Bayesian Estimation for Discrete Choice Models

Methodological Foundations

Michel Bierlaire

December 20, 2025

Report TRANSP-OR xxxxxx
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
École Polytechnique Fédérale de Lausanne

Contents

1 Introduction

This document presents the methodological foundations of Bayesian estimation for discrete choice models. It is designed as a conceptual and theoretical companion to the use of `Biogeme` for Bayesian estimation, complementing the software documentation and user guide.

The intended audience consists of readers who are already familiar with random utility theory, likelihood-based inference, and standard discrete choice models, in particular those accustomed to maximum likelihood estimation. The goal of this document is to provide the theoretical background necessary to understand how Bayesian estimation operates in the context of discrete choice modeling and how the results produced by `Biogeme` should be interpreted.

The objective is not to offer a comprehensive or fully rigorous treatment of Bayesian statistics. Rather, the focus is on introducing the essential concepts and computational principles that underlie Bayesian estimation of discrete choice models, with particular emphasis on models that are too complex to be handled analytically. Throughout the document, methodological choices are motivated by their practical implications for model specification, estimation, and interpretation in `Biogeme`.

In particular, the document aims to clarify:

- the conceptual differences between Bayesian inference and maximum likelihood estimation, and the implications of these differences for parameter uncertainty and inference;
- why simulation-based methods are unavoidable for realistically specified discrete choice models, especially in the presence of random coefficients and hierarchical structures;
- how Markov chain Monte Carlo (MCMC) methods are used to approximate posterior distributions when closed-form solutions are unavailable;
- why modern gradient-based algorithms, such as Hamiltonian Monte Carlo (HMC), are particularly well suited to Bayesian estimation of discrete choice models.

Implementation details, software configuration, and step-by-step user instructions are intentionally excluded from this document. These aspects are documented separately in the `Biogeme` user guide, while the present text focuses on the underlying statistical and computational principles needed to use Bayesian estimation in `Biogeme` in an informed and critical manner.

2 Bayesian inference for discrete choice models

Bayesian inference provides a probabilistic framework for learning about unknown model parameters from observed choice data. Uncertainty about these parameters is represented explicitly through probability distributions, rather than summarized by a single point estimate.

Let $\mathcal{D} = \{\mathtt{y}_n, \mathbf{x}_n\}_{n=1}^{N}$ denote the observed data, where $\mathtt{y}_n$ is the observed choice made in situation n and $\mathbf{x}_n$ is the corresponding vector of explanatory variables. Let θ denote the vector of parameters of the discrete choice model, such as alternative-specific constants and taste coefficients entering the utility functions.

In the Bayesian framework, the parameters θ are treated as unknown random quantities and assigned a prior distribution $p(\theta)$. This prior encodes information or beliefs about plausible parameter values before observing the data. The likelihood function,

$$L(\mathcal{D} \mid \theta) = \prod_{n=1}^{N} p(\mathtt{y}_n \mid \mathbf{x}_n, \theta),$$

summarizes the information provided by the observed choices through the assumed behavioral model.

Bayes' theorem combines prior information and empirical evidence to produce the posterior distribution:

$$p(\theta \mid \mathcal{D}) = \frac{L(\mathcal{D} \mid \theta)\, p(\theta)}{\int L(\mathcal{D} \mid \theta')p(\theta')\, d\theta'}, \tag{1}$$

where the denominator is a normalizing constant ensuring that the posterior integrates to one. This quantity, sometimes referred to as the marginal likelihood or model evidence, is typically intractable for discrete choice models of practical interest.

The posterior distribution is the central object of Bayesian inference. It captures all information about the parameters that is available after observing the data and forms the basis for estimation, prediction, and policy analysis. In contrast to frequentist estimation, which yields a single point estimate and relies on asymptotic arguments for uncertainty quantification, Bayesian inference produces a full probability distribution over parameters. This enables direct probability statements, coherent predictive inference, and systematic propagation of parameter uncertainty into all derived quantities.

