Non-parametric Bayesian Estimation of Mixture Binary Model

Junchen Feng

October 26, 2012

1 Motivation

1.1 Why Mixture Model

In the seminal Hirano, Imbens, Ridder(2003) paper[5], they claimed that series logit estimator is a consistent non-parametric estimator of the propensity score. Intuitively(from Hirano's slide), the argument reasons as the following:

Let $p(Y = 1|X) = m(x) = \frac{1}{1+e^{-k(x)}} = \Lambda(k(x))$. We can approximate m(x) to any degree by a polynomials, but the sum may not be bounded between $[0, 1]$.Since $\Lambda^{-1}(x) = -log(\frac{1}{y} - 1)$ is invertible, there is a definite mapping between k(x) and m(x). Instead of approximating m(x), we can approximate k(x) arbitrarily well by polynomials and then convert to m(x).

However, this argument does not work for mixture model because if the true data generating process is a mixture logit. Let it be a simple two component mixture:

$$p(Y = 1|X) = m'(x) = a_1 \frac{1}{1 + e^{-k_1(x)}} + (1 - a_1) \frac{1}{1 + e^{-k_2(x)}}$$

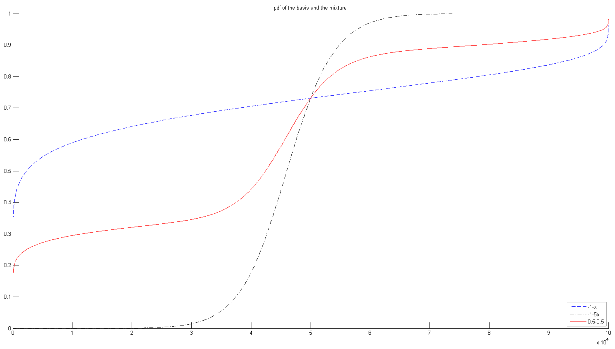

Figure 1: pdf of the mixture logit

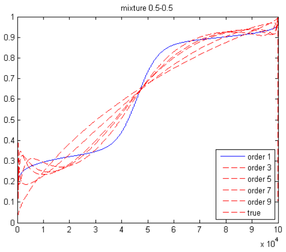

Figure 2: order 5 polynomial fit

Although $m'(x)$ will still have a inverse function$\Lambda'^{-1}(x)$, it will NOT be a logit function. To see this graphically, let's consider the following data generating process:

$$P(Y=1|X) = 0.5 * \frac{1}{1+e^{-1-10x}} + 0.5 * \frac{1}{1+e^{-1-0.5x}}$$

From the figure 2, one could tell that the mixture model has a different shape from the logit model, which graphically shows that the inverse function will not be a logit function. If one fits a five order polynomials, the bias is obvious. The

1.2 Why Bayesian Framework

It is all but natural to bring up MLE in estimating this model. However, there are a few of problems with MLE. The most critical concern is the potential multi-modality of the likehood, which breaches the fundamental identification assumption of the MLE. Of course one could try to combine MLE with grid search on the initial value, but how would one go about define the grid? Most of the times, we know very little about the behaviour of the likelihood.

Since EM is coming from the MLE family, the same critique also applies. Therefore, Heckman and Singer(1984)[6]'s proposal of the step function alike frequentist approach is no exception. Moreover, it gives a false sense of certainty about the number of mixture model, if one considers the possibility of being stuck in a local optima.

Bayesian framework won't be able to solve these problems per se. If the functional form is correctly specified and the MCMC runs infinitely long, all the problem above is solved since the whole parameter space will be covered. This is, of course, not realistic. However, it gives the analyst a honest assessment of the uncertainty of the models, if one uses a flexible prior to mitigate the functional form assumption and runs MCMC patient enough.

In the light of the argument above, I will introduce a Dirichlet process based Bayesian algorithm. The idea is first proposed by Ferguson(1973) [1] and later developed by Escobar(1994)[3] and Escobar and West(1995)[2].

