

Unveiling the Ways of Bayes

An overview of Bayes estimators for choice models



Ricardo A Daziano

daziano@cornell.edu

Cornell University

Contents

- 1 Elements of Bayesian econometrics
- 2 Gibbs sampling for probit models
- 3 Role of the prior exemplified
- 4 Gibbs sampler of ICLV
- 5 Metropolis Hastings for logit models
- 6 Bayesian confidence regions
- 7 MH within Gibbs for mixed logit



Going Bayesian? Why?

- Bayes estimators are gradient free



Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free



Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free
- Actually, no maximization is involved (most of the time)



Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free
- Actually, no maximization is involved (most of the time)
- Interesting for non-convex likelihood functions and weakly identified models



Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free
- Actually, no maximization is involved (most of the time)
- Interesting for non-convex likelihood functions and weakly identified models
- Works particularly well for latent variables (hybrid choice models, missing data, **utility**)



Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free
- Actually, no maximization is involved (most of the time)
- Interesting for non-convex likelihood functions and weakly identified models
- Works particularly well for latent variables (hybrid choice models, missing data, **utility**)
- Asymptotic properties coincide with MLE

Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free
- Actually, no maximization is involved (most of the time)
- Interesting for non-convex likelihood functions and weakly identified models
- Works particularly well for latent variables (hybrid choice models, missing data, **utility**)
- Asymptotic properties coincide with MLE
- Simulation-aided inference (repeated sampling), avoiding MSLE bias



Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free
- Actually, no maximization is involved (most of the time)
- Interesting for non-convex likelihood functions and weakly identified models
- Works particularly well for latent variables (hybrid choice models, missing data, **utility**)
- Asymptotic properties coincide with MLE
- Simulation-aided inference (repeated sampling), avoiding MSLE bias
- Bayes works for **small samples**, and data that are not samples



Going Bayesian? Why?

- Bayes estimators are gradient free
- Hessian free
- Actually, no maximization is involved (most of the time)
- Interesting for non-convex likelihood functions and weakly identified models
- Works particularly well for latent variables (hybrid choice models, missing data, **utility**)
- Asymptotic properties coincide with MLE
- Simulation-aided inference (repeated sampling), avoiding MSLE bias
- Bayes works for **small samples**, and data that are not samples
- Neat intuition: scientific method + interval estimation problem + decision making theory



Uncertainty

- Uncertainty, broadly defined, accounts for a world that is probabilistic in nature
- The study of uncertainty has become a paramount topic for several fields in economics, statistics, and psychology
- Decision-making process of consumers: we need to take into account **behavioral uncertainty** (random utility)
- The introduction of Bayesian tools adds another dimension to handling uncertainty

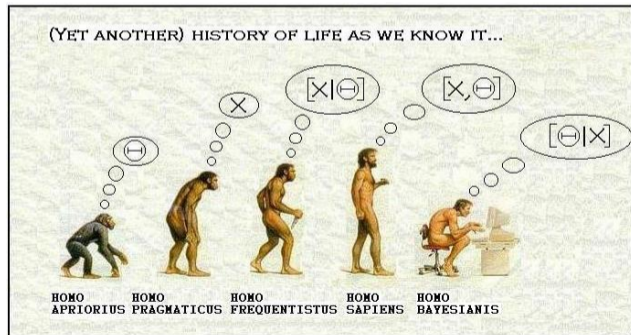


The Bayesian approach

- We recognize that we are **uncertain** about the **true state of the world** (which is expressed by the true parameters of the econometric model)
- **Frequentist** (classic) approach: true parameters are fixed but unknown constants
- Bayesian: true parameters are **random variables**
- This notion is fundamental for Bayesian inference and is derived from the concept of **subjective probabilities**
 - Beliefs about the occurrence of a particular event (probability laws under uncertainty)

The Bayesian approach cont'd

Bayesian approach: updating our vision of the world in the light of new evidence (learning)

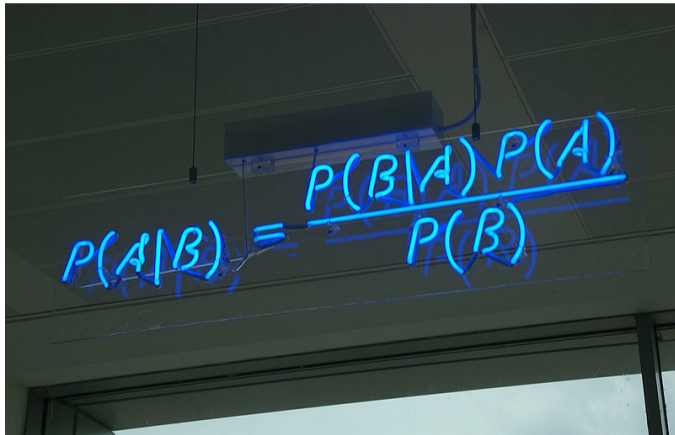


Bayesian inference

- Parameters of a model $(\mathcal{Y}, \mathcal{P} = P_{\theta})$ are assumed to have a **prior** statistical distribution $p(\theta)$
- $p(\theta)$ describes the probability distribution of θ before the observation of \mathbf{y}
- The combination of the **prior distribution** $p(\theta)$ with the information coming in via the sample data $\mathbf{y} \in \mathcal{Y}$ determines the **posterior distribution** of the parameters $p(\theta|\mathbf{y})$

Inference about the parameters in Bayesian econometrics: we can introduce **prior knowledge** or beliefs and apply the rules of probability directly

Bayes' theorem


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' theorem

The posterior and prior distributions are related following **Bayes' theorem** according to

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

Note that $p(\mathbf{y}|\boldsymbol{\theta}) \equiv \ell(\mathbf{y}; \boldsymbol{\theta})$ by definition.

