Hamiltonian Sequential Monte Carlo with Application to Consumer Choice Behavior*

Martin Burda† Remi Daviet‡

May 24, 2022

Abstract

Practical use of nonparametric Bayesian methods requires the availability of efficient algorithms for posterior inference. The inherently serial nature of traditional Markov Chain Monte Carlo (MCMC) methods imposes limitations on their efficiency and scalability. In recent years there has been a surge of research activity devoted to developing alternative implementation methods that target parallel computing environments. Sequential Monte Carlo (SMC), also known as a particle filter, has been gaining popularity due to its desirable properties. SMC uses a genetic mutation-selection sampling approach with a set of particles representing the posterior distribution of a stochastic process. We propose to enhance the performance of SMC by utilizing Hamiltonian transition dynamics in the particle transition phase, in place of random walk used in the previous literature. We call the resulting procedure Hamiltonian Sequential Monte Carlo (HSMC). Hamiltonian transition dynamics has been shown to yield superior mixing and convergence properties relative to random walk transition dynamics in the context of MCMC procedures. The rationale behind HSMC is to translate such gains to the SMC environment. HSMC will facilitate practical estimation of models with complicated latent structures, such as nonparametric individual unobserved heterogeneity, that are otherwise difficult to implement. We demonstrate the behavior of HSMC in a challenging simulation study and contrast its favourable performance with SMC and other alternative approaches. We then apply HSMC to a panel discrete choice model with nonparametric consumer heterogeneity, allowing for multiple modes, asymmetries, and data-driven clustering, providing insights for consumer segmentation, individual level marketing, and price micromanagement.

JEL: C11, C14, C15, C23, C25
Keywords: Particle filtering, Bayesian nonparametrics, mixed panel logit, discrete choice

*We would like to thank the Editor, two anonymous referees, and the participants of the 11th International Conference on Computational and Financial Econometrics (CFE), University of London, UK, 2017, and the European Seminar on Bayesian Econometrics (ESOBE), New Orleans, USA, 2018, for insightful comments and suggesions. Computations were performed on the Niagara supercomputer at the SciNet HPC Consortium. SciNet is funded by: the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

†Department of Economics, University of Toronto, 150 St. George St., Toronto, ON M5S 3G7, Canada; Email: martin.burda@utoronto.ca

‡University of Wisconsin, Madison, 975 University Avenue, Madison, WI 53706, USA

1 Introduction

In Bayesian statistics, parameters are treated as random variables and all forms of uncertainty are expressed in terms of probability. A nonparametric Bayesian model is a model whose parameter space has infinite dimensionality. For any given finite data set, only a finite subset of the available parameters is invoked whereby its dimensionality is allowed to grow with the sample size (Orbanz and Teh, 2010). Nonparametric (and semiparametric) models allow us to avoid the often arbitrary and possibly unverifiable assumptions inherent in parametric models. A recent detailed exposition of Bayesian nonparametric methods is provided in Ghosal and van der Vaart (2017).

A critical issue for the practical use of nonparametric Bayesian models is the availability of efficient algorithms for posterior inference. The last several decades have witnessed an explosive growth of numerical implementation methods in Bayesian analysis. The cornerstone of such methods has been Markov Chain Monte Carlo (MCMC) and its variants. Nonetheless, the inherently *serial* nature of MCMC, whereby a new draw of the desired parameter chain can only be taken conditional on completing the preceding draw, imposes limitations on the implementational efficiency and scalability of such methods. Yet, the speed of microprocessor cores measured by their GHz frequency has been virtually stable since the mid-2000s, following decades of rapid growth (Rupp, 2018). During the last ten years or so, improvements in computing performance have not originated from processor speed but rather from *parallelization*[4].

In recent years there has been a surge of research activity devoted to developing alternative implementation methods that target (massively) parallel computing environments. In this paper we focus on one particular stream of research in this area: Sequential Monte Carlo (SMC), also known as a particle filter (Doucet et al., 2001). SMC uses a genetic mutation-selection sampling approach with a set of particles representing the posterior distribution of a stochastic process. SMC is highly parallelizable as the core computational load involving the model likelihood is performed by individual particles independently of one another. Due to their computational complexity, Bayesian nonparametric methods stand to benefit substantially from such approaches.

