1  Bayesian Discrete Choice.

1. Quick Review of Bayes

2. Bayesian analysis of linear regression

3. Bayesian MNP

4. Target Marketing
2 Quick Review of Bayes.

- See Cameron and Trivedi for an overview of Bayes.

- In Bayesian econometrics, the econometrician acts as a rational decision maker, just like the agents in economic theory.

- The econometrician starts off with a prior distribution $p(\theta)$ about the model parameters.

- The econometrician observes some data $y = [y_1, \ldots, y_n]$.

- The econometrician has a model, $f(y|\theta)$ which is the probability of observing $y$ conditional on the parameters $\theta$.

- The econometrician's posterior probability by Bayes Theorem is:
\[ p(\theta|y) = \frac{p(\theta)f(y|\theta)}{\int p(\theta)f(y|\theta)d\theta} \]

- In the last decade, there has been an explosion in the applications of Bayesian approaches in statistics and econometrics.

- In Markov chain monte carlo, the econometrician simulates the posterior distribution \( p(\theta|y) \).

- This involves simulating a markov chain where the invariant (or long run) distribution is exactly equal to the posterior.

- The output of this simulation is a sequence of pseudo random numbers \( \theta^{(1)}, \ldots, \theta^{(S)} \).
Let \( f(\theta^{(1)}, \ldots, \theta^{(S)}) \) denote the density that puts weight \( \frac{1}{S} \) on each simulation draw.

Then under suitably regularity conditions

\[
f(\theta^{(1)}, \ldots, \theta^{(S)}) \rightarrow^d p(\theta \mid y)
\]

This posterior distribution expresses the econometricians beliefs about the parameters after seeing the data.

We could use the posterior to:

1. Construct a 95 percent credible set (i.e. a set that has 95 percent posterior probability). This is analogous to a confidence interval.

2. Simulate the distribution of functions of the parameters, \( g(\theta) \) as \( \frac{1}{S} \sum_s g(\theta^{(s)}) \).
3. Construct a predictive distribution for forecasting, i.e. simulate the model for each pseudo-random draw $\theta^{(1)}, \ldots, \theta^{(S)}$

- There are a few advantages to the Bayesian approach.

- The first is that it is very elegant numerically.

- There are some problems in latent variable, time series and other models that can only be solved through the use of Bayes.

- The second is that Bayes is exact in finite samples (up to our ability to approximate the posterior).

- Asymptotic theory depended on first order Taylor series expansion.
Essentially, you are hoping that the first and second derivatives capture the behavior of the function.

This can be a very poor approximation in finite samples in some cases.

In Bayes, such linearizations are not required.

However, you do need to be able to simulate the posterior accurately.

Third, Bayes fits into decision theory and can help to guide rational decision making.

For example, Rossi et. al. (1997) study the problem of target marketing.
• You see each household in a scanner panel data set choose a handful of times.

• You can use Bayes’ theorem to update on the random coefficients of each household i.

• e.g. if there are 10,000 households, you have a posterior for the preferences parameters for each of the 10,000 household conditional on its purchase history.

• Using this posterior, you can form your posterior beliefs about the profits from sending a coupon to an individual household.

• Fourth, in Bayes you can form posteriors over models.
• Suppose that you have $m = 1, \ldots, M$ probability models, $f_m(y|\theta)$.

• Then $f(y|\theta) = \sum_{m=1}^{M} p_m f_m(y|\theta)$ where $p_m$ is the probability of model $m$.

• You can use Bayes theorem to express your posterior probability $p_m$ about model $m$.

• This is a very elegant way to handle non-nested model and might be superior to classical approaches to non-nested testing.

• Finally, in Bayesian econometrics, you can work with models that are not identified or that do not exhibit normal asymptotics.

• A flat likelihood does not affect your ability to construct a posterior.
3 Markov Chains.

- Two common ways to conduct MCMC are Gibbs sampling and Metropolis.

- A normal random walk metropolis works as follows.

- First, the econometrician comes up with a rough guess $\theta^0$ at the MLE.

- Second, come up with a rough guess at $I^0$ at the information matrix using the hessian of the MLE.

