infty$ and $E_{f_X}h = \int_X h(x)f_X(x)dx$. For $g \in L^1(f_X)$

$$\bar{g}_n := \frac{1}{n}\sum_{i=0}^{n-1} g(X_i) \overset{a.s}{\to} E_{f_X}[g(x)]$$

3.2.3 Geometric ergodicity

For all $x \in \mathbf{X}$ and all n=1,2,.., if exist a function $M : \mathbf{X} \to [0, \infty)$ and a constant $0 < \rho < 1$

$$||P^n(x, \cdot) - \phi(\cdot)|| \le M(x)\rho^n$$

Then the Markov chain X is geometrically ergodic.

- Drift condition

Define $V : \mathbf{X} \to [0, \infty)$ is unbounded off compact sets, constants $\lambda \in [0, 1]$, and $L \in \mathbb{R}$. If the *drift function* V exist and

$$E[V(X_{n+1}|X_n = x)] \leq \lambda V(x) + L$$

Then the Markov chain X is geometrically ergodic.

3.2.4 Central Limit Theorems

Assume that X is geometrically ergodic and $g \in L^2(f_X)$. Define $\sigma^2 = E_{f_X} g^2 - (E_{f_X} g)^2$ and $\kappa^2 = \sigma^2 + 2\sum_{k=1}^{\infty} c_k < \infty$. As $n \to \infty$

$$\sqrt{n}(\bar{g}_n - E_{f_X} g) \xrightarrow{d} N(0, \kappa^2)$$

3.2.5 Minorization regenration and alteernative CLT

3.3 MCMC and EM algorithm

We can extend data augmentation methods by some techniques both from MCMC and EM (Expectation and Maximization) algorithm (McLachlan & Krishnan 2008). Define Y as the observed data, which are sampled from a family of pdfs $\{p(y|\theta) : \theta \in \Theta\}$. To find the feature (parameter) of the true distribution, the target is the θ-marginal posterior density $\pi(\theta|y)$ and corresponding expectations $E[\pi(\theta|y)]$. Suppose the **missing data**, $z \in Z \subset \mathbb{R}^q$ can be identified. The **complete data** posterior density as

$$\pi(\theta, z|y) = \frac{p(y, z|\theta)\pi(\theta)}{\int_\Theta \int_Z p(y, z|\theta)\pi(\theta) dz d\theta} = \frac{p(y, z|\theta)\pi(\theta)}{\int_\Theta p(y|\theta)\pi(\theta) d\theta} = \frac{p(y, z|\theta)\pi(\theta)}{c(y)}.$$

where $p(y, z|\theta)$ represents the joint density, $\pi(\theta)$ denotes the prior density, $c(y)$ is the marginal density of the data. we can check that the θ-marginal pdf is invariant:

$$\int_Z \pi(\theta, z|y) dz = \frac{\pi(\theta)}{c(y)} \int_Z p(y, z|\theta) dz = \frac{p(y|\theta)\pi(\theta)}{c(y)} = \pi(\theta|y).$$

This question can be solved by deterministic EM algorithm. Let θ^i denote the current guess to the mode of the observed posterior $p(\theta|y)$. $p(\theta|z, y)$ is the augmented posterior. $p(z|\theta^i, y)$ represents the conditional predictive distribution of the latent data Z, conditional on the current guess to the posterior mode. In the E-step, we can find the Q-function, the conditional expectation of the complete-data log likelihood function as

$$Q(\theta, \theta^i) = \int_Z \log(p(\theta|z, y))p(z|\theta^i, y)dz$$

In the M-step, Q-function is maximized with respect to θ to obtain θ^{i+1} by Coordinate-Ascent Algorithm. Continue the E-step and M-step until convergent. However, for the high-dimension θ the Q-function is hard to differentiate.

Once we can straightforwardly sample from two full posterior densities $\pi(z|\theta, y)$ and $\pi(\theta|z, y)$, then the target density $\pi(\theta|y)$ can be simulated by MCMC (Meng & van Dyk 1999). EM algorithm works well with exponential families when the log-concave property holds. But if $p(z|\theta, y)$ is not "sufficiently smooth", the iterates θ^i may converge to some local maxima, minima, or saddle stationary point of $p(\theta|y)$. In small samples, the posterior or likelihood may not consistent with the normal approximation. Unlike the EM algorithm, the DA methods base on the entire likelihood function or posterior distribution. It can converge to the global maximizer and can also improves the inference in both small and large sample situation (Tanner 1993).