SMC algorithms were initially developed to solve filtering problems that arise in nonlinear state space models (Doucet et al., 2001). In economics, the SMC approach has become a popular method of inference

[4]Recent trends include shared memory multi-core CPUs, GPUs, and distributed memory high-performance clusters.

for dynamic systems that benefit from real-time updating of the posterior approximation via recursive importance sampling updates (Kim et al., 1998; Fernández-Villaverde and Rubio-Ramírez, 2007; Creal, 2012; Lopes and Carvalho, 2013; Herbst and Schorfheide, 2014; Blevins, 2016). Chopin (2002) adapted SMC to conduct posterior inference for a static Euclidean parameter vector. This approach was further extended by Fearnhead (2004), Ulker et al. (2010), Carvalho et al. (2010), Bouchard-Côté et al. (2017), and Griffin (2017) to static nonparametric mixture models, which is also our modeling context. The extent to which SMC is parallelizable in a related parametric environment and the corresponding computational gains are elaborated in Durham and Geweke (2014).

Starting with a particle distribution representing the prior distribution, SMC is typically implemented in three phases to update the prior through the likelihood: (i) particle reweighting (correction phase), (ii) particle resampling (selection phase), and (iii) particle transition (mutation phase). We detail each phase further below. In this paper, we propose to enhance the performance of SMC by utilizing Hamiltonian transition dynamics in the particle mutation phase, in place of random walk transitions used in SMC in the previous literature in the computationally expensive mutation phase. We call the resulting procedure Hamiltonian Sequential Monte Carlo (HSMC). Hamiltonian transition dynamics have been shown to yield superior mixing and convergence properties relative to random walk transition dynamics in the context of serial MCMC procedures (Neal, 2011). In particular, Hamiltonian dynamics use information about the first derivative of the likelihood function and construct a proposal draw using a sequence of steps, unlike random walk (RW) one-step proposals that do not use derivative information. The rationale behind HSMC is to extend such gains to the SMC environment. We compare the behavior of HSMC with SMC and other alternative methods in a challenging simulation study, demonstrating favourable performance of HSMC. We further apply HSMC to a panel discrete choice model with a nonparametric distribution of unobserved consumer heterogeneity. The results can provide practitioners with valuable input into consumer segmentation analysis, individual level marketing, and price micromanagement.

The remainder of this paper is organized as follows. In section 2 we provide the background for MCMC methods and in section 3 a review of SMC. Hamiltonian transition dynamics are detailed in section 4. We discuss Bayesian nonparametrics in section 5. Within the context of a Bayesian nonparametric mixture model, we introduce HSMC combining SMC with Hamiltonian dynamics in section 6. Convergence diagnostics are detailed in section 7. We contrast HSMC with SMC and other methods in simulation study

in section 8. We then apply both SMC and HSMC to a nonparametric discrete choice model in section 9, comparing the performance of both approaches. Section 10 concludes.

2 Markov Chain Monte Carlo

Consider a general class of models that is parametrized by a Euclidean vector $\theta \in \Theta$ with posterior density $\pi(\theta)$ assumed known up to θ and an integrating constant[5]. Formally, this class of models can be characterized by a family $\mathcal{P}_\theta$ of probability measures on a measurable space $(\Theta, \mathcal{B})$ where $\mathcal{B}$ is the Borel $\sigma-$algebra. The purpose of Markov Chain Monte Carlo (MCMC) methods is to formulate a Markov chain on the parameter space Θ for which, under certain conditions, $\pi(\theta) \in \mathcal{P}_\theta$ is the invariant (also called "equilibrium") distribution. The Markov chain of draws of θ can be used to construct simulation-based estimates of the required integrals and functionals $h(\theta)$ of θ that are expressed as integrals. These functionals include objects of interest for inference on θ such as quantiles of $\pi(\theta)$.

The Markov chain sampling mechanism specifies a method for generating a sequence of random variables $\{\theta_r\}_{r=1}^R$, starting from an initial point θ_0, in the form of conditional distributions for the draws $\theta_{r+1}|\theta_r \sim Q(\theta_r)$. Under relatively weak regularity conditions (Robert and Casella, 2004), the average of the Markov chain converges to the expectation under the stationary distribution:

$$\lim_{R\to\infty} \frac{1}{R}\sum_{r=1}^{R} h(\theta_r) = E_\pi[h(\theta)].$$

A Markov chain with this property is called ergodic. As a means of approximation the analyst relies on large but finite number of draws $R \in \mathbb{N}$ which can be selected in applications based on various criteria.