2 Dirichlet Process

2.1 Description of the Process

To addresses the problem of Bayesian non-parametric estimation, Ferguson(1973) [1] proposed to model the data with the Dirichlet process.
The Dirichlet process has two parameters, A_0, G_0. G_0 is a probability measure and A_0 is a constant. G_0 is a belief of what G is. It is also the mean distribution of the process. It works like a location parameter. The parameter A_0 is the measure of strength in the above belief, acting like a dispersion parameter. [3]

Let π be a sequence of variables generated by the Dirichlet process A_0, G_0. It follows a general Polya urn scheme, namely:

$$\pi_1 \sim G_0$$

$$\pi_n|\pi_1,\cdots,\pi_{n-1}\{ \begin{array}{ll} =\pi_j & \text{with probability} \frac{1}{A_0+n-1} \\ \sim G_0 & \text{with probability} \frac{A_0}{A_0+n-1} \end{array}$$

The limiting distribution of the Dirichlet process will be uniform distribution on the set space verify with Ferguson. It thus ensures that there is positive probability to visit all elements of the set space. In another word, if the functional form is correctly specified, AND if we run the MCMC chain long enough, the correct mixture will be found. After the mixture is found, how quickly it populates the estimated parameter space depends on A_0.

Since A_0 affects the dispersion, the expected number of unique elements produced by the Dirichlet process is a function of A_0 (Antoniak, 1994)[4]

$$E(\text{Number of distinctive} \quad \pi|) = \sum_{i=1}^{n} \frac{A_0}{A_0+i-1} \sim A_0 ln(1+\frac{n}{A_0})$$

Sum up, assuming each draw from the process is i.i.d., the joint probability of the Dirichlet process $\{\pi_1, \pi_2, \cdots, \pi_n\}$ is

$$dF(\pi_1, \pi_2, \cdots, \pi_n) = \prod_{i=1}^{N} \frac{A_0 G_0 + \sum\limits_{j=1}^{i-1} \delta(\pi_j, \pi_i)}{A_0 + i - 1}$$

where

$$\delta(X, B) = \{ \begin{array}{ccc} 1 & \text{when} x & \in B \\ 0 & \text{when} x & \notin B \end{array}$$

2.2 Posterior Dirichlet Process

Let Y be the data whose density is $\phi(Y, \pi)$. By Bayes Law, the posterior density of the Dirichlet process is

$$f(\pi_i | \pi_j, j \neq i, Y) = \frac{\phi(Y, \pi_i) A_0 G_0 + \sum\limits_{j \neq k} \phi(Y, \pi_j) \delta(\pi_j, \pi_i)}{A(Y) + \sum\limits_{j \neq i} \phi(Y, \pi_i)}$$

where

$$A(Y) = \sum_{k=1}^{N} \phi(Y, \pi_k)(A_0 G_0)$$

The derivation of the posterior Dirichlet process is given in appendix A.

According to the posterior density, the drawing scheme is

$$\pi_i | \pi_j, j \neq i, Y \{ \begin{array}{ll} = \pi_j & \text{with probability} \frac{\phi(Y, \pi_j)}{A(Y) + \sum\limits_{j \neq k} \phi(Y, \pi_k)} \\ \sim h(\pi_i | Y) = \frac{A_0 G_0 \phi(Y, \pi_i)}{A(Y)} & \text{with probability} \frac{A(Y)}{A(Y) + \sum\limits_{j \neq k} \phi(Y, \pi_k)} \end{array}$$

3 Random Effect Model

Although the goal of the paper is set out to estimate a binary model, it is easier to use a normal model to illustrate the behavior of the Dirichlet Bayesian algorithm. Plus, logistic regression does not have a closed form expression under this setup, thus it is much difficult to program.

3.1 Model Set up

The random effect model is linear additive, which follows a conjugate normal inverse gamma set up.

$$
\begin{aligned}
Y &= Xb + \delta + \epsilon \\
\epsilon &\sim N(0, V) \\
b &\sim N(\mu, \tau V) \\
V &\sim IG(s, S)
\end{aligned}
$$

However, instead of drawing from a continuing distribution, such as normal or student t, the random effect δ is drawn from a discrete distribution (D). As Ferguson(1973) [1] demonstrated, by increasing the number of clusters, the Dirichlet process can approximate any distribution arbitrarily well. Therefore, it is a general solution to this class of problems.

The current model is normal conjugate so that the posterior Dirichlet process will have closed form expressions. Escobar(1994)[3] proposed a Monte Carlo Simulation method that allows for model set-up that does not yield a closed form solution.