Since $p(\mathbf{y})$ is constant, for inference purposes Bayes' theorem is rewritten as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

The prior distribution $p(\theta)$

- A prior reflects **knowledge** and **beliefs** (notion of subjective probability)
- Priors are usually chosen inside a family of parametric probability distributions
 - The posterior distribution may become the prior for a subsequent problem
 - Recall that $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$
 - **Conjugate family of distributions**: the chosen prior is such that the posterior falls within the same family
- The use of Bayesian inference is particularly interesting for **small samples** where the role of the prior distribution is potentially relevant

The prior distribution $p(\theta)$ cont'd

- In general, even for small samples the relative importance of the prior distribution is proportional to its **precision**
 - The effect of the prior gradually disappears as the prior variance increases
 - The importance of the prior distribution disappears as the sample size increases
- 1 A **diffuse** or **noninformative prior** is a distribution that is widely dispersed
 - 2 A flat prior distribution with an infinite integral is called an **improper prior**

Conjugacy

- **We know the exact distribution of the posterior:** inference based on generation of random numbers!
- (Equivalent of closed-form solution)

Posterior	Likelihood	Prior
Beta	Binomial	Beta
Beta	Negative Binomial	Beta
Gamma	Poisson	Gamma
Gamma	Exponential	Gamma
Beta	Geometric	Beta
Normal	Normal (unknown mean)	Normal
Inverse Gamma	Normal (unknown variance)	Inverse Gamma
Normal/Gamma	Normal (unknown mean & variance)	Normal/Gamma
Dirichlet	Multinomial	Dirichlet



Let's start simple: binary probit

- (And let's make sure we adopt the same notation)
- Random utility of 2 alternatives for consumer i :

$$U_{i1} = \mathbf{x}'_{i1}\boldsymbol{\beta} + \varepsilon_{i1}$$

$$U_{i2} = \mathbf{x}'_{i2}\boldsymbol{\beta} + \varepsilon_{i2}$$

- In a probit model, we can work with a one-dimensional normally distributed latent variable:

$$u_i \sim \mathcal{N}((\mathbf{x}_{i1} - \mathbf{x}_{i2})'\boldsymbol{\beta}, 1)$$

- Stacking individuals and considering a design matrix in differences:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$$

Choice indicator and probabilities

- We don't observe u_i but $y_i \sim \text{Bernoulli}(P_{i1})$
- Choice probability (alternative 1):

$$P_{i1} = \text{Pr}(U_{i1} \geq U_{i2}) = \Phi((\mathbf{x}_{i1} - \mathbf{x}_{i2})' \boldsymbol{\beta})$$

- Likelihood:

$$\ell(\boldsymbol{\beta}; \mathbf{y} | \mathbf{X}) = \prod_{i=1}^N [\Phi((\mathbf{x}_{i1} - \mathbf{x}_{i2})' \boldsymbol{\beta})]^{y_i} [1 - \Phi((\mathbf{x}_{i1} - \mathbf{x}_{i2})' \boldsymbol{\beta})]^{1-y_i}$$

- Frequentist: we find $\hat{\boldsymbol{\beta}}_{\text{MLE}} = \arg \max \ell(\boldsymbol{\beta}; \mathbf{y} | \mathbf{X})$

Toward Bayesian inference of the binary probit

- Start with the likelihood, different notation: $\ell(\boldsymbol{\beta}; \mathbf{y} | \mathbf{X}) = p(\mathbf{y} | \boldsymbol{\beta}, \mathbf{X})$
- Add prior $p(\boldsymbol{\beta})$, for example a **normal prior**: $p(\boldsymbol{\beta}) \sim \mathcal{N}(\check{\boldsymbol{\beta}}, \check{V}_{\boldsymbol{\beta}})$

$$p(\boldsymbol{\beta}) = (2\pi)^{-\frac{k}{2}} |\check{V}_{\boldsymbol{\beta}}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \check{\boldsymbol{\beta}})' \check{V}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \check{\boldsymbol{\beta}})\right)$$

- Posterior inference: $p(\boldsymbol{\beta} | \mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\beta})p(\mathbf{y} | \boldsymbol{\beta}, \mathbf{X})$
- Sometimes $p(\boldsymbol{\beta})p(\mathbf{y} | \boldsymbol{\beta}, \mathbf{X})$ is known (conjugacy)
- Sometimes we don't find a known posterior



If we knew \mathbf{u} then we find conjugacy

- In this hypothetical case, the posterior of interest is $p(\boldsymbol{\beta}|\mathbf{u}, \mathbf{X})$
- Bayesian inference: $p(\boldsymbol{\beta}|\mathbf{u}) \propto p(\boldsymbol{\beta})p(\mathbf{u}|\boldsymbol{\beta}, \mathbf{X})$
- Normal prior (same as before): $p(\boldsymbol{\beta}) \sim \mathcal{N}(\check{\boldsymbol{\beta}}, \check{\mathbf{V}}_{\boldsymbol{\beta}})$
- If \mathbf{u} were observed, we find a normal likelihood:

$$p(\mathbf{u}|\boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^N (2\pi)^{-1/2} \exp\left(-\frac{1}{2}(u_i - \mathbf{x}'_i\boldsymbol{\beta})'(u_i - \mathbf{x}'_i\boldsymbol{\beta})\right)$$

Binary probit: Deriving the posterior (aka algebra)