The conditional distribution $Q(\theta_r)$ is typically derived from a given (economic) model and its corresponding posterior $\pi(\theta)$. In many cases of interest, the model likelihood in $\pi(\theta)$ has a complicated form which precludes direct sampling from $\pi(\theta)$. In such case, the Metropolis-Hastings (M-H) principle is often utilized for drawing $\theta_{r+1}|\theta_r$ from $Q(\theta_r)$; see Chib and Greenberg (1995) for a detailed overview. Suppose we have a proposal-generating density $q(\theta^*_{r+1}|\theta_r)$ where θ^*_{r+1} is a proposed state given the current state θ_r of the Markov chain. The M-H principle stipulates that θ^*_{r+1} be accepted as the next state θ_{r+1} with the acceptance probability

$$\alpha(\theta_r, \theta^*_{r+1}) = \min\left[\frac{\pi(\theta^*_{r+1})q(\theta_r|\theta^*_{r+1})}{\pi(\theta_r)q(\theta^*_{r+1}|\theta_r)}, 1\right], \tag{1}$$

[5]For the sake of simplicity in the notation of this section we suppress the dependence of the posterior on data and other parameters not directly sampled.

otherwise $\theta_{r+1} = \theta_r$. Then the Markov chain satisfies the so-called detailed balance condition

$$\pi(\theta_r)q(\theta^*_{r+1}|\theta_r)\alpha(\theta_r, \theta^*_{r+1}) = \pi(\theta^*_{r+1})q(\theta_r|\theta^*_{r+1})\alpha(\theta^*_{r+1}, \theta_r)$$

which is sufficient for ergodicity. $\alpha(\theta^*_{r+1}, \theta_r)$ is the probability of the move $\theta_r|\theta^*_{r+1}$ if the dynamics of the proposal generating mechanism were to be reversed. The proposal-generating density $q(\theta^*_{r+1}|\theta_r)$ is often chosen to be sampled easily. The popular Gibbs sampler arises as a special case when the M-H sampler is factored into conditional densities. The proposal draws $\theta^*_{r+1}|\theta_r$ from $q(\theta^*_{r+1}|\theta_r)$ in (1) are generated in one step.

3 Sequential Monte Carlo

A key challenge for MCMC methods, in particular in high-dimensional parameter spaces, is to find a good proposal density for the acceptance probability (1). Sequential Monte Carlo, also known as particle filter, encompasses a set of simulation-based methods that address this problem by constructing proposal densities sequentially with a number of desirable properties.

The essence of SMC can be summarized as follows (Herbst and Schorfheide, 2016). Let $p(Y|\theta)$ denote the likelihood and $p(\theta)$ the prior density. The notation $Y_{1:N} = (Y_1, \ldots, Y_N)$ will be used as shorthand for vectors. Let $\{\phi_m\}_{m=1}^{S_\phi}$ be a sequence that slowly increases from zero to one, with $\phi_{S_\phi} = 1$ for some finite S_ϕ. A sequence of posteriors can be constructed by sequentially adding observations to the likelihood function,

$$\pi_m(\theta) = \frac{p(Y_{1:\lfloor \phi_m N \rfloor}|\theta)p(\theta)}{\int p(Y_{1:\lfloor \phi_m N \rfloor}|\theta)p(\theta)d\theta}, \ \phi_m \uparrow 1, \ m = 1, \ldots, S_\phi, \tag{2}$$

where $\lfloor x \rfloor$ is the largest integer that is less than or equal to x. If ϕ_1 is close to zero then the $p(\theta)$ can provide a proposal density for π_1. SMC seeks to efficiently exploit $\pi_m(\theta)$ as a suitable proposal density for $\pi_{m+1}(\theta)$. As a result, SMC algorithms generate weighted draws from the sequence of posteriors $\{\pi_m(\theta)\}_{m=1}^{S_\phi}$. The weighted draws are called *particles*. Denote the overall number of particles by H. At any given stage m, the posterior $\pi_m(\theta)$ is represented by a swarm of weighted particles $\left\{\theta^j_m, w^j_m\right\}_{j=1}^{H}$, where w^j_m is the weight of particle j at stage m, in the sense that for the Monte Carlo average,

$$\overline{h}_m = \frac{1}{H}\sum_{j=1}^{H} w^j_m h(\theta^j) \overset{a.s.}{\to} E_{\pi_m}\left[h(\theta_m)\right].$$

Given the set of particles at stage $m-1$, SMC proceeds in three steps: (i) *correction*: reweighting of the stage $m-1$ particles to reflect the posterior at stage m; (ii) *selection*: resampling the particles with

probable elimination of low-weight particles and multiplication of high-weight particles; and (iii) *mutation*: propagating the particles forward using a Markov transition kernel. The details on each step are given further below in section 6.