The EM algorithm converges at a linear rate(Dempster et al. 1977), with the rate depending on the proportion of information about θ in $(\theta|Y)$ is observed. In contrast, MCMC can converge at a geometric speed, More discussions of EM and DA can be found in Dyk & Meng (2010).

Example 4 The multivariate location-scale Student's t density with known degree of freedom $\nu > 0$.

Assume the true distribution of response $Y_i \sim t_{\nu,\mu,\sigma^2}$,i=1,..,m.

$$p(y|\mu, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\sigma^2}\Gamma(\frac{\nu}{2})}(1 + \frac{(y-\mu)^2}{\nu\sigma^2})^{-\frac{\nu+1}{2}}$$

Let the prior density $\pi(\mu, \sigma^2) \propto 1/\sigma^2$. The target posterior density is

$$\begin{aligned}\pi(\mu, \sigma^2|y) &\propto \pi(\mu, \sigma^2)\prod_{i=1}^{m} p(y_i|\mu, \sigma^2) \\ &\propto (\sigma^2)^{-\frac{m+2}{2}}\prod_{i=1}^{m}(1 + \frac{(y_i-\mu)^2}{\nu\sigma^2})^{-\frac{\nu+1}{2}}\end{aligned}$$

Meng & van Dyk (1999) propose a solution by DA algorithm. Let $Z = \mathbb{R}_+^p$ and make a i.i.d. paired vectors $(Y_{1:m}, Z_{1:m})$, where $(Y_i|Z_i, \mu, \sigma^2) \sim N(\mu, \sigma^2/z_i)$, $(Z_i|\mu, \sigma^2) \sim Gamma(\nu/2, \nu/2)$. Then the joint density is $p(y, z|\mu, \sigma^2) = \prod_{i=1}^m p(y_i|z_i, \mu, \sigma^2)p(z_i|\mu, \sigma^2)$

Checking the (μ, σ^2)-marginal density confirms that the condition holds.

$$\begin{aligned}\int_Y p(z, y|\mu, \sigma^2)dz =& \prod_{i=1}^m \int_{\mathbb{R}_+} \frac{\sqrt{z_i}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{z_i}{2\sigma^2}(y_i-\mu)^2\right] \frac{(\frac{\nu}{2})^{(\frac{\nu}{2})}}{\Gamma(\frac{\nu}{2})} z_i^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu}{2}z_i\right] dz_i \\ =& \prod_{i=1}^m \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\sigma^2}\Gamma(\frac{\nu}{2})}(1+\frac{(y_i-\mu)^2}{\nu\sigma^2})^{-\frac{\nu+1}{2}}\end{aligned}$$

Then we can rearrange the joint pdf.

$$\begin{aligned}p(\mu, \sigma^2, z|y) &\propto \pi(\mu, \sigma^2)p(z, y|\mu, \sigma^2) \\ &\propto \frac{1}{\sigma^2} \prod_{i=1}^m \frac{\sqrt{z_i}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{z_i}{2\sigma^2}(y_i-\mu)^2\right] \frac{(\frac{\nu}{2})^{(\frac{\nu}{2})}}{\Gamma(\frac{\nu}{2})} z_i^{\frac{\nu}{2}-1} \exp\left[-\frac{\nu}{2}z_i\right]\end{aligned}$$

and get the first full conditional pdf

$$\pi(z|\mu, \sigma^2, y) \propto \frac{(\frac{(y-\mu)^2}{2\sigma^2}+\frac{\nu}{2})^{(\frac{\nu+1}{2})}}{\Gamma(\frac{\nu+1}{2})} z^{\frac{\nu+1}{2}-1} \exp\left[-\left(\frac{(y-\mu)^2}{2\sigma^2}+\frac{\nu}{2}\right)z\right]$$