- Prior times likelihood:

$$\propto p(\boldsymbol{\beta})p(\mathbf{u}|\boldsymbol{\beta}, \mathbf{X})$$

$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \check{\boldsymbol{\beta}})' \check{\mathbf{V}}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \check{\boldsymbol{\beta}})\right) \prod_{i=1}^N \exp\left(-\frac{1}{2}(u_i - \mathbf{x}'_i \boldsymbol{\beta})'(u_i - \mathbf{x}_i \boldsymbol{\beta})\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[(\boldsymbol{\beta} - \check{\boldsymbol{\beta}})' \check{\mathbf{V}}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \check{\boldsymbol{\beta}}) + (\mathbf{u} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{u} - \mathbf{X}\boldsymbol{\beta})\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\beta}' \check{\mathbf{V}}_{\boldsymbol{\beta}}^{-1} + \mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'(\check{\mathbf{V}}_{\boldsymbol{\beta}}^{-1}\check{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{u}) - (\dots)\boldsymbol{\beta}\right]\right)$$

- (We can drop additive terms that do not depend on $\boldsymbol{\beta}$)

Binary probit: Deriving the posterior

Consider:

- $\hat{V}_\beta = (\check{V}_\beta^{-1} + \mathbf{X}'\mathbf{X})^{-1}$

- $\hat{\beta} = \hat{V}_\beta (\check{V}_\beta^{-1}\check{\beta} + \mathbf{X}'\mathbf{u})$

$$\propto \exp\left(-\frac{1}{2}\left[\beta' \hat{V}_\beta^{-1} \beta - \beta' (\check{V}_\beta^{-1} \check{\beta} + \mathbf{X}'\mathbf{u}) - (\check{V}_\beta^{-1} \check{\beta} + \mathbf{X}'\mathbf{u}) \beta\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\beta' \hat{V}_\beta^{-1} \beta - \beta' \hat{V}_\beta^{-1} \hat{V}_\beta (\check{V}_\beta^{-1} \check{\beta} + \mathbf{X}'\mathbf{u}) - \hat{V}_\beta^{-1} \hat{V}_\beta (\check{V}_\beta^{-1} \check{\beta} + \mathbf{X}'\mathbf{u}) \beta\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\beta' \hat{V}_\beta^{-1} \beta - \beta' \hat{V}_\beta^{-1} \hat{\beta} - \hat{\beta}' \hat{V}_\beta^{-1} \beta\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\beta' \hat{V}_\beta^{-1} \beta - \beta' \hat{V}_\beta^{-1} \hat{\beta} - \hat{\beta}' \hat{V}_\beta^{-1} \beta + \hat{\beta}' \hat{V}_\beta^{-1} \hat{\beta}\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}(\beta - \hat{\beta})' \hat{V}_\beta^{-1} (\beta - \hat{\beta})\right)$$

We found conjugacy!

- normal prior \times normal likelihood = **normal posterior**
- In this case:

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \hat{V}_{\boldsymbol{\beta}})$$

- Posterior draws can be generated from this known normal
- Posterior mean (\equiv Bayes point estimate):

$$\hat{\boldsymbol{\beta}} = (\check{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{X}'\mathbf{X})^{-1} (\check{V}_{\boldsymbol{\beta}}^{-1}\check{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{u})$$

- Do you recognize the posterior mean? (think of your econometrics course)

It's a (Bayesian) regression model!

- With a diffuse prior (large prior variance), the posterior mean becomes:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$$

- In fact, for an ordinary regression $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ with a diffuse prior, the posterior mean is $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- This result is not a surprise, we are assuming that **we know \mathbf{u}**
- This is where Bayesian concepts strike again: we will use a trick called **data augmentation**

Going back to the model + a useful fact

- We know that $\mathbf{u} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I})$
- But observed choices constrain the values of \mathbf{u}
 - 1 If $y_i = 1$, then $u_i = U_{i1} - U_{i2} > 0$
 - 2 If $y_i = 0$, then $u_i \leq 0$
- The distribution of $p(\mathbf{u}|\mathbf{y}, \mathbf{X}, \boldsymbol{\beta})$ is a **truncated normal**
 - 1 $u_i \sim \mathcal{N}(\mathbf{x}'_i\boldsymbol{\beta}, 1)\mathbb{I}(u_i > 0)$ if $y_i = 1$
 - 2 $u_i \sim \mathcal{N}(\mathbf{x}'_i\boldsymbol{\beta}, 1)\mathbb{I}(u_i \leq 0)$ if $y_i = 0$
- How can we use this?



Data augmentation

- A **latent variable** can be treated as an **additional parameter**
- **u** is now a parameter, meaning that now our posterior is $p(\beta, \mathbf{u} | \mathbf{y}, \mathbf{X})$
- There is no conjugacy for this new posterior

Full conditional distributions

① $p(\beta | \mathbf{u}, \mathbf{y}, \mathbf{X})$

② $p(\mathbf{u} | \beta, \mathbf{y}, \mathbf{X})$

- If each of the full conditional distributions have known priors (conjugacy) we can use **Gibbs sampling**