- A sequence of psueorandom values $\theta^{(1)}, ..., \theta^{(S)}$ is drawn as follows. Given $\theta^{(s)}$, we draw $\theta^{(s+1)}$ as follows:
1. First, draw a candidate value \( \tilde{\theta} \sim N(\theta^{(s)}, I^0) \)

2. Second, compute \( \alpha = \min\{ \frac{p(\tilde{\theta})f(y|\tilde{\theta})}{p(\theta^{(s)})f(y|\theta^{(s)})}, 1 \} \)

3. Set \( \theta^{(s+1)} = \tilde{\theta} \) with probability \( \alpha \) and \( \theta^{(s+1)} = \theta^{(s)} \) with probability \( \alpha \).

• Implementing this algorithm simply requires the econometrician to evaluate the likelihood repeatedly and draw normal deviates.

• A second algorithm for constructing a Markov Chain is Gibbs sampling.

• Partition parameters into \( \theta_1, \ldots, \theta_d \) blocks
• Let \( p_k(\theta_k | \theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_d) \) denote the conditional distribution of the \( k \)th block of parameters given the others.

• In some applications, this distribution can be convenient to form even if the entire likelihood is quite complicated!

• Starting with an initial value \( \theta^0 \), Gibbs sampling works as follows. Given \( \theta^{(s)} \)

1. Draw \( \theta_1^{(s+1)} \sim p_1(\theta_1|\theta_2^{(s)}, \theta_2^{(s)}, \ldots, \theta_d^{(s)}) \)

2. Draw \( \theta_2^{(s+1)} \sim p_2(\theta_1|\theta_1^{(s+1)}, \theta_3^{(s)}, \theta_4^{(s)}, \ldots, \theta_d^{(s)}) \)

3. Draw \( \theta_3^{(s+1)} \sim p_3(\theta_1|\theta_1^{(s+1)}, \theta_2^{(s+1)}, \theta_4^{(s)}, \ldots, \theta_d^{(s)}) \)

\[ \vdots \]
d. Draw $\theta_d^{(s+1)} \sim p_d(\theta_1^{(s+1)}, \theta_2^{(s+1)}, \theta_4^{(s+1)} ..., \theta_{d-1}^{(s+1)})$

d+1 Return to 1.
4 Bayesian Analysis of Regression

- MCMC/Gibbs sampling are particularly powerful in problems with latent variables.

- In discrete choice, utility is latent to the econometrician.

- In a multinomial probit, if utility was observed by the econometrician, estimating parameters would boil down to linear regression.

- For our analysis, it will be useful to consider the Bayesian analysis of linear regression.

- A key step in Rossi et. al.’s paper essentially involve the Bayesian analysis of a normal linear regression.
• The analysis here follows Geweke’s textbook, Contemporary Bayesian Econometrics (an excellent intro to the subject).

• \( y \) is a \( T \times 1 \) vector of dependent variables

• \( X \) is a \( T \times k \) matrix of covariates (nonstochastic)

• Assume that error terms are normal, homoskedastic.

• \( y|\beta, h, X \sim N(X\beta, h^{-1}I_T) \)

\[
p(y|\beta, h, X) = (2\pi)^{-T/2}h^{T/2}\exp(-h(y-X\beta)'(y-X\beta)/2))
\]

• \( h \) is called the precision parameter (inverse of variance)
• $I_T$ is a scalar covariance matrix

• This is our likelihood function.

• Recall that Bayes theorem implies that the posterior distribution of the model parameters is the prior times the likelihood.

• We need to specify a prior distribution for our model parameters.

• $p(\beta) = N(\beta, H^{-1})$

$$p(\beta) = (2\pi)^{-k/2}|H|^{1/2} \exp\left(-((\beta - \beta)'H^{-1}(\beta - \beta))\right)$$

• The prior on beta is normal with prior mean, $\beta$ and prior precision $H$
• The prior distribution on $h$ is $\sigma^2 h \sim \chi^2(\nu)$

$$p(h) = \left[ 2^{\nu/2} \Gamma(\nu/2) \right]^{-1} (\sigma)^{\nu/2} h^{(\nu-2)/2} \exp(-\sigma^2 h/2)$$

• Remark- this is essentially a gamma distribution, rewritten in a manner that will be convenient for reasons below.