4 Hamiltonian Dynamics

Hamiltonian (or Hybrid) Monte Carlo (HMC) is a class of MCMC methods featuring multi-step distant proposals whose path follows the evolution of Hamiltonian dynamics. HMC has its roots in the physics literature where it was introduced for simulating molecular dynamics (Duane et al., 1987). It has since become popular in a number of application areas including statistical physics (Gupta et al., 1988; Akhmatskaya et al., 2009), computational chemistry (Tuckerman et al., 1993), and as a generic tool for Bayesian statistical inference (Neal, 1993, 2011; Ishwaran, 1999; Liu, 2004; Beskos et al., 2010). HMC is most applicable in situations when a suitable importance sampler is not available or practical to implement and one would thus typically need to rely on random walk sampling. HMC has been shown to yield samples far more efficient than obtained by the random walk Metropolis-Hastings mechanism (Rasmussen, 2003; Neal, 2011).

In contrast to the one-step proposals drawn in MCMC, Hamiltonian Monte Carlo (HMC) uses a *sequence* of steps in constructing the proposal whereby the last step in the sequence becomes the proposal draw. The proposal sequence is generated using difference equations of the law of motion yielding high acceptance probability even for proposals that are relatively distant from the current draw in the parameter space. This facilitates efficient exploration of the parameter space with the resulting Markov chain.

Consider a vector of parameters of interest $\theta \in \mathbb{R}^d$ distributed according to the posterior density $\pi(\theta)$. Let $\gamma \in \mathbb{R}^d$ denote a vector of auxiliary parameters with $\gamma \sim N(0, M)$, distributed Gaussian with mean vector 0 and covariance matrix M, independent of θ. Denote the joint density of (θ, γ) by $\pi(\theta, \gamma)$. Then the negative of the logarithm of the joint density of (θ, γ) is given by the Hamiltonian equation[6]

$$H(\theta, \gamma) = -\ln \pi(\theta) + \frac{1}{2} \ln \left((2\pi)^d |M| \right) + \frac{1}{2} \gamma' M^{-1} \gamma. \tag{3}$$

Hamiltonian Monte Carlo (HMC) is formulated in the following three steps that we will describe in detail

[6] In the physics literature, θ denotes the position (or state) variable and $-\ln \pi(\theta)$ describes its potential energy, while γ is the momentum variable with kinetic energy $\gamma' M^{-1} \gamma / 2$, yielding the total energy $H(\theta, \gamma)$ of the system, up to a constant of proportionality. M is a constant, symmetric, positive-definite "mass" matrix which is often set as a scalar multiple of the identity matrix.

further below:

1. Draw an initial auxiliary parameter vector $\gamma_r^0 \sim N(0, M)$;
2. Transition from (θ_r, γ_r) to $(\theta_r^L, \gamma_r^L) = (\theta_{r+1}^*, \gamma_{r+1}^*)$ according to the Hamiltonian dynamics;
3. Accept $(\theta_{r+1}^*, \gamma_{r+1}^*)$ with probability $\alpha(\theta_r, \gamma_r; \theta_{r+1}^*, \gamma_{r+1}^*)$, otherwise keep (θ_r, γ_r) as the next MC draw.

Step 1 provides a stochastic initialization of the system akin to a random walk (RW) draw. This step is necessary in order to make the resulting Markov chain $\{(\theta_r, \gamma_r)\}_{r=1}^R$ irreducible and aperiodic (Ishwaran, 1999). In contrast to RW, this so-called refreshment move is performed on the auxiliary variable γ as opposed to the original parameter of interest θ, setting $\theta_r^0 = \theta_r$. In terms of the HMC sampling algorithm, the initial refreshment draw of γ_r^0 forms a Gibbs step on the parameter space of (θ, γ) accepted with probability 1. Since it only applies to γ, it will leave the target joint distribution of (θ, γ) invariant and subsequent steps can be performed conditional on γ_r^0 (Neal, 2011).