Gibbs sampler for binary probit

- Start with arbitrary \mathbf{u}
- At iteration (g):
 - 1 Update $\beta^{(g+1)}$ by drawing from $p(\beta|\mathbf{u}^{(g)}, \mathbf{y}, \mathbf{X})$ (**normal**)
 - 2 Update $\mathbf{u}^{(g+1)}$ by drawing from $p(\mathbf{u}|\beta^{(g+1)}, \mathbf{y}, \mathbf{X})$ (**truncated normal**)
- Repeat G times, after a preset burn-in period (discarding initial draws)
- The sequence $\{\beta^{(g)}\}$ are draws from its posterior
- Bayes point estimate:

$$\hat{\beta}_{\text{Bayes}} = \frac{1}{G} \sum_{g=1}^G \beta^{(g)}$$

A Swiss example

Axhausen et al. (2008) collected data on transportation choices in Switzerland using two discrete choice experiments. **SP 1** was a mode choice experiment (car/rail), whereas **SP 2** was an unlabelled rail route choice experiment:

Mode choice car – rail (main study version)

Travel costs:	18 Fr.	Travel costs:	23 Fr.
Total travel time:	40 minutes	Travel time:	30 minutes
... congested:	10 minutes	Headway:	30 minutes
... uncongested:	30 minutes	No. of changes:	0 times

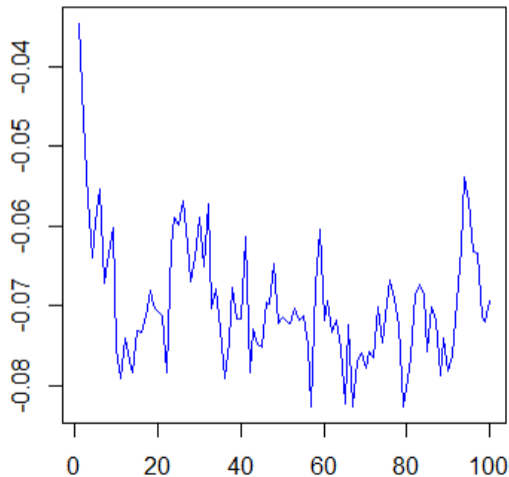
← Your choice →

Route choice rail (main study version)

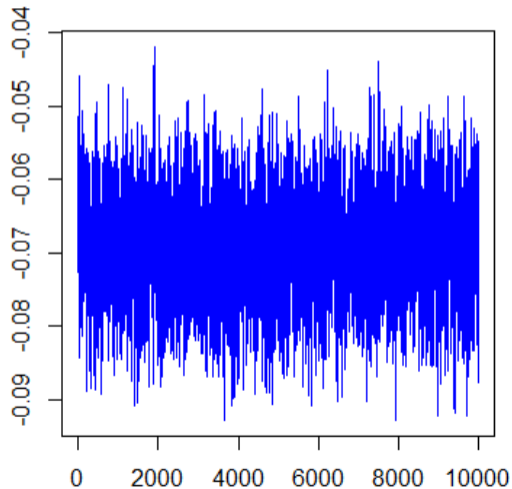
Travel costs:	20 Fr.	Travel costs:	23 Fr.
Travel time:	40 minutes	Travel time:	30 minutes
Headway:	15 minutes	Headway:	30 minutes
No. of changes:	1 times	No. of changes:	0 times

← Your choice →

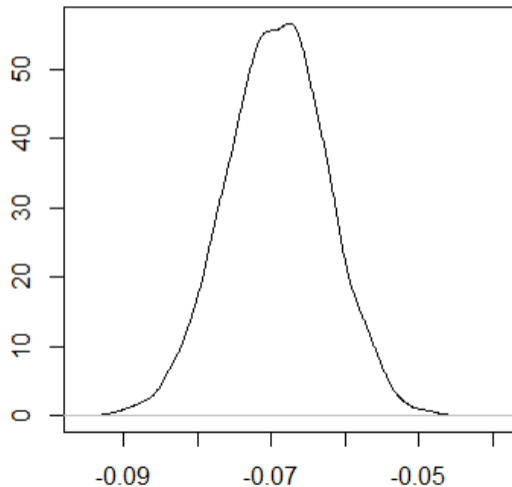
Binary probit rail route choice: burn-in draws travel cost



Binary probit rail route choice: converged draws travel cost



Binary probit rail route choice: posterior density travel cost



Binary probit rail route choice: frequentist vs Bayesian

Table: Binary probit estimates, Swiss value of time data

	Bayesian estimates		Classical estimates	
	post. mean	post. st. dev	point estimate	s.e.
ASC	-0.014	0.025	-0.014	0.025
TC	-0.069	0.007	-0.069	0.007
TT	-0.033	0.002	-0.033	0.002
HW	-0.022	0.001	-0.022	0.001
CH	-0.667	0.023	-0.666	0.024