3 Simulation and Monte Carlo approximation

Analytical evaluation of integrals involving probability distributions quickly becomes infeasible, even for relatively simple models. This is particularly true in Bayesian discrete choice models, where posterior distributions are high-dimensional, non-Gaussian, and defined only up to a normalizing constant. Monte Carlo simulation provides a general and powerful alternative by replacing analytical integration with numerical averaging.

Let X be a random variable with density f_X, and let $X_1, \ldots, X_R$ be independent and identically distributed draws from this density. For any measurable function $g(\cdot)$ such that $\mathbb{E}[|g(X)|] < \infty$, the law of large numbers implies

$$\frac{1}{R}\sum_{r=1}^{R} g(X_r) \xrightarrow[R\to\infty]{\text{a.s.}} \mathbb{E}[g(X)] = \int g(x)\, f_X(x)\, dx.$$

The indicator-function example,

$$\frac{1}{R}\sum_{r=1}^{R} \mathbf{1}\{X_r \in [a, b]\} \xrightarrow[R\to\infty]{\text{a.s.}} \int_a^b f_X(x)\, dx,$$

is a special case illustrating how probabilities can be approximated by sample frequencies.

Monte Carlo methods therefore approximate expectations with respect to complex distributions by simple averages over simulated draws. The accuracy of these approximations improves as the number of simulations increases, independently of the dimensionality of the problem.

In Bayesian inference, Monte Carlo simulation plays a central role. Posterior expectations of parameters, predictive probabilities, and other quantities of interest are approximated by averaging the corresponding functions over draws from the posterior distribution. However, for discrete choice models, the posterior distribution $p(\theta \mid \mathcal{D})$ cannot be sampled from directly, because it is known only up to a proportionality constant and typically has a complex shape.

This difficulty motivates the use of more advanced simulation techniques, such as Markov chain Monte Carlo methods, which generate dependent draws whose stationary distribution coincides with the posterior. These methods form the computational backbone of Bayesian estimation for discrete choice models and are discussed in the following sections.

4 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods are a class of simulation techniques designed to generate dependent draws whose empirical distribution converges to a target posterior distribution. Rather than sampling independently, MCMC constructs a stochastic process that

explores the parameter space in such a way that regions of high posterior probability are visited more frequently.

An MCMC algorithm produces a sequence

$$\theta^{(1)}, \theta^{(2)}, \ldots$$

forming a Markov chain whose stationary distribution is the posterior $p(\theta \mid \mathcal{D})$. Under standard regularity conditions, ergodic averages of this sequence converge to posterior expectations. Consequently, once the chain has converged, Monte Carlo averages of the draws can be used to approximate posterior means, variances, and other quantities of interest.

The initial iterations of an MCMC chain are typically influenced by the arbitrary starting values and do not yet reflect the stationary distribution. These early draws are therefore discarded during a *burn-in* or *warm-up* phase. In addition to reducing the impact of initialization, the warm-up phase is often used to tune internal algorithmic parameters, such as proposal scales or step sizes. Only draws generated after this phase are retained for posterior inference.

It is standard practice to run multiple MCMC chains in parallel, each initialized from dispersed starting points in the parameter space. If all chains converge to the same distribution and mix well, their trajectories should be statistically indistinguishable after warm-up. Consistency across chains therefore provides strong evidence of convergence and adequate exploration of the posterior distribution.

Unlike standard Monte Carlo sampling, draws generated by MCMC are generally correlated. As a consequence, a sequence of N MCMC draws typically contains less information than N independent samples from the same distribution. This reduction in information is quantified by the *effective sample size* (ESS), which measures the number of independent draws that would yield an estimator with the same variance.

Consider a scalar parameter θ and a sequence of N post–warm-up draws $\{\theta^{(t)}\}_{t=1}^{N}$. Let ρ_k denote the lag-k autocorrelation of the chain that measures the linear dependence between draws separated by k iterations of the Markov chain:

$$\rho_k = \frac{\sum_{t=1}^{N-k} \left(\theta^{(t)} - \bar{\theta}\right)\left(\theta^{(t+k)} - \bar{\theta}\right)}{\sum_{t=1}^{N} \left(\theta^{(t)} - \bar{\theta}\right)^2}.$$

Under standard assumptions, the effective sample size is defined as

$$\mathrm{ESS} = \frac{N}{1 + 2\sum_{k=1}^{\infty} \rho_k}. \tag{2}$$

The denominator captures the cumulative impact of autocorrelation: positive autocorrelation inflates the variance of Monte Carlo estimators and reduces the amount of independent information contained in the sample.