Define $\hat{\mu} = \frac{1}{z.}\sum_{j=1}^m y_j z_j$ The second full conditional pdf is

$$\begin{aligned}\pi(\mu|\sigma^2, y, z) &\propto \exp\left[-\frac{\sum z_j}{2\sigma^2}\mu^2 + \frac{\sum z_j y_j}{\sigma^2}\mu\right] \\ &\propto \exp\left[-\frac{z.}{2\sigma^2}(\mu - \frac{\sum y_j z_j}{z.})^2\right] \\ &\propto \frac{\sqrt{z.}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{z.}{2\sigma^2}(\mu-\hat{\mu})^2\right]\end{aligned}$$

Define $\hat{\sigma}^2 = \frac{1}{z.}\sum_{j=1}^m z_j(y_j-\hat{\mu})^2$. The third full conditional pdf can get by

$$\begin{aligned}\pi(\sigma^2|y, z) &= \int_{\mathbb{R}} p(\mu, \sigma^2, z|y)d\mu \\ &\propto \frac{(\frac{z.\hat{\sigma}^2}{2})^{\frac{m+1}{2}}}{\Gamma(\frac{m+1}{2})}(\frac{1}{\sigma^2})^{\frac{m+1}{2}+1} \exp\left[-\frac{z.\hat{\sigma}^2}{2\sigma^2}\right]\end{aligned}$$

Thus, we can simulate (μ, σ^2) by three steps:

1. Draw $(Z_i|\mu, \sigma^2, y)$ from $Gamma(\frac{\nu+1}{2}, \frac{1}{2}(\frac{(y_i-\mu)^2}{\sigma^2}+\nu))$.

2. Draw $(\sigma^2|y,z)$ from $IG(\frac{m+1}{2}, \frac{z.\hat{\sigma}^2}{2})$.

3. Draw $(\mu|\sigma^2,y,z)$ from $N(\hat{\mu}, \frac{\sigma^2}{z.})$.

4 Probit Models

Probit models is one type of Generalized Linear Models. Suppose the binary response $Y_1,..,Y_n \sim Bern(p_i)$ is the observed data with sample size n. $X_i^T = (X_{i1},..,X_{ip})$ known p covariates. $\beta_{k\times 1}$ is the unknown vector of parameters we want to estimate.

To construct a regression model, suppose a known CDF $H(\cdot)$ as the link function with the linear part $\mathbf{x}_i^T\boldsymbol{\beta}$ Then, let $p_i = H(\mathbf{x}_i^T\boldsymbol{\beta})$.

Denote $\pi(\boldsymbol{\beta})$ as the prior density. The target conditional density is

$$\pi(\boldsymbol{\beta}|y) = \frac{\pi(\boldsymbol{\beta})}{C(y)}\prod_{i=1}^{k} H(\mathbf{x}_i^T\boldsymbol{\beta})^{y_i}(1-H(\mathbf{x}_i^T\boldsymbol{\beta}))^{1-y_i}$$

where $c(y)$ is free from β

$$C(y) = \int \pi(\boldsymbol{\beta})\prod_{i=1}^{k} H(\mathbf{x}_i^T\boldsymbol{\beta})^{y_i}(1-H(\mathbf{x}_i^T\boldsymbol{\beta}))^{1-y_i}d\boldsymbol{\beta}$$

When $H(\cdot)$ represents the standard normal CDF, it obtains the probit model. When $H(\cdot)$ is logistic CDF, it is the logistic model.

While the logistic models have a close form, probit models cannot compute in closed form and are hard to integral.

Sampling from the posterior density $\pi(\boldsymbol{\beta}|y)$ with high dimension is also problematic. (why)

Albert & Chib (1993) propose a DA algorithm to estimate the exact posterior distribution of β. This simulation-based approach introduces latent variables $Z_1,..Z_N \overset{iid}{\sim} N(\mathbf{x}_i^T\boldsymbol{\beta},1)$. The relationship between observed data and missing data is $y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \le 0 \end{cases}$.