Step 2 constructs a sequence $\{\theta_r^k, \gamma_r^k\}_{k=1}^L$ according to the Hamiltonian dynamics starting from the current state (θ_r^0, γ_r^0) and setting the last member of the sequence as the HMC new state proposal $(\theta_{r+1}^*, \gamma_{r+1}^*) = (\theta_r^L, \gamma_r^L)$. The role of the Hamiltonian dynamics is to ensure that the M-H acceptance probability (1) for $(\theta_{r+1}^*, \gamma_{r+1}^*)$ is kept close to 1. As will become clear shortly, this corresponds to maintaining the difference $-H(\theta_{r+1}^*, \gamma_{r+1}^*) + H(\theta_r^0, \gamma_r^0)$ close to zero throughout the sequence $\{\theta_r^k, \gamma_r^k\}_{k=1}^L$. This property of the transition from (θ_r, γ_r) to $(\theta_{r+1}^*, \gamma_{r+1}^*)$ can be achieved by conceptualizing θ and γ as functions of continuous time t and specifying their evolution using the Hamiltonian dynamics equations[7]

$$\frac{d\theta_i}{dt} = \frac{\partial H(\theta, \gamma)}{\partial \gamma_i} = \left[M^{-1}\gamma\right]_i \ , \tag{4}$$

$$\frac{d\gamma_i}{dt} = -\frac{\partial H(\theta, \gamma)}{\partial \theta_i} = \nabla_{\theta_i} \ln \pi(\theta), \tag{5}$$

for $i = 1, \ldots, d$, where ∇_{θ_i} denotes the derivative of $\ln \pi(\theta)$ with respect to θ_i. For any discrete time interval of duration s, (4)–(5) define a mapping T_s from the state of the system at time t to the state at time $t + s$. For practical applications of interest these differential equations (4)–(5) in general cannot be solved analytically and instead numerical methods are required. The Stormer-Verlet (or leapfrog) numerical

[7] In the physics literature, the Hamiltonian dynamics describe the evolution of (θ, γ) that keeps the total energy $H(\theta, \gamma)$ constant.

integrator (Leimkuhler and Reich, 2004) is one such popular method, discretizing the Hamiltonian dynamics as

$$\gamma(t+\varepsilon/2) = \gamma(t) + (\varepsilon/2)\nabla_\theta \ln \pi(\theta(t)), \tag{6}$$

$$\theta(t+\varepsilon) = \theta(t) + \varepsilon M^{-1}\gamma(t+\varepsilon/2), \tag{7}$$

$$\gamma(t+\varepsilon) = \gamma(t+\varepsilon/2) + (\varepsilon/2)\nabla_\theta \ln \pi(\theta(t+\varepsilon)), \tag{8}$$

for some small $\varepsilon \in \mathbb{R}$. From this perspective, γ plays the role of an auxiliary variable that parametrizes (a functional of) $\pi(\theta, \cdot)$ providing it with an additional degree of flexibility to maintain the acceptance probability close to one for every k. Even though $\ln \pi(\theta_r^k)$ can deviate substantially from $\ln \pi(\theta_r^0)$, resulting in favorable mixing for θ, the additional terms in γ in (3) compensate for this deviation maintaining the overall level of $H(\theta_r^k, \gamma_r^k)$ close to constant over $k = 1, \ldots, L$ when used in accordance with (6)–(8), since $\frac{\partial H(\theta,\gamma)}{\partial \gamma_i}$ and $\frac{\partial H(\theta,\gamma)}{\partial \theta_i}$ enter with the opposite signs in (4)–(5). In contrast, without the additional parametrization with γ, if only $\ln \pi(\theta_r^k)$ were to be used in the proposal mechanism as is the case in RW style samplers, the M-H acceptance probability would often drop to zero relatively quickly.

Step 3 applies a Metropolis correction to the proposal $(\theta_{r+1}^*, \gamma_{r+1}^*)$. In continuous time, or for $\varepsilon \to 0$, (4)–(5) would keep $-H(\theta_{r+1}^*, \gamma_{r+1}^*) + H(\theta_r, \gamma_r) = 0$ exactly resulting in $\alpha(\theta_r, \theta_{r+1}^*) = 1$ but for discrete $\varepsilon > 0$, in general, $-H(\theta^*, \gamma^*) + H(\theta, \gamma) \neq 0$ necessitating the Metropolis step. A key feature of HMC is that the generic M-H acceptance probability (1) can be expressed in a simple tractable form using only the posterior density $\pi(\theta)$ and the auxiliary parameter Gaussian density $\phi(\gamma; 0, M)$. The transition from (θ_r^0, γ_r^0) to (θ_r^L, γ_r^L) via the proposal sequence $\{\theta_r^k, \gamma_r^k\}_{k=1}^L$ taken according to the discretized Hamiltonian dynamics (6)–(8) is a deterministic proposal, placing a Dirac delta probability mass $\delta(\theta_r^k, \gamma_r^k) = 1$ on each (θ_r^k, γ_r^k) conditional on (θ_r^0, γ_r^0). The system (6)–(8) is time reversible and symmetric in (θ, γ), which implies that the forward and reverse transition probabilities $q(\theta_r^L, \gamma_r^L | \theta_r^0, \gamma_r^0)$ and $q(\theta_r^0, \gamma_r^0 | \theta_r^L, \gamma_r^L)$ are equal: this simplifies the Metropolis-Hastings acceptance ratio in (1) to the Metropolis form $\pi(\theta_{r+1}^*, \gamma_{r+1}^*)/\pi(\theta_r^0, \gamma_r^0)$. From the definition of the Hamiltonian $H(\theta, \gamma)$ in (3) as the negative of the log-joint densities, the joint density of (θ, π) is given by