In practice, the infinite sum in (2) is truncated using data-driven rules, and effective sample sizes are estimated separately for each parameter or function of parameters. Modern Bayesian software, including PyMC and ArviZ, reports multiple ESS measures, such as bulk and tail ESS, which assess sampling efficiency in the central region and in the tails of the posterior distribution.

High autocorrelation leads to small effective sample sizes and thus larger Monte Carlo error. Reliable Bayesian inference therefore requires not only convergence of the chains but also sufficiently large effective sample sizes for the quantities of interest.

We refer the reader to Geyer (1992), Gelman et al. (2014) and Vehtari et al. (2021) for comprehensive discussions of MCMC theory, diagnostics, and best practices.

5 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC), originally introduced in the physics and lattice field theory literature and later formalized for statistical inference by Duane et al. (1987) and Neal (2011), addresses the inefficiency of random-walk MCMC methods by exploiting gradient information of the log posterior to generate long, coherent proposals.

Let $\mathsf{f}(\boldsymbol{\theta}) \propto \mathsf{p}(\boldsymbol{\theta} \mid \mathcal{D})$. HMC introduces an auxiliary momentum variable $\mathbf{p}$ and defines the Hamiltonian

$$\mathsf{H}(\boldsymbol{\theta}, \mathbf{p}) = -\log \mathsf{f}(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{p}^\top \mathbf{M}^{-1}\mathbf{p},$$

where $\mathbf{M}$ is a positive definite mass matrix.

Sampling proceeds by simulating Hamiltonian dynamics in the augmented space $(\boldsymbol{\theta}, \mathbf{p})$, followed by a Metropolis correction step to ensure exactness.

Leapfrog integration. Hamiltonian Monte Carlo relies on simulating Hamiltonian dynamics in an augmented parameter–momentum space. Since the associated differential equations cannot be solved analytically in realistic models, they are approximated numerically using the *leapfrog* (or Störmer–Verlet) integrator. The leapfrog scheme alternates half-steps for the momentum and full-steps for the parameters, using gradients of the log posterior to guide the trajectory.

A key property of the leapfrog integrator is that it is both *time-reversible* and *volume-preserving*. These properties are essential: they ensure that the numerical approximation can be embedded within a Metropolis–Hastings accept–reject step that exactly preserves the target posterior distribution, despite the presence of numerical integration error. As the step size decreases, the Hamiltonian error decreases and the acceptance probability approaches one, allowing HMC to propose long, coherent moves through the parameter space with high efficiency.

The No-U-Turn Sampler. In standard Hamiltonian Monte Carlo, the efficiency of the algorithm depends critically on the choice of the trajectory length (equivalently, the number of leapfrog steps). If the trajectory is too short, the sampler behaves like a random walk; if it is too long, computation is wasted as the trajectory starts retracing its steps.

The No-U-Turn Sampler (NUTS) resolves this issue by dynamically and automatically determining when to stop the Hamiltonian trajectory. It builds trajectories forward and backward in time and terminates the expansion when a *U-turn* is detected, that is, when the momentum begins to point back toward previously visited states. This adaptive strategy eliminates the need for manual tuning of the trajectory length while retaining the theoretical guarantees of HMC.

Combined with automatic step-size and mass-matrix adaptation during warm-up, NUTS provides a robust, efficient, and largely tuning-free algorithm. For this reason, it has become the default sampler in modern Bayesian software such as Stan and PyMC, and is the algorithm used by Biogeme for Bayesian estimation.