The joint mass function is

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\beta})}{C(y)} \prod_{i=1}^{N} [p(y_i|\beta)] f(z_i|\beta) \\ &\propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \left[\Phi(\mathbf{x}_i^T\boldsymbol{\beta})^{y_i}(1-\Phi(\mathbf{x}_i^T\boldsymbol{\beta}))^{1-y_i}\right] \phi(z_i|\beta) \\ &= \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \left[1_{z_i>0}1_{y_i=1} + 1_{z_i\leq 0}1_{y_i=0}\right] \phi(z_i|\beta)\end{aligned}$$

where Φ is a standard normal CDF, ϕ is a normal pdf with means equal $\mathbf{x}_i^T\boldsymbol{\beta}$ and variances equal one.

Assign a flat prior to β, we can get the first conditional posterior density:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{Z}) \propto \prod_{i=1}^{N} \phi(Z_i; \mathbf{x}_i^T\boldsymbol{\beta}, 1)$$

For a multivariate linear model $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N_N(0, \mathbf{I})$, the least squares estimates have $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$ and $\hat{\sigma}_{\beta}^2 = (\mathbf{X}^T\mathbf{X})^{-1}$. Thus, we can draw $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{Z}$ from $N_k(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}, (\mathbf{X}^T\mathbf{X})^{-1})$

The second conditional posterior pdf is

$$Z_i|\boldsymbol{y}, \boldsymbol{\beta} \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, 1) \begin{cases} \text{truncated at the left by 0} & \text{if } y_i = 1 \\ \text{truncated at the right by 0} & \text{if } y_i = 0 \end{cases}$$

5 Logistic Models

In logistic models, the link function $p_i = H(z_i) = \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}$. The log odds of success $H^{-1}(p_i) = \log(\frac{p_i}{1-p_i})$

The joint pmf is

$$\begin{aligned}\pi(\boldsymbol{\beta}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\beta})}{C(y)} \prod_{i=1}^{N} [p(y_i|\beta)] \\ &\propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \left[(\frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}})^{y_i} (1 - \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}})^{1-y_i} \right] \\ &= \pi(\boldsymbol{\beta}) \prod_{i=1}^{N} \left[\frac{e^{y_i\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}} \right]\end{aligned}$$

5.1 Single-layer DA methods

Polson et al. (2013) propose a DA approach which only need a single layer of latent variables, which involve the Pólya–Gamma distribution. Denote $W_i \sim PG(1, |\mathbf{x}_i^T\boldsymbol{\beta}|)$,i=1:n, then (no absolute value in polson2013)

$$f(\omega|\beta) = \cosh(\frac{1}{2}\mathbf{x}^T\boldsymbol{\beta})\exp\left[-\frac{1}{2}(\mathbf{x}^T\boldsymbol{\beta})^2\omega\right] g(\omega)$$

where $\cosh c = \frac{1}{2}(e^c + e^{-c}) = \frac{1+e^{2c}}{2e^c}$, $g(\omega)$ is free of β s.t.

$$g(\omega) = \sum_{n}^{\infty}(-1)^n \frac{2n+1}{\sqrt{2\pi\omega^3}}\exp[-\frac{(2n+1)^2}{8\omega}]\mathbb{I}_{(0,\infty)}(\omega)$$

Since y is observed data, $\pi(\omega|\beta, y) = f(\omega|\beta)$.

Assign the prior of $\boldsymbol{\beta}$ is $\sim N_p(\mathbf{b}, \mathbf{B})$

$$\begin{aligned}
\pi(\boldsymbol{\beta}, \boldsymbol{\omega}|\boldsymbol{y}) &= \frac{\pi(\boldsymbol{\beta})}{C(y)}\prod_{i=1}^{N}[p(y_i|\beta)]f(\omega_i|\beta) \\
&= \frac{\pi(\boldsymbol{\beta})}{C(y)}\prod_{i=1}^{N}\left[\frac{e^{y_i\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\right]\cosh(\frac{1}{2}\boldsymbol{x_i^T}\boldsymbol{\beta})\exp\left[-\frac{1}{2}(\mathbf{x}_i^T\boldsymbol{\beta})^2\omega_i\right]g(\omega_i) \\
&\propto \phi_p(\mathbf{b}, \mathbf{B})\prod_{i=1}^{N}\frac{e^{y_i\mathbf{x}_i^T\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}\cdot\frac{1+e^{\mathbf{x}_i^T\boldsymbol{\beta}}}{2e^{\frac{1}{2}\mathbf{x}_i^T\boldsymbol{\beta}}}\exp\left[-\frac{1}{2}(\mathbf{x}_i^T\boldsymbol{\beta})^2\omega\right] \\
&= 2^{-n}\phi_p(\mathbf{b}, \mathbf{B})\prod_{i=1}^{N}\exp\left[(y_i-\frac{1}{2})\mathbf{x}_i^T\boldsymbol{\beta} - \frac{1}{2}(\mathbf{x}_i^T\boldsymbol{\beta})^2\omega_i\right]
\end{aligned}$$