• This form for the prior is chosen because of conjugacy, i.e. the posterior distribution can be written in an analytically convenient manner.

• Now recall that the posterior is proportional to the prior time the likelihood.

• That is, $p(\beta, h|X, y) \propto p(\beta)p(h)p(y|\beta, h, X)$
• Combining the equations above yields $p(\beta, h|X, y) \propto$:

$$
(2\pi)^{-T/2} h^{T/2} \left[ \frac{2^\nu/2 \Gamma(\nu/2)}{\nu/2} \right]^{-1} |H|^{1/2} \\
(s)^{\nu/2} h^{(T+\nu-2)/2} \exp(-s^2 h/2) \\
\exp \left[ -(\beta - \beta')' H^{-1} (\beta - \beta)/2 \\
- h(y - X\beta)'(y - X\beta)/2 \right]
$$

• Now recall from Cameron and Trivedi, the idea behind Gibbs sampling is to ”block” the parameters into a set of convenient conditional distributions.

• In this case, we will want to block $p(\beta|h, X, y)$ and $p(h|\beta, X, y)$

• Let’s first derive $p(\beta|h, X, y)$. 
• It is obvious from the above expression, that \( \beta \) is going to be normally distributed.

• We will want to complete the the square inside the \( \exp() \) to rewrite the expression in the form
\[
(\beta - \tilde{\beta})' \tilde{H}^{-1}(\beta - \tilde{\beta})
\]

• Then \( \tilde{\beta} \) will be the posterior mean and \( \tilde{H} \) the posterior precision

• Distributing terms and completing the square in \( \beta \) yields that:
\[
(\beta - \bar{\beta})' \bar{H}^{-1}(\beta - \bar{\beta}) + h(y - X\beta)'(y - X\beta) = (\beta - \tilde{\beta})' \tilde{H}^{-1}(\beta - \tilde{\beta}) + h(y^0 - X\beta)'(y^0 - X\beta)
\]
where \( y^0 = \) fitted ols values
\[
= (\beta - \tilde{\beta})' \tilde{H}^{-1}(\beta - \tilde{\beta}) + Q
\]
\[
\tilde{H} = \bar{H} + hX'X
\]
\[
\tilde{\beta} = \tilde{H}^{-1}(\bar{H}^{-1}\bar{\beta} + hX'Xb)
\]
where \( b \) is the ols estimate of \( \beta \)

\( Q \) is a constant that does not depend on \( \beta \)
• Note that the posterior precision is a weighted average of the prior precision and the ols estimate of the precision

• The posterior estimate of $\tilde{\beta}$ involves a weighted average of the prior mean and the ols estimate.

• The weights depend on the posterior precision, the prior precision and the ols estimate of the precision.

• As the sample size becomes sufficiently large, the data will "swamp" the prior.

• The number of terms in the likelihood is a function of $T$ and grows with the sample size

• The number of terms in the prior remains fixed.
• Next, we have to derive the posterior in $h$

• If you look at the prior times the likelihood, it is obvious that the posterior will be of the form $p(h|\beta, X, y)$ of $h^\alpha \exp(-h\omega)$

• If we can derive $\alpha$ and $\omega$ we can express the posterior

• By some straightforward, albeit tedious, algebra we can write the posterior distribution as $p(h|\beta, X, y)$:

\[
\tilde{s}^2 h \sim \chi^2(\tilde{v}) \\
\tilde{s} = s + (y^0 - X\beta)'(y^0 - X\beta) \\
\tilde{v} = v + T
\]

• A Gibbs sampler generates a pseudo-random sequence $(h^{(s)}, \beta^{(s)})_{s = 1, \ldots, S}$ using the following
markov chain,

1. Given \((h^{(s)}, \beta^{(s)})\), draw \(\beta^{(s+1)} \sim p(\beta|h^{(s)}, X, y)\)

2. Given \(\beta^{(s+1)}\) draw \(h^{(s+1)} \sim p(h|\beta^{(s+1)}, X, y)\)

3. Return to 1

- This example illustrates the importance of conjugacy.

- That is, choosing our prior and likelihoods to be the "right" functional form greatly simplifies the analysis of the posterior distribution.