Leapfrog integration. Hamiltonian Monte Carlo relies on simulating Hamiltonian dynamics in an augmented parameter–momentum space. Since the associated differential equations cannot be solved analytically in realistic models, they are approximated numerically using the *leapfrog* (or Störmer–Verlet) integrator. The leapfrog scheme alternates half-steps for the momentum and full-steps for the parameters, using gradients of the log posterior to guide the trajectory.

A key property of the leapfrog integrator is that it is both *time-reversible* and *volume-preserving*. These properties are essential: they ensure that the numerical approximation can be embedded within a Metropolis–Hastings accept–reject step that exactly preserves

the target posterior distribution, despite the presence of numerical integration error. As the step size decreases, the Hamiltonian error decreases and the acceptance probability approaches one, allowing HMC to propose long, coherent moves through the parameter space with high efficiency.

The No-U-Turn Sampler. In standard Hamiltonian Monte Carlo, the efficiency of the algorithm depends critically on the choice of the trajectory length (equivalently, the number of leapfrog steps). If the trajectory is too short, the sampler behaves like a random walk; if it is too long, computation is wasted as the trajectory starts retracing its steps.

The No-U-Turn Sampler (NUTS) resolves this issue by dynamically and automatically determining when to stop the Hamiltonian trajectory. It builds trajectories forward and backward in time and terminates the expansion when a *U-turn* is detected, that is, when the momentum begins to point back toward previously visited states. This adaptive strategy eliminates the need for manual tuning of the trajectory length while retaining the theoretical guarantees of HMC.

Combined with automatic step-size and mass-matrix adaptation during warm-up, NUTS provides a robust, efficient, and largely tuning-free algorithm. For this reason, it has become the default sampler in modern Bayesian software such as Stan and PyMC. It is the algorithm used by Biogeme for Bayesian estimation.

In summary, HMC and NUTS:

- scale well to high-dimensional parameter spaces,
- handle strong posterior correlations efficiently,
- avoid random-walk behavior,
- are well suited to hierarchical and latent-variable models.

These properties make them particularly effective for Bayesian discrete choice models, including mixtures and panel data.

6 Bayesian treatment of random coefficients

In a Bayesian framework, random coefficients are treated as *latent variables* and are inferred jointly with the structural (population-level) parameters of the model. This contrasts with classical logit mixtures estimation, where random coefficients are integrated out analytically or numerically to obtain a marginal likelihood.

6.1 Model structure

Let $\mathcal{D} = \{\mathrm{y}_n, \mathbf{x}_n\}_{n=1}^{N}$ denote the observed data, where y_n is the observed choice for decision-maker (or observation) n, and $\mathbf{x}_n$ collects the associated explanatory variables. Let $\boldsymbol{\eta}_n$ denote the vector of individual-specific random coefficients associated with observation n, and let $\boldsymbol{\beta}$ denote the vector of population-level parameters governing the distribution of these random coefficients (e.g. means, standard deviations, or covariance parameters).

The hierarchical structure of the model can be written as

$$\boldsymbol{\eta}_n \mid \boldsymbol{\beta} \sim \mathrm{p}(\boldsymbol{\eta}_n \mid \boldsymbol{\beta}), \qquad \mathrm{y}_n \mid \boldsymbol{\eta}_n \sim \mathrm{p}(\mathrm{y}_n \mid \boldsymbol{\eta}_n), \qquad \boldsymbol{\beta} \sim \mathrm{p}(\boldsymbol{\beta}),$$

where $\mathrm{p}(\boldsymbol{\beta})$ denotes the prior distribution over the structural parameters.

6.2 Joint posterior distribution

Rather than working with the marginal likelihood

$$p(y_n \mid \boldsymbol{\beta}) = \int p(y_n \mid \boldsymbol{\eta}_n) p(\boldsymbol{\eta}_n \mid \boldsymbol{\beta}) \, d\boldsymbol{\eta}_n,$$

Bayesian inference targets the *joint posterior* distribution of all unknown quantities:

$$p(\boldsymbol{\beta}, \boldsymbol{\eta}_{1:N} \mid \mathcal{D}) \;\propto\; p(\boldsymbol{\beta}) \prod_{n=1}^{N} p(\boldsymbol{\eta}_n \mid \boldsymbol{\beta}) p(y_n \mid \boldsymbol{\eta}_n).$$

This joint posterior is explored using Markov chain Monte Carlo (MCMC) methods. Each MCMC draw provides a joint realization of the population parameters $\boldsymbol{\beta}$ and the latent coefficients $\boldsymbol{\eta}_{1:N}$.