When ω_i is known, $\pi(\boldsymbol{\beta}, \boldsymbol{\omega}|\boldsymbol{y}) = \pi(\boldsymbol{\beta}|\boldsymbol{\omega}, \boldsymbol{y})$. Let $\boldsymbol{\Omega} = diag_n(\omega_{\mathbf{i}})$; $\boldsymbol{\kappa} = \mathbf{y_{1:n}} - \frac{1}{2}$ Then we get to conditional pmfs of $\pi(\boldsymbol{\beta}|\boldsymbol{\omega}, \boldsymbol{y})$

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\boldsymbol{\omega}, \boldsymbol{y}) &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\beta'}(\mathbf{x'}\boldsymbol{\Omega}\mathbf{x} + \mathbf{B^{-1}})\boldsymbol{\beta} - \mathbf{2}\boldsymbol{\beta'}(\mathbf{x'}\boldsymbol{\kappa} + \mathbf{B^{-1}b})\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\mathbf{v}_\omega^{-1}[\boldsymbol{\beta'}\mathbf{I}\boldsymbol{\beta} - \mathbf{2}\boldsymbol{\beta'}\mathbf{v}_{\boldsymbol{\omega}}(\mathbf{x'}\boldsymbol{\kappa} + \mathbf{B^{-1}b})]\right\} \\
&\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{m}_{\boldsymbol{\omega}})'\mathbf{v}_{\boldsymbol{\omega}}^{-\mathbf{1}}(\boldsymbol{\beta} - \mathbf{m}_{\boldsymbol{\omega}})\right\}
\end{aligned}$$

where $\mathbf{m}_\omega = \underset{(\mathbf{p\cdot p})}{\mathbf{v}_\omega}(\underset{(\mathbf{p\cdot n})}{\mathbf{x^T}}(\mathbf{y_{1:n}} - \frac{\mathbf{1}}{\mathbf{2}}) + \mathbf{B^{-1}b})$; $\mathbf{v}_\omega = (\mathbf{x^T}\boldsymbol{\Omega}\mathbf{x} + \mathbf{B^{-1}})^{-\mathbf{1}}$. Then we confirm that $\boldsymbol{\beta}|\boldsymbol{\omega}, \boldsymbol{y} \sim N_p(m_\omega, V_\omega)$

To sample from the posterior distribution using the Pólya–Gamma method, simply iterate two steps:

$$(\omega_i|\boldsymbol{\beta}) \sim PG(n_i, \mathbf{x}_i^T\boldsymbol{\beta})$$
$$(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\omega}) \sim N(\mathbf{m}_\omega, \mathbf{v}_\omega)$$

5.2 Other Bayesian methods

One method is similar with Albert & Chib (1993). Let $y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i \leq 0 \end{cases}$. The *latent utilities* $z_i = \mathbf{x}_i^T\boldsymbol{\beta} + \varepsilon_i$, $\varepsilon_i \sim Logistic(1)$

The standard approach has been to add another layer of auxiliary variables to handle the logistic error model on the latent-utility scale. One strategy is to represent the logistic distribution as a normal-scale mixture

$$(\epsilon_i|\phi_i) \sim N(0, \phi_i); \quad \phi_i = (2\lambda_i)^2; \lambda_i \sim KS(1) \quad \text{Kolmogorov–Smirnov distribution}$$

Alternatively, one may approximate the logistic error term as a discrete mixture of normals.

$$(\epsilon_i|\phi_i) \sim N(0, \phi_i); \quad \phi_i = \sum_{k=1}^{K} \omega_k \delta_{\phi^{(k)}}$$

where δ_ϕ indicates a *Dirac measure* at ϕ. The weights ω_k and the points $\phi^{(k)}$ in the discrete mixture are fixed for a given choice of k so that *the Kullback–Leibler divergence* from the true distribution of the random utilities is minimized. Frühwirth-Schnatter and Frühwirth (2010) found that the choice of $K = 10$ leads to a good approximation.