3.2 Bayesian Estimation

The following setup is suggested by Escobar and West(1995)[2]. The element of the Dirichlet process π_i consists of the parameter for mean u_i and the parameter for variance V_i. The prior distribution of the Dirichlet process ($G(*)$) is a normal inverse gamma model

$$u_i|V_i \sim N(m, \tau V_i)$$
$$V_i^{-1} \sim G(\frac{s}{2}, \frac{S}{2})$$

The data follows a normal mean distribution

$$y_i \sim N(u_i, V_i^{-1})$$

The element is drawn from posterior process

$$\begin{aligned}
\pi_i &\sim q_0 G_i(\pi_i) + \sum_{j=1, j\neq i}^{n} q_j \delta_{\pi_j}(\pi_i) \\
G_i(\pi_i) &\sim N(m_i^*, V_i^*) IG(s_i^*, S_i^*) \\
&\text{where} \\
m_i^* &= \frac{m + \tau y_i}{1+\tau} \\
V_i^* &= \frac{\tau}{1+\tau} V_i \\
s_i^* &= \frac{s_i + 1}{2} \\
S_i^* &= \frac{S + \frac{(y_i - m)^2}{1+\tau}}{2} \\
q_0 &\propto at(m, \frac{(1+\tau)S}{s}, s) \\
q_j &\propto \phi(y_i, \mu_j, V_j) \\
q_0 + \sum_{j\neq i} q_j &= 1
\end{aligned}$$

The derivation of the posterior distribution is given on appendix B

Since the model allows the variance to be different among observations($V_i \neq V$), for the algorithm to converge, one has to model m and τ, which has a prior distribution of

$$\begin{aligned}
m &\sim N(a, A) \\
\tau^{-1} &\sim G(\frac{w}{2}, \frac{W}{2})
\end{aligned}$$

and a posterior distribution of

$$
\begin{aligned}
m &\sim N(a^*, A^*) \\
\tau^{-1} &\sim G(\frac{w^*}{2}, \frac{W^*}{2}) \\
&\text{where} \\
\bar{V}^{-1} &= \sum(V_j^{-1}) \\
A^* &= \frac{A}{A + \tau\bar{V}}\tau\bar{V} \\
a^* &= \frac{\tau\bar{V}}{A + \tau\bar{V}}a + \tau\bar{V}\sum(V_j^{-1}\mu_j) \\
w^* &= w + n \\
W^* &= W + \sum\frac{(u_j - m)^2}{V_j}
\end{aligned}
$$

The derivation of the posterior distribution is given in appendix C

3.3 Empirical Result

For simplicity, I begin the experiment with a simple normal mean model, following Escobar and West(1995) [2].

$$y_i = \mu_i + \epsilon$$

where μ_i is the random effect and drawn from a discrete distribution which is generated by a Dirichlet process.

I start the algorithm with a flat prior on the normal inverse gamma distribution.

$$\mu_i \sim N(0, 1000)$$
$$\frac{1}{\tau} \sim G(1, 1)$$

However, Escobar and West(1995) suggests initializing the draws of couplet μ_i, V_i from the posterior Dirichlet process based on prior distribution of normal inverse gamma model. It results in a under-dispersion of the initial elements. To correct the wrong distribution of the initial elements, it is necessary to draw more from the baseline distribution. Since the current algorithm does not learn the dispersion parameter A_0, I set it to be modestly large $A_0 = 5$, which enables the algorithm to learn the true clustering but resulted in an over-dispersion of the posterior draws.

In the following figure, I plot the true distribution of the random effect, the prior distribution of the random effect and the posterior distribution of random effect, result of a million long Markov Chain. Obviously, the algorithm is learning the true distribution, there are two clustering around the two groups of the random effect. However, the learning is far from perfect.

4 Appendix

4.1 A: Derivation of Posterior Dirichlet Process

$$\begin{aligned}
f(\pi_i|\pi_j, j \neq i, Y) &= \frac{f(Y|\pi_i, \pi_j, j \neq i) f(\pi_i|\pi_j, j \neq i)}{f(Y)} \\
&= \frac{f(Y|\pi_i, \pi_j, j \neq i) f(\pi_i|\pi_j, j \neq i)}{\sum_{k=1}^{N} [f(Y|\pi_k, \pi_j, j \neq k) f(\pi_k|\pi_j, j \neq k)]} \\
\sum_{k=1}^{N} [f(Y|\pi_k, \pi_j, j \neq k) f(\pi_k|\pi_j, j \neq k)] &= \frac{1}{A_0 + n - 1} \sum_{k=1}^{N} [\phi(Y, \pi_k)(A_0 G_0 + \sum_{j \neq k} \delta(\pi_j, \pi_k))]
\end{aligned}$$