- Standard texts in Bayesian statistics (e.g. Berger/Bernardo and Smith) have appendices that lay out conjugate distributions.
The multinomial probit is closely related to the analysis of the normal linear model.

The multinomial probit is defined as:

\[ y_{ij} = x_{ij} \beta + \varepsilon_{ij} \]  
\[ \text{var}(\varepsilon_{i1}, ..., \varepsilon_{iJ}) = h^{-1}I_J \]  
\[ c_{ij} = 1\{y_{ij} > y_{ij'} \text{ for } j' \neq j\} \]

In the above, \( y_{ij} \) is the utility of person \( i \) for alternative \( j \)

\( \varepsilon_{ij} \) is the stochastic preference shock

\( x_{ij} \) are covariates that enter into \( i \)'s utility

\( c_{ij} = 1 \) if \( i \) chooses \( j \)
If the $y_{ij}$ were known, then we could use the Gibbs sampler above to estimate $\beta$ and $h$.

However, the $y_{ij}$ are latent variables and therefore we do data augmentation.

The idea behind data augmentation is simple— we integrate out the distribution of the variables that we do not see.

Following the notation in Cameron and Trivedi, let $f(\theta|y, y^*)$ denote the posterior conditional on the observed variables, $y$ and the latent variables, $y^*$.

Let $f(y^*|y, \theta)$ denote the distribution of the latent variable conditional on $y$ and parameters.

Then the posterior can be written as:

$$p(\theta|y) = \int f(\theta|y, y^*)f(y^*|y, \theta)dy^*$$
• Taking account of the latent variable simply involves an additional Gibbs step.

• The distribution of the latent utility $y_{ij}$ is a truncated normal distribution.

• If $c_{ij} = 1$, $y_{ij}$ is a truncated normal with mean parameter $\beta$, precision $h$ and lower truncation point $\max\{y_{ij'}, j' \neq j\}$.

• If $c_{ij} = 0$, $y_{ij}$ is a truncated normal with mean parameter $\beta$, precision $h$ and upper truncation point $\max\{y_{ij}\}$.

• The Gibbs sampler for the multinomial probit simply adds the data augmentation step above:
• A Gibbs sampler generates a pseudo-random sequence \((h^{(s)}, \beta^{(s)}, \left\{ y_{ij}^{(s)} \right\}_{i \in I, j \in J})\) \(s = 1, ..., S\) using the following markov chain

1. Given \((h^{(s)}, \beta^{(s)})\), draw \(\beta^{(s+1)} \sim p(\beta|h^{(s)}, X, y, C)\)

2. Given \(\beta^{(s+1)}\) draw \(h^{(s+1)} \sim p(h|\beta^{(s+1)}, X, y, C)\)

3. For each \(I\), draw \(y_{i1}^{(s+1)} \sim p(h|\beta^{(s+1)}, X, y_{i2}^{(s)}, ..., y_{iJ}^{(s)}, C)\)

4. Draw \(y_{i2}^{(s+1)} \sim p(h|\beta^{(s+1)}, X, y_{i1}^{(s+1)}, ..., y_{iJ}^{(s)}, C)\)

5. ...

6. Draw \(y_{iJ}^{(s+1)} \sim p(h|\beta^{(s+1)}, X, y_{i1}^{(s+1)}, ..., y_{iJ-1}^{(s+1)}, C)\)

7. Return to 1
6 Target Marketing

- In "The Value of Purchase History Data in Target Marketing" Rossi et. al. attempt to estimate household level preference parameters.

- This is of interest as a marketing problem.

- CMI checkout coupon uses purchase information to customize coupons to a particular household.

- In principal, the entire purchase history (from consumer loyalty cards) could be used to customize coupons (and hence prices)

- If a household level preference parameter can be forecasted with high precision, this is essentially first degree price discrimination!
• Even with short purchase histories, they find that profits are increased 2.5 fold through the use of purchase data compared to blanket couponing strategies.

• Even one observation can boost profits from couponing by 50%.

• This application is of interest to economists as well.

• The methods in this paper allow us to account for consumer heterogeneity in a very rich manner.

• This might be useful to examine the distribution of welfare consequences of a policy intervention (e.g. a merger or market regulation).