- Diffuse prior ensures results are similar between Bayes and classical

Binary probit, prior influence on the posterior of β_{TC}

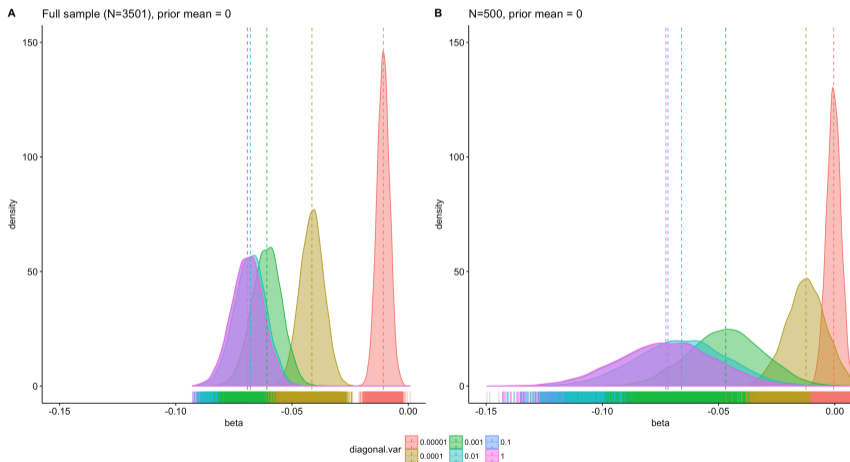
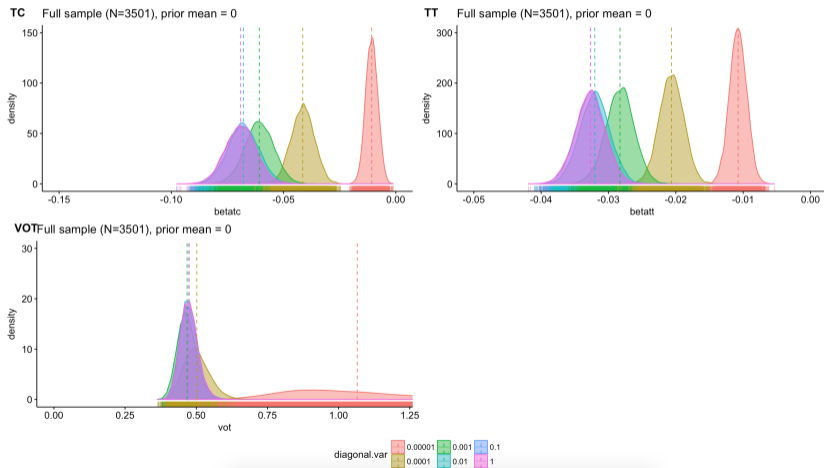


Figure: Posterior of β_{TC} , varying prior variance and sample size

Effect of the prior on WTP (probit estimates)



More about WTP posteriors and CIs

Energy Economics 44 (2014) 166–177



Contents lists available at ScienceDirect

Energy Economics

journal homepage: www.elsevier.com/locate/eneco



Accounting for uncertainty in willingness to pay for environmental benefits



Ricardo A. Daziano^{a,*}, Martin Achtnicht^b

^a School of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14853, United States

^b Centre for European Economic Research (ZEW), L7.1, D-68161 Mannheim, Germany

ARTICLE INFO

Article history:

Received 17 December 2012

Received in revised form 18 December 2013

Accepted 24 March 2014

Available online 15 April 2014

JEL classification:

C25

D12

Q51

Keywords:

Discrete choice models

Willingness to pay

Credible sets

ABSTRACT

Previous literature on the distribution of willingness to pay has focused on its heterogeneity distribution without addressing exact interval estimation. In this paper we derive and analyze Bayesian confidence sets for quantifying uncertainty in the determination of willingness to pay for carbon dioxide abatement. We use two empirical case studies: household decisions of energy-efficient heating versus insulation, and purchase decisions of ultra-low-emission vehicles. We first show that deriving credible sets using the posterior distribution of the willingness to pay is straightforward in the case of deterministic consumer heterogeneity. However, when using individual estimates, which is the case for the random parameters of the mixed logit model, it is complex to define the distribution of interest for the interval estimation problem. This latter problem is actually more involved than determining the moments of the heterogeneity distribution of the willingness to pay using frequentist econometrics. A solution that we propose is to derive and then summarize the distribution of point estimates of the individual willingness to pay under different loss functions.

© 2014 Elsevier B.V. All rights reserved.



From binary to multinomial probit

- $\mathbf{U}_i \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, very hard to estimate in frequentist setting (GHK)
- Ordinary regression with unknown variance has a normal-gamma conjugate posterior
- Normal prior-posterior for $\boldsymbol{\beta}$
- The gamma distribution works for σ^{-1}
- The Wishart distribution is the multidimensional version of the gamma and works for $\boldsymbol{\Sigma}^{-1}$
- Sketch of Gibbs sampler for multinomial probit
 - 1 Draw $\boldsymbol{\beta}^{(g+1)}$ from normal
 - 2 Draw $\boldsymbol{\Sigma}^{(g+1)}$ from inverse Wishart
 - 3 Draw $\mathbf{U}^{(g+1)}$ from truncated normal



Combining Gibbs sampling and GHK

TRANSPORTATION SCIENCE

Vol. 48, No. 4, November 2014, pp. 671–683
ISSN 0041-1655 (print) | ISSN 1526-5447 (online)

informs

<http://dx.doi.org/10.1287/trsc.2013.0464>
© 2014 INFORMS

Forecasting Adoption of Ultra-Low-Emission Vehicles Using Bayes Estimates of a Multinomial Probit Model and the GHK Simulator

Ricardo A. Daziano

School of Civil and Environmental Engineering, Cornell University, Ithaca, New York 14853, daziano@cornell.edu

Martin Achtnicht

Centre for European Economic Research, D-68161 Mannheim, Germany, achtnicht@zew.de

In this paper we use Bayes estimates of a multinomial probit model with fully flexible substitution patterns to forecast consumer response to ultra-low-emission vehicles. In this empirical application of the probit Gibbs sampler, we use stated-preference data on vehicle choice from a Germany-wide survey of potential light-duty-vehicle buyers using computer-assisted personal interviewing. We show that Bayesian estimation of a multinomial probit model with a full covariance matrix is feasible for this medium-scale problem and provides results that are very similar to maximum simulated likelihood estimates. Using the posterior distribution of the parameters of the vehicle choice model as well as the GHK simulator, we derive the choice probabilities of the different alternatives. We first show that the Bayes point estimates of the market shares reproduce the observed values. Then we define a base scenario of vehicle attributes that aims to represent an average of the current vehicle choice situation in Germany. Consumer response to qualitative changes in the base scenario is subsequently studied. In particular, we analyze the effect of increasing the network of service stations for charging electric vehicles as well as for refueling hydrogen. The result is the posterior distribution of the choice probabilities that represent adoption of the energy-efficient technologies.



Integrated choice and latent variable (ICLV) model

- Choice model with latent variables \mathbf{z} as independent variables
- Just as utility (a latent variable), we need structural and measurement equations
 - Structural parameters \mathbf{b} (+ nuisance parameters Ψ)
 - Measurement parameters λ (loading factors)
- Painful to estimate in a frequentist setting
- Relatively simple in a Bayesian setting with a probit kernel
 - 1 Posterior of interest $p(\beta, \Sigma, \mathbf{b}, \Psi, \lambda | \mathbf{y}, \mathbf{I}, \mathbf{X}, \mathbf{w})$
 - 2 Don't forget about data augmentation! **Augmented posterior:**

$$p(\mathbf{U}, \mathbf{z}, \beta, \Sigma, \mathbf{b}, \Psi, \lambda | \mathbf{y}, \mathbf{I}, \mathbf{X}, \mathbf{w})$$

ICLV Gibbs sampler

We know the trick already: conditional on \mathbf{z} , structural parameters \mathbf{b} are estimated using a Bayesian regression (normal sampling), we know how to deal with the multinomial probit kernel

Sketch of ICLV Gibbs sampler

- 1 Draw $\beta^{(g+1)}$ from normal (regression)
- 2 Draw $\Sigma^{(g+1)}$ from inverse Wishart
- 3 Draw $\mathbf{U}^{(g+1)}$ from truncated normal
- 4 Draw $\mathbf{z}^{(g+1)}$ from appropriate normal
- 5 Draw $\mathbf{b}^{(g+1)}$ from normal (regression)
- 6 Draw $\lambda^{(g+1)}$ from normal (regression)
- 7 Draw $\Psi^{(g+1)}$ from inverse Wishart



ICLV sampler: all the details

Transportation Research Part B 76 (2015) 1–26



Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb



Inference on mode preferences, vehicle purchases, and the energy paradox using a Bayesian structural choice model



Ricardo A. Daziano *

School of Civil and Environmental Engineering, Cornell University, 305 Hollister Hall, Ithaca, NY 14853, United States

ARTICLE INFO

Article history:

Received 31 January 2014
Received in revised form 24 February 2015
Accepted 26 February 2015
Available online 19 March 2015

Keywords:

Bayesian microeconometrics
Discrete choice models
Structural equation modeling
Energy paradox

ABSTRACT

Discrete choice modeling is experiencing a reemergence of research interest in the inclusion of latent variables as explanatory variables of consumer behavior. There are several reasons that motivate the integration of latent attributes, including better-informed modeling of random consumer heterogeneity and treatment of endogeneity. However, current work still is at an early stage and multiple simplifying assumptions are usually imposed. For instance, most previous applications assume all of the following: independence of taste shocks and of latent attributes, exclusion restrictions, linearity of the effect of the latent attributes on the utility function, continuous manifest variables, and an a priori bound for the number of latent constructs. We derive and apply a structural choice model with a multinomial probit kernel and discrete effect indicators to analyze continuous latent segments of travel behavior, including inference on the energy paradox. Our estimator allows for interaction and simultaneity among the latent attributes, residual correlation, nonlinear effects on the utility function, flexible substitution patterns, and temporal correlation within responses of the same individual. Statistical properties of the Bayes estimator that we propose are exact and are not affected by the number of latent attributes.

© 2015 Elsevier Ltd. All rights reserved.

Bayesian logit models

- We keep normal prior for β
- Conditional logit posterior:

$$p(\beta | \mathbf{y}, \mathbf{X}) \propto \exp\left(-(\beta - \check{\beta})' \check{V}_{\beta}^{-1} (\beta - \check{\beta})\right) \prod_{i=1}^N \left(\frac{\exp(\mathbf{x}'_{ij}\beta)}{\sum_j \exp(\mathbf{x}'_{ij}\beta)}\right)^{y_{ij}}$$

- In general, there is **no conjugacy** when having an EV kernel (Bayesians prefer probit due to conjugacy)
- We cannot use Gibbs sampling either, we need to explore the parameter space using a different tool
- For completely unknown posteriors, **Metropolis-Hastings** (MH) is used



Sketch of MH for conditional logit – at iteration (g):

- ① Generate a candidate draw β^{cand} from transition probability $q(\beta^{cand}|\beta^{curr})$ (if normal centered at β^{cand} : **random walk**)
- ② Evaluate prior at β^{cand} and β^{curr}
- ③ Evaluate likelihood at β^{cand} and β^{curr}
- ④ Calculate **acceptance ratio** α
- ⑤ Take a draw from the uniform density
- ⑥ Accept candidate ($\beta^{curr} = \beta^{cand}$) if uniform draw $< \alpha$



Bayes estimators of logit-type models

Resource and Energy Economics 35 (2013) 429–450



Conditional-logit Bayes estimators for consumer valuation of electric vehicle driving range



Ricardo A. Daziano*

School of Civil and Environmental Engineering, Cornell University, 305 Hollister Hall, Ithaca, NY 14853, United States

ARTICLE INFO

Article history:

Received 16 February 2012
Received in revised form 1 May 2013
Accepted 3 May 2013
Available online 14 May 2013

JEL classification:

C25
D12
Q42
Q50

Keywords:

Bayesian discrete choice
Battery electric vehicles
Range anxiety
Heterogeneity distributions

ABSTRACT

Range anxiety – consumers' concerns about limited driving range – is generally considered an important barrier to the adoption of electric vehicles. If consumers cannot overcome these fears it is unlikely that they will consider purchasing an electric car. Hence, a successful introduction of low emission vehicles in the market requires a full understanding of consumer valuation of driving range. By analyzing experimental data on vehicle purchase decisions in California, I derive and study the statistical behavior of Bayes estimates that summarize consumer concerns toward limited driving range. These estimates are superior to marginal utilities as parameters of interest in a discrete demand model of vehicle choice. One of the empirical results is the posterior distribution of the willingness to pay for electric vehicles with improved batteries offering better driving range. Credible intervals for this willingness to pay, as well as both parametric and nonparametric heterogeneity distributions, are also analyzed.

© 2013 Elsevier B.V. All rights reserved.

Another example (logit-type hazard models)

Transportation Research Part A 96 (2017) 154–167



Contents lists available at [ScienceDirect](#)

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra



Bayesian estimation of hazard models of airline passengers' cancellation behavior



Esther Chiew^a, Ricardo A. Daziano^{a,*}, Laurie A. Garrow^b

^a Cornell University, School of Civil and Environmental Engineering, 301 Hollister Hall, Ithaca, NY 14853, United States

^b Georgia Institute of Technology, School of Civil and Environmental Engineering, 790 Atlantic Drive, Atlanta, GA 30332-0355, United States

ARTICLE INFO

Article history:

Received 18 February 2016

Received in revised form 8 November 2016

Accepted 16 December 2016

Available online 6 January 2017

Keywords:

Air travel demand

Cancellation probabilities

Bayes estimation

Credible intervals

ABSTRACT

This study explores the use of Bayesian methods to estimate hazard models of airline passengers' cancellation behavior. We show how the discrete time proportional odds (DTPO) cancellation model can be rewritten as an equivalent fixed parameter discrete choice model that can be easily estimated using Bayesian methods and extended to random parameters that account for unobserved heterogeneity. The use of Bayesian methods allows us to address several limitations of existing airline cancellation models. First, because of the random parameter reformulation, it is possible to calculate individual-specific cancellation probabilities. Second, unlike existing DTPO models that forecast average cancellation probabilities only, our model can be used to forecast both means and a measure of variance (credible intervals) associated with an individual's cancellation probability. We apply the Bayesian estimation method to a dataset of tickets purchased over a two-year period by employees of a university in Atlanta, Georgia. During this time period, the major carrier in Atlanta terminated an agreement in which it allowed employees to purchase discounted fares that could be refunded or exchanged without a fee. The data allow us to investigate how passenger cancellation behavior changed when these fares were discontinued. Cancellations are reduced on average 3.3% when customers must pay to exchange their tickets. For a simulated hypothetical flight the coefficient of variation of cancellation is 43% when the state rate was offered, and 83% without state rates.

© 2016 Elsevier Ltd. All rights reserved.

Interval estimation problem

Definition: Credible region or Bayesian confidence region. A set $\mathcal{C} \subseteq \Theta$ such that

$$P(\boldsymbol{\theta} \in \mathcal{C}) = \int_{\mathcal{C}} p(\boldsymbol{\theta}|\mathbf{y})d\mu(\boldsymbol{\theta}) = 1 - \alpha,$$

where $(1 - \alpha)$ is a credibility level.

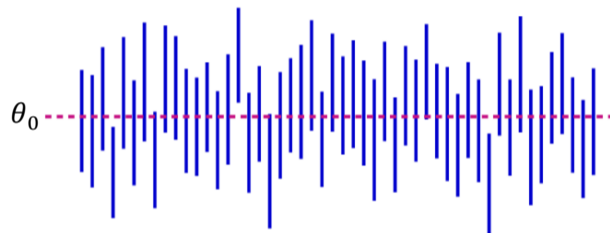
- Note that a credible region is a fixed area containing $\boldsymbol{\theta}$ with a specified coverage probability $(1 - \alpha)$, conditional on the observed data \mathbf{y}

Difference with frequentist confidence region

- The frequentist confidence region is a completely different concept
 - 1 Under a classical perspective θ is fixed: there is no sense in constructing a region based on its distribution
 - 2 A non-Bayesian confidence region is constructed using the **unobserved sampling distribution** of the estimator

A classical confidence region is **asymptotic**: the region depends on the distribution of unobserved realizations of the data that cannot be obtained for small samples; this distribution can be described using large sample theory

Frequentist confidence interval

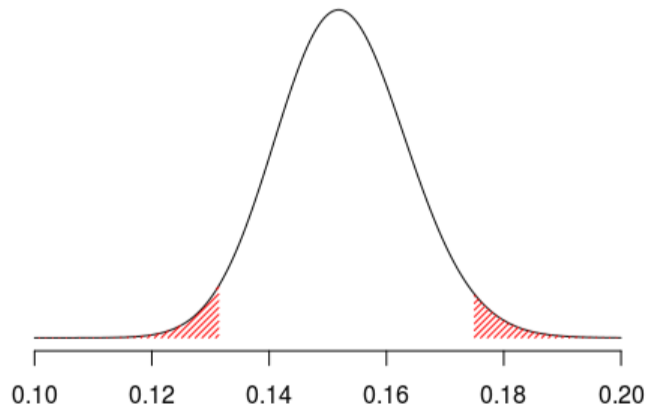


Constructing credible intervals

- **Quantile intervals:** take the $(\alpha/2)^{\text{th}}$ and $(1 - \alpha/2)^{\text{th}}$ quantile values of the sorted MCMC draws to find CI thresholds
- Accurate when the posterior distribution is symmetric
- **Highest Posterior Density (HPD) intervals:** shortest possible interval with a fixed probability $1 - \alpha$



Credible interval (posterior mass)



Logit credible intervals

Table: Interval estimates for the binary logit model

	Bayesian estimates 95% HPD int.	Classical estimates 95% conf. int.
ASC	[-0.105,0.062]	[-0.103,0.065]
TC	[-0.160,-0.107]	[-0.159,-0.105]
TT	[-0.069,-0.052]	[-0.068,-0.052]
HW	[-0.041,-0.034]	[-0.042,-0.034]
CH	[-1.237,-1.070]	[-1.237,-1.069]

Mixed Logit: MH within Gibbs

- Consider $\beta_i \sim \mathcal{N}(\mu, \Sigma)$
- Individual-level parameter σ_i is a **latent variable**
- Bayes: **data augmentation**
- If we know β_i then $p(y_{it} | \mathbf{X}_{it}, \beta_i)$ is simply a conditional logit
- Mixing density $f(\beta_i | \mu, \Sigma)$ acts as prior
- Only additional priors for μ and Σ required (**hyperparameters**)

Mixed Logit: sketch of the sampler

Posterior:

$$p(\boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{y}, \mathbf{X}) \propto p(\boldsymbol{\mu})p(\boldsymbol{\Sigma}) \prod_{i=1}^N p(\mathbf{y}_i | \boldsymbol{\beta}_i, \mathbf{X}_i) f(\boldsymbol{\beta}_i | \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- 1 Normal prior on $\boldsymbol{\mu}$; conjugate prior (normal update)
- 2 Inverse Wishart prior on $\boldsymbol{\Sigma}$; conjugate prior (IW update)
- 3 Normal prior for $\boldsymbol{\beta}$; MH update



Fresh from the oven: our working paper

Bayesian Estimation of Mixed Multinomial Logit Models: Advances and Simulation-Based Evaluations

Prateek Bansal, Rico Krueger, Michel Bierlaire, Ricardo A. Daziano, Taha H. Rashidi

(Submitted on 7 Apr 2019 (v1), last revised 12 Apr 2019 (this version, v2))

Variational Bayes (VB) methods have emerged as a fast and computationally-efficient alternative to Markov chain Monte Carlo (MCMC) methods for Bayesian estimation of mixed multinomial logit (MMNL) models. It has been established that VB is substantially faster than MCMC at practically no compromises in predictive accuracy. In this paper, we address two critical gaps concerning the usage and understanding of VB for MMNL. First, extant VB methods are limited to utility specifications involving only individual-specific taste parameters. Second, the finite-sample properties of VB estimators and the relative performance of VB, MCMC and maximum simulated likelihood estimation (MSLE) are not known. To address the former, this study extends several VB methods for MMNL to admit utility specifications including both fixed and random utility parameters. To address the latter, we conduct an extensive simulation-based evaluation to benchmark the extended VB methods against MCMC and MSLE in terms of estimation times, parameter recovery and predictive accuracy. The results suggest that all VB variants perform as well as MCMC and MSLE at prediction and recovery of all model parameters with the exception of the covariance matrix of the multivariate normal mixing distribution. In particular, VB with nonconjugate variational message passing and the delta-method (VB-NCVMP-Delta) is relatively accurate and up to 15 times faster than MCMC and MSLE. On the whole, VB-NCVMP-Delta is most suitable for applications in which fast predictions are paramount, while MCMC should be preferred in applications in which accurate inferences are most important.

What we are doing in this working paper

- Mixed logit with inter and intra-consumer heterogeneity
- Goal: fast(er) predictions (thinking of recommender systems)
- MCMC: Huang's half-t prior for Σ (Akinc and Vandebroek, 2018)
- **Variational Bayes**: computationally-efficient alternative to MCMC
 - 1 VB is substantially faster than MCMC at practically no compromises in predictive accuracy
 - 2 Approximate Bayesian inference based on optimization rather than sampling
 - 3 Locally-optimal, (exact) analytical solution to an approximation of the posterior
- VB up to 15 times faster than MCMC and MSLE

Summarizing

- In Bayesian econometrics: the analyst wishes to learn about uncertain parameters
- Applying the rules of probability, the analyst can update prior knowledge in regards to new evidence (posterior)
- The prior reflects both knowledge and beliefs (subjective probabilities)
- If possible, priors are chosen to ensure conjugacy
- Bayes estimates do depend on the likelihood (and it's the same frequentist likelihood)
- Bayes point estimates do exist: usually posterior mean
- Posterior standard deviations measure the precision of the Bayes point estimates



Summarizing cont'd

- Bayesian econometrics works particularly well for latent variables
- Complex models that face convergence issues in MLE can be estimated using Bayesian tools (no optimization)
- Conditional estimates at the individual level are a direct result of estimation
- So, what is preventing us from adopting Bayes?
 - 1 Learning curve
 - 2 Perception of lack of flexible, easy-to-use software



Other uses of Bayes tools



Transportation Research Part C: Emerging Technologies

Available online 3 October 2018

In Press, Corrected Proof [?](#)



A framework to integrate mode choice in the design of mobility-on-demand systems

Yang Liu ¹ ✉, Prateek Bansal ¹ ✉, Ricardo Daziano ✉, Samitha Samaranyake ✉

Un gros merci!