Notice that

$$\begin{aligned}
\sum_{k=1}^{N} \sum_{j \neq k} \phi(Y, \pi_k) \delta(\pi_j, \pi_k) &= \sum_{j \neq k} \phi(Y, \pi_k) \sum_{k=1}^{N} \delta(\pi_j, \pi_k) \\
&= \sum_{j \neq k} \phi(Y, \pi_k) * 1
\end{aligned}$$

The density above is

$$\begin{aligned}
f(Y) &= \frac{1}{A_0 + n - 1} \sum_{k=1}^{N} [\phi(Y, \pi_k)(A_0 G_0 + \sum_{j \neq k} \delta(\pi_j, \pi_k))] \\
&= \frac{1}{A_0 + n - 1} [\sum_{k=1}^{N} \phi(Y, \pi_k)(A_0 G_0) + \sum_{j \neq k} \phi(Y, \pi_k)] \\
&= \frac{1}{A_0 + n - 1} [A(Y) + \sum_{j \neq k} \phi(Y, \pi_k)] \\
A(Y) &= \sum_{k=1}^{N} \phi(Y, \pi_k)(A_0 G_0)
\end{aligned}$$

4.2 B: Posterior Dirichlet Process of the Normal Mean Model

(1) $p(Y)$

$$\begin{aligned}
p(y) &= \int\limits_{0}^{\infty}\int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi V_i}}exp(-\frac{(y-u_i)^2}{2V_i})\frac{1}{\sqrt{2\pi\tau V_i}}exp(-\frac{(u_i-m)^2}{2\tau V_i})\frac{(\frac{S}{2})^{\frac{s}{2}}}{\Gamma(\frac{s}{2})}(\frac{1}{V_i})^{\frac{s}{2}-1}exp(-\frac{1}{V_i}\frac{S}{2})dV_i du_i \\
&= \int\limits_{0}^{\infty} \frac{(\frac{S}{2})^{\frac{s}{2}}}{\Gamma(\frac{s}{2})}\frac{1}{\sqrt{2\pi\frac{V_i}{\tau+1}}}(\frac{1}{V_i})^{\frac{s}{2}-1}exp[-\frac{1}{2}(\frac{(y-m)^2}{(\tau+1)}+S)\frac{1}{V_i}]\int\limits_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\frac{\tau}{1+\tau}V_i}}exp(-\frac{(u_i-y-\frac{m}{\tau})^2}{2\tau 1+\tau V_i})du_i dV_i \\
&= \frac{\Gamma(\frac{1+s}{2})(\frac{S}{2})^{\frac{s}{2}}(\frac{1}{2}(\frac{(y-m)^2}{(\tau+1)}+S))^{-\frac{s+1}{2}}}{\Gamma(\frac{s}{2})\sqrt(2(1+\tau)\pi)}\int\limits_{0}^{\infty}\frac{(\frac{1}{2}(\frac{(y-m)^2}{(\tau+1)}+S))^{\frac{s+1}{2}}}{\Gamma(\frac{s+1}{2})}(\frac{1}{V_i})^{\frac{s+1}{2}-1}exp[-\frac{1}{2}(\frac{(y-m)^2}{(\tau+1)}+S)\frac{1}{V_i}]dV_i \\
&= t(m,\frac{(1+\tau)S}{s},s)
\end{aligned}$$

(2) Posterior Distribution G_i

$$\begin{aligned}
p(\mu|Y,V) &= p(Y|\mu,V)p(\mu)p(V) \\
&\sim exp[-\frac{1}{2}(y-\mu)'(V)^{-1}(y-\mu)]exp[-\frac{1}{2}(\mu-m)'(\tau V)^{-1}(\mu-m)] \\
&\sim exp(-\frac{1}{2})exp[u'(\frac{\tau}{1+\tau}V)'(\frac{\tau}{1+\tau}V)^{-1}(\frac{\tau}{1+\tau}V)u-2(m'(\tau V)^{-1}+Y'v^{-1})(\frac{\tau}{1+\tau}V)^{-1}(\frac{\tau}{1+\tau}V \\
&\sim N(m^*,V^*) \\
V^* &= \frac{\tau}{1+\tau}V
\end{aligned}$$

$$m^* = \frac{m+\tau y}{\tau+1}$$

5 Appendix C: Posterior Distribution of the Hyper Parameters

(1) mean

$$\begin{aligned}
p(m|\tau,\pi) &= p(\pi|\tau^{-1},m)p(m)p(\tau^{-1}) \\
&\sim \prod_{j=1}^{N} exp[-\frac{1}{2}(\mu_i - m)'(\tau V_j)^{-1}(\mu_i - m)]exp[(m-a)A^{-1}(m-a)] \\
&\sim exp[-\frac{1}{2}m'(\sum_{j=1}^{N}\tau_j^{-1} + A^{-1})m - 2(\sum_{j=1}^{N}(\tau V_j)^{-1}\mu_j + A^{-1}a)'m]
\end{aligned}$$

Let $\bar{V}^{-1} = \sum_{j=1}^{N}\tau_j^{-1}$ and let $x = \frac{A}{A+\tau\bar{V}}$. Note that

$$\begin{aligned}
x\tau\bar{V} &= \frac{A\tau\bar{V}}{A+\tau\bar{V}} \\
&= (A^{-1} + \tau^{-1}\bar{V}^{-1})^{-1}
\end{aligned}$$

$$\begin{aligned}
p(m|\tau,\pi) &\sim exp[-\frac{1}{2}m'(x\tau\bar{V})(x\tau\bar{V})^{-1}(x\tau\bar{V})m - 2(\sum_{j=1}^{N}(\tau V_j)^{-1}\mu_j + A^{-1}a)'(x\tau\bar{V})(x\tau\bar{V})^{-1}m] \\
&\sim exp(-\frac{1}{2}(m - (\sum_{j=1}^{N}(\tau V_j)^{-1}\mu_j + A^{-1}a)'(x\tau\bar{V}))(x\tau\bar{V})^{-1}(m - (\sum_{j=1}^{N}(\tau V_j)^{-1}\mu_j + A^{-1}a)(x\tau\bar{V})))
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{j=1}^{N}(\tau V_j)^{-1}\mu_j + A^{-1}a)(x\tau\bar{V}) \qquad & = x\bar{V}\sum_{j=1}^{N}(V_j)^{-1} + (x\tau\bar{V})A^{-1}a \\
= x(\bar{V}\sum_{j=1}^{N}(V_j)^{-1}) + \frac{\tau\bar{V}}{\tau\bar{V}+A}a & \\
= x(\bar{V}\sum_{j=1}^{N}(V_j)^{-1}) + (1-x)a &
\end{aligned}$$

(2) variance

$$\begin{aligned}
p(\tau|m,\pi) = p(\pi|\tau^{-1},m)p(m)p(\tau^{-1}) & \\
& \sim (\frac{1}{\tau})^{\frac{k}{2}}\prod_{j=1}^{N} exp[-\frac{1}{2}(\mu_i - m)'(2\tau V_j)^{-1}(\mu_i - m)](\frac{1}{\tau})^{\frac{w}{2}-1} exp(-\frac{W}{2}\frac{1}{\tau}) \\
& \sim (\frac{1}{\tau})^{\frac{k}{2}+\frac{w}{2}-1} exp[(\sum_{j=1}^{N}\frac{(\mu_i - m)'(V_j)^{-1}(\mu_i - m)}{2} + \frac{W}{2})\frac{1}{\tau}]
\end{aligned}$$

References

[1] Thomas S. Ferguson,‘A Bayesian Analysis of Some Non-parametric Problems”, *The Annals of Statistics*, Vol. 1, No.2(Mar. 1973), 209-230

[2] Michael D. Escobar and Mike West, “Bayesian Density Estimation and Inference Using Mixture”, *Journal of the American Statistical Association*, Vol. 90, No.430(Jun., 1995), 577-588

[3] Michael D. Escobar, “Estimating Normal Means with a Dirichlet Process Prior”, *Journal of the American Statistical Association*, Vol. 89, No.425

(Mar. 1994), pp.268-277

[4] Charles E Antoniak, “Mixtures of Dirichlet Process with Application to Bayesian Nonparametric Problems”, *The Annals of Statistics, Vol.2, No.6(Nov., 1974)*, 1152-1174

[5] Keisuke Hirano, Guido W. Imbens and Geert Ridder, “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score”, Vol. 71, No. 4(Jul., 2003), 1161-1189

[6] J. Heckman and B. Singer, “A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data”, Vol. 52, No. 2(Mar., 1984), 271-320

[7] Dean A. Follmann and Diane Lambert,”Generalizing Logistic Regression by Nonparametric Mixing”, Vol. 84, No. 405,(Mar., 1989), 295-300