• Beyond that, these methods demonstrate the power of Bayesian methods in latent variable problems.
7 Random Coefficients Model

- Multinomial probit with panel data on household level choices

\[ y_{h,t} = X_{h,t} \beta_h + \varepsilon_{h,t} \]
\[ \varepsilon_{h,t} \sim N(0, \Lambda) \]
\[ \beta_h = \Delta z_h + v_h, \quad v_h \sim N(0, V_\beta) \]

- Households \( h = 1, ..., H \) and time \( t = 1, ..., T \)

- \( X_{h,t} \) covariates and \( z_h \) demographics

- Note that household specific random coefficients \( \beta_h \) remain fixed over time

- \( I_{h,t} \) observed choice
• The posterior distributions are derived in Appendix A.

• Formally, the derivations are very close to our multinomial probit model above.

• Gibbs sampling is used to simulate the posterior distribution of $\Lambda, \Delta, V_\beta$

8 Predictive Distributions

• The authors wish to give different coupons to different households.

• A rational (Bayesian) decision maker would form her beliefs about household $h$’s preference parameters given her posterior about the model parameters.
• This will involve, as we show below, forming a predictive distribution for $\beta_h$ given the econometrician’s information set.

• As a first case, suppose that the econometrician only knew $z_h$, the demographics of household $h$.

• From our model, $p(\beta_h|z_h, \Delta, V_\beta)$ is $N(\Delta z_h, V_\beta)$.

• Given the posterior $p(\Delta, V_\beta|Data)$, the econometrician’s predictive distribution for $\beta_h$ is:

$$p(\beta_h|z_h, Data) = \int p(\beta_h|z_h, \Delta, V_\beta)p(\Delta, V_\beta|Data)$$

• We can simulate $p(\beta_h|z_h, Data)$ using Gibbs sampling given our posterior simulations $\Delta^{(s)}, V_\beta^{(s)}$, $s = 1, \ldots, S$:

$$\frac{1}{S} \sum_{s=1}^{S} p(\beta_h|z_h, \Delta^{(s)}, V_\beta^{(s)})$$
• We could draw random $\beta_h$ from $p(\beta_h|z_h, \text{Data})$.

• For each $\Delta^{(s)}$, $V^{(s)}_{\beta}$, draw $\beta^{(s)}_h$ from $p(\beta_h|z_h, \Delta^{(s)}, V^{(s)}_{\beta})$.

• Given $\beta^{(s)}_h$, $s = 1, \ldots, S$, we could then simulate purchase probabilities.

• Draw $\varepsilon^{(s)}_{ht}$ from $\varepsilon_{h,t} \sim N(0, \Lambda^{(s)})$.

• The posterior purchase probability for $j$, given $X_{ht}$ and $z_h$ is:

$$
\frac{1}{S} \sum_s 1 \left\{ X_{jht} \beta_h + \varepsilon^{(s)}_{jht} > X_{j'ht} \beta_h + \varepsilon^{(s)}_{j'ht} \text{ for } j' \neq j \right\}
$$

• This would allow us to simulate the purchase response to different couponing strategies for a specific household $h$. 
• The paper runs through different couponing strategies given different information set (e.g. full or choice only information sets).

• The key ideas are similar- form a predictive distribution for h’s preferences and simulate purchase behavior in an analogous fashion.

• In the case of a full purchase information history, we could use the raw Gibbs output since the markov chain will simulate $\beta_h^{(s)}$ $s = 1, ..., S$.

• This could then be used to simulate choice behavior as in the example above (given draws of $\varepsilon_{ht}^{(s)}$).
9 Data

• AC Neilson scanner panel data for tuna in Springfield Missouri.

• 400 households, 1.5 years, 1-61 purchases.

• Brands and covariates in Table 2.

• Demographics Table 3.

• Table 4, delta coefficients.

• Poorer people prefer private label.

• Goodness of fit moderate for demographic coefficients.
• Figures 1 and 2, household level coefficient estimates with different information sets

• Table 5, return to different marketing strategies.

• Bottom line, you gain .5 cents to 1.0 cents per customer through better estimates.

• With a lot of customers, this could be quite profitable.