The discrete mixture of normals is an approximation, but it outperforms the scale mixture of normals in terms of effective sampling rate, as it is much faster.

	Albert and Chib (1993)	Polson et al. (2013)
Gaussian	location mixture	scale mixture
Latent variables	truncated normals	Polya-Gamma

One may also arrive at the hierarchy above by manipulating the random utility derivation of McFadden (1974)

The dRUM. One must use a table of different weights and variances representing different normal mixtures, to approximate a finite collection of type-III logistic distributions, and interpolate within this table to approximate the entire family.

Another approximation: the use of a Student-t link function as a close substitute for the logistic link. This also introduces a second layer of latent variables, in that the Student-t error model for z_i is represented as a scale mixture of normals.

5.3 Other Frequentist methods

5.3.1 MLE methods

5.3.2 EM Algoritm

5.4 Implementation

5.5 Summary

Our data-augmentation scheme differs from each of these approaches in several ways.

it does not appeal directly to the random-utility interpretation of the logit model. Instead, it represents the logistic CDF as a mixture with respect to an infinite convolution of gammas.

the method is exact, in the sense of making draws from the correct joint posterior distribution, rather than an approximation to the posterior that arises out of an approximation to the link function.

like the Albert and Chib (1993) method, it requires only a single layer of latent variables.

In binary logit models, the Pólya–Gamma is more efficient than all previously proposed data-augmentation schemes.

the Pólya–Gamma method always had a higher effective sample size than the two default Metropolis samplers we tried.

the Pólya–Gamma method truly shines when the model has a complex prior structure.

References

Agresti, A. (2012), *Categorical Data Analysis*, John Wiley & Sons, Incorporated, Somerset, UNITED STATES.
URL: *http://ebookcentral.proquest.com/lib/psu/detail.action?docID=1168529*

Agresti, A. & Hitchcock, D. B. (2005), 'Bayesian inference for categorical data analysis', *Statistical Methods and Applications* **14**(3), 297–330.
URL: *https://doi.org/10.1007/s10260-005-0121-y*

Albert, J. H. & Chib, S. (1993), 'Bayesian Analysis of Binary and Polychotomous Response Data', *Journal of the American Statistical Association* **88**(422), 669–679.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476321*

Casella, G. & Berger, R. L. (2002), *Statistical inference*, 2nd ed edn, Thomson Learning, Australia ; Pacific Grove, CA.

Choi, H. M. (2014), Convergence analysis of Gibbs samplers for Bayesian regression models, Ph.D., University of Florida, United States – Florida. ISBN: 9781339148977.
URL: *http://search.proquest.com/docview/1727735455/abstract/9DD422F654E3470DPQ/1*

Choi, H. M. & Hobert, J. P. (2013), 'The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic', *Electronic Journal of Statistics* **7**, 2054–2064. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.
URL: *http://projecteuclid.org/euclid.ejs/1377005819*

Choi, H. M. & Hobert, J. P. (2016), 'A comparison theorem for data augmentation algorithms with applications', *Electronic Journal of Statistics* **10**(1), 308–329. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.
URL: *https://projecteuclid.org/euclid.ejs/1455715964*

Choi, H. M. & Román, J. C. (2017), 'Analysis of Polya-Gamma Gibbs sampler for Bayesian logistic analysis of variance', *Electronic Journal of Statistics* **11**(1), 326–337. Publisher: The Institute of Mathematical Statistics and the Bernoulli Society.
URL: *http://projecteuclid.org/euclid.ejs/1486458017*

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum Likelihood from Incomplete Data Via the EM Algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22.