$$\pi(\theta, \gamma) = \exp\left[-H(\theta, \gamma)\right] = \pi(\theta)\left((2\pi)^d |M|\right)^{-1/2} \exp\left(-\frac{1}{2}\gamma' M^{-1}\gamma\right). \tag{9}$$

Hence, the Metropolis acceptance probability takes the form

$$\begin{aligned}\alpha(\theta_r, \gamma_r; \theta_{r+1}^*, \gamma_{r+1}^*) &= \min\left[\frac{\pi(\theta_{r+1}^*, \gamma_{r+1}^*)}{\pi(\theta_r^0, \gamma_r^0)}, 1\right] \\ &= \min\left[\exp\left(-H(\theta_{r+1}^*, \gamma_{r+1}^*) + H(\theta_r^0, \gamma_r^0)\right), 1\right].\end{aligned}$$

The expression for $\alpha(\theta_r, \gamma_r; \theta^*_{r+1}, \gamma^*_{r+1})$ shows, as noted above, that the HMC acceptance probability is given in terms of the difference of the Hamiltonian equations $H(\theta^0_r, \gamma^0_r) - H(\theta^*_{r+1}, \gamma^*_{r+1})$. The closer can this difference be kept to zero, the closer the acceptance probability approaches one. A key feature of the Hamiltonian dynamics (4)–(5) in Step 2 is that they maintain $H(\theta, \gamma)$ constant over the parameter space in continuous time conditional on $H(\theta^0_r, \gamma^0_r)$ obtained in Step 1, while their discretization (6)–(8) closely approximates this property for discrete time steps $\varepsilon > 0$ with a global error of order ε^2 corrected by the Metropolis update in Step 3 (Neal, 2011). The goal of the Hamiltonian proposal transition dynamics is thus to maintain the proposal acceptance probability at or close to one even for a relatively long proposal sequence.

The acceptance ratio can only be maintained at exactly one if the proposal trajectory evolution were continuous. However, due its discretization into individual steps, the acceptance probability in general deviates from one due to discretization errors. The length of the proposal sequence can then be tuned using $\varepsilon > 0$ and L to achieve a desired acceptance rate, analogously to the RW environment. The Hamiltonian dynamics approximately keeps the joint density $\pi(\theta, \gamma)$ of θ and γ constant, permitting changes in the marginal density $\pi(\theta)$. Due to this feature, the proposal sequence does not move along a "straight" trajectory in the parameter space Θ of θ, but rather along a "curve". This ensures that the proposal sequence does not travel "too far" into the tails and stays in regions with non-zero probability.

Each proposal sequence in HMC and its extensions starts with a "refreshment" of the kinetic auxiliary variable γ newly drawn from the Gaussian distribution $N(0, M)$ where M is the mass matrix. This draw determines the "direction" in which the proposal sequence will propagate through the parameter space. The stochastic nature of γ prevents the chain from getting stuck at the original point or too close to it.

4.1 Constraints

Parameter space constraints can be incorporated into the HMC proposal mechanism via "hard walls" representing a barrier against which the proposal sequence, simulating a particle movement, bounces off elastically. Constraints thus do not provide grounds for proposal rejection, eliminating any associated redundancies. Heuristically, the constraint is checked at each step of the proposal sequence and if it is violated for any dimension of the parameter vector then the proposal dynamics are negated for that dimension and the trajectory of the sequence is thus reflected off the hard wall posed by the constraint. This facilitates efficient

exploration of the parameter space even in parameter spaces that are constrained in a complex way.

5 Bayesian Nonparametric Mixture Modeling

Consider an exchangeable