6.3 Implicit integration and relation to logit mixtures

Posterior expectations of functions of $\boldsymbol{\beta}$ or predictive quantities are obtained by Monte Carlo averaging over posterior draws. In particular, the integration over the random coefficients that appears explicitly in classical mixtures of logit models is performed *implicitly* by averaging over the sampled values of $\boldsymbol{\eta}_{1:N}$:

$$p(y_n \mid \mathcal{D}) \approx \frac{1}{S} \sum_{s=1}^{S} p\left(y_n \mid \boldsymbol{\eta}_n^{(s)}\right),$$

where $\{\boldsymbol{\eta}_n^{(s)}\}_{s=1}^{S}$ are draws from the posterior. This avoids explicit numerical integration and eliminates simulation noise associated with classical simulated maximum likelihood estimators.

6.4 Panel data

When panel data are available, each decision-maker i is observed over multiple choice occasions $t = 1, \ldots, T_i$. In that case, the random coefficients are indexed at the individual level and shared across observations:

$$\boldsymbol{\eta}_i \mid \boldsymbol{\beta} \sim p(\boldsymbol{\eta}_i \mid \boldsymbol{\beta}), \qquad y_{it} \mid \boldsymbol{\eta}_i \sim p(y_{it} \mid \boldsymbol{\eta}_i).$$

The joint posterior becomes

$$p(\boldsymbol{\beta}, \boldsymbol{\eta}_{1:I} \mid \mathcal{D}) \;\propto\; p(\boldsymbol{\beta}) \prod_{i=1}^{I} p(\boldsymbol{\eta}_i \mid \boldsymbol{\beta}) \prod_{t=1}^{T_i} p(y_{it} \mid \boldsymbol{\eta}_i).$$

This formulation naturally captures intertemporal correlation in choices through the shared latent coefficients $\boldsymbol{\eta}_i$, without requiring any modification of the estimation machinery. The Bayesian hierarchical model thus provides a unified and coherent framework for cross-sectional and panel logit mixtures models.

7 Identification and posterior geometry

Identification issues arise when the likelihood function is weakly informative with respect to some parameters or linear combinations of parameters. In such situations, substantial changes in these directions of the parameter space produce only negligible changes in the likelihood, so that the data provide little information to pin down unique parameter values. In frequentist

estimation, this typically manifests itself through a nearly singular or ill-conditioned Hessian matrix. In Bayesian estimation, the same phenomenon appears in the geometry of the posterior distribution.

When identification is weak, the posterior tends to be highly anisotropic. Some directions in parameter space remain very wide, reflecting large uncertainty along combinations of parameters that are poorly identified by the data. These wide directions are often accompanied by strong posterior correlations, as multiple parameters can trade off against one another without materially affecting model fit. From a computational perspective, such posterior geometries are challenging: Markov chains tend to explore these elongated regions slowly, leading to high autocorrelation and a low effective sample size, even when formal convergence diagnostics appear acceptable.

A useful way to characterize posterior geometry is through the eigenvalue decomposition of the posterior covariance matrix. The eigenvalues measure the variance of the posterior distribution along orthogonal directions in parameter space. Large eigenvalues correspond to directions in which the posterior is very diffuse, indicating weak identification, while small eigenvalues correspond to tightly constrained, well-identified directions. The ratio of the largest to the smallest eigenvalue, known as the condition number, provides a compact summary of posterior anisotropy. A large condition number signals near-linear dependencies among parameters and should be interpreted as a warning sign of weak or partial identification.

Finally, comparing posterior dispersion to prior dispersion provides additional insight into the source of identification. When the posterior variance of a parameter is similar to its prior variance, the data have added little information beyond the prior, suggesting that identification is driven primarily by prior assumptions. Conversely, substantial shrinkage from prior to posterior indicates that the likelihood is informative and that the parameter is identified by the data. Such comparisons are particularly valuable in Bayesian estimation, as they make explicit whether inference is supported by empirical evidence or mainly by prior structure.

8 Posterior inference and prediction

Once samples from the joint posterior distribution are available, statistical inference and prediction reduce to Monte Carlo integration. Let $\{\theta^{(s)}\}_{s=1}^{S}$ denote posterior draws of the model parameters, where θ collects all unknown quantities of interest, including structural parameters and, when relevant, latent random coefficients.

Posterior summaries such as means, medians, posterior modes, and credible intervals are obtained by evaluating the corresponding functionals over the posterior draws. For a scalar parameter θ, for example, the posterior mean and a $(1-\alpha)$ highest density interval are approximated as

$$\mathbb{E}[\theta \mid \mathcal{D}] \approx \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}, \qquad \mathrm{HDI}_{1-\alpha}(\theta) \approx \text{quantiles of } \{\theta^{(s)}\}_{s=1}^{S},$$

with analogous expressions for other summary statistics. Convergence diagnostics and effective sample sizes computed from the MCMC output provide quantitative measures of the reliability of these estimates.

Prediction in a Bayesian framework is based on the posterior predictive distribution. For a new observation with covariates x^*, the predictive probability of outcome y^* is given by

$$p(y^* \mid x^*, \mathcal{D}) = \int p(y^* \mid x^*, \theta) p(\theta \mid \mathcal{D}) \, d\theta,$$

which is approximated by Monte Carlo averaging over posterior draws:

$$p(y^* \mid x^*, \mathcal{D}) \approx \frac{1}{S}\sum_{s=1}^{S} p(y^* \mid x^*, \theta^{(s)}).$$

The same principle applies to derived quantities such as elasticities, value-of-time measures, or welfare changes induced by counterfactual policies. Each posterior draw defines a complete model realization, from which these quantities can be computed deterministically. The posterior distribution of the derived quantity is then obtained by evaluating it across all draws, yielding both point estimates and credible intervals that fully account for parameter uncertainty.

Posterior predictive simulation extends this approach further by generating synthetic outcomes from the predictive distribution. By sampling parameters from the posterior and outcomes from the corresponding likelihood, one obtains simulated datasets that reflect both stochastic choice behavior and uncertainty about the underlying model parameters. This provides a natural framework for model checking, forecasting, and scenario analysis.

Overall, Bayesian posterior inference and prediction offer a coherent and unified approach in which estimation, prediction, and policy evaluation are all based on the same probabilistic foundation. Parameter uncertainty is propagated automatically into predictions and derived outputs, enabling uncertainty-aware decision support without requiring ad hoc approximations or asymptotic arguments.

9 Concluding remarks

Bayesian estimation provides a coherent and flexible framework for discrete choice modeling. Its combination with modern MCMC algorithms enables the estimation of models that are difficult or impractical to handle using classical maximum likelihood.

The methodological concepts presented in this document underpin the practical implementation described in the companion software-oriented documentation.

References

Duane, S., Kennedy, A. D., Pendleton, B. J. and Roweth, D. (1987). Hybrid monte carlo, *Physics Letters B* **195**(2): 216–222. DOI: 10.1016/0370-2693(87)91197-X.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2014). *Bayesian Data Analysis*, Texts in Statistical Science Series, third edition edn, Chapman & Hall/CRC.

Geyer, C. J. (1992). Practical markov chain monte carlo, *Statistical Science* **7**(4): 473–483. DOI: 10.1214/ss/1177011137.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. L. Jones and X.-L. Meng (eds), *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC, pp. 113–162. Also available as arXiv:1206.1901.
URL: *https://arxiv.org/abs/1206.1901*

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\widehat{\mathsf{R}}$ for assessing convergence of MCMC, *Bayesian Analysis* **16**(2): 667–718. DOI: 10.1214/20-BA1221.