Doksum, K. A. (2015), *Mathematical Statistics : Basic Ideas and Selected Topics, Volumes I-II Package*, Chapman and Hall/CRC.
URL: *https://www.taylorfrancis.com/books/mathematical-statistics-peter-bickel-kjell-doksum/10.1201/9781315369266*

Dyk, D. A. v. & Meng, X.-L. (2010), 'Cross-Fertilizing Strategies for Better EM Mountain Climbing and DA Field Exploration: A Graphical Guide Book', *Statistical Science* **25**(4), 429–449. Publisher: Institute of Mathematical Statistics.
URL: *http://projecteuclid.org/euclid.ss/1300108229*

Gelfand, A. E. & Smith, A. F. M. (1990), 'Sampling-Based Approaches to Calculating Marginal Densities', *Journal of the American Statistical Association* **85**(410), 398–409. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
URL: *http://www.jstor.org/stable/2289776*

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2020), *Bayesian Data Analysis, Third Edition*, CRC Press.
URL: *http://www.stat.columbia.edu/ gelman/book/*

Greenland, S. (2001), 'Putting Background Information about Relative Risks into Conjugate Prior Distributions', *Biometrics* **57**(3), 663–670. Publisher: [Wiley, International Biometric Society].
URL: *http://www.jstor.org/stable/3068401*

Hoff, P. D. (2009), *A First Course in Bayesian Statistical Methods*, Springer Texts in Statistics, Springer-Verlag, New York.
URL: *https://www.springer.com/us/book/9780387922997*

McLachlan, G. & Krishnan, T. (2008), *The EM Algorithm and Extensions.* OCLC: 845575460.
URL: *https://nbn-resolving.org/urn:nbn:de:101:1-201411014724*

Meng, X.-L. & van Dyk, D. (1999), 'Seeking efficient data augmentation schemes via conditional and marginal augmentation', *Biometrika* **86**(2), 301–320.
URL: *https://doi.org/10.1093/biomet/86.2.301*

Neal, R. M. (2003), 'Slice Sampling', *The Annals of Statistics* **31**(3), 705–741. Publisher: Institute of Mathematical Statistics.
URL: *http://www.jstor.org/stable/3448413*

O'Hagan, A. & Forster, J. J. (2004), *Kendall's Advanced Theory of Statistics, volume 2B: Bayesian Inference, second edition*, Vol. 2B, Arnold.
URL: *https://eprints.soton.ac.uk/46376/*

Piegorsch, W. W. & Casella, G. (1996), 'Empirical Bayes Estimation for Logistic Regression and Extended Parametric Regression Models', *Journal of Agricultural, Biological, and Environmental Statistics* **1**(2), 231–249. Publisher: [International Biometric Society, Springer].
URL: *http://www.jstor.org/stable/1400367*

Polson, N. G., Scott, J. G. & Windle, J. (2013), 'Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables', *Journal of the American Statistical Association* **108**(504), 1339–1349.
URL: *http://www.tandfonline.com/doi/abs/10.1080/01621459.2013.829001*

Shorten, C. & Khoshgoftaar, T. M. (2019), 'A survey on Image Data Augmentation for Deep Learning', *Journal of Big Data* **6**(1), 60.
URL: *https://doi.org/10.1186/s40537-019-0197-0*

Swendsen, R. H. & Wang, J.-S. (1987), 'Nonuniversal critical dynamics in Monte Carlo simulations', *Physical Review Letters* **58**(2), 86–88. publisher: American Physical Society.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.58.86*

Tanner, M. A. (1993), *Tools for Statistical Inference*, Springer Series in Statistics, Springer US, New York, NY.
URL: *http://link.springer.com/10.1007/978-1-4684-0192-9*

Tanner, M. A. & Wong, W. H. (1987), 'The Calculation of Posterior Distributions by Data Augmentation', *Journal of the American Statistical Association* **82**(398), 528–540. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].
URL: *http://www.jstor.org/stable/2289457*

Tierney, L. (1994), 'Markov Chains for Exploring Posterior Distributions', *The Annals of Statistics* **22**(4), 1701–1728. Publisher: Institute of Mathematical Statistics.
URL: *http://www.jstor.org/stable/2242477*

Zellner, A. & Rossi, P. E. (1984), 'Bayesian analysis of dichotomous quantal response models', *Journal of Econometrics* **25**(3), 365–393.
URL: *http://www.sciencedirect.com/science/article/pii/0304407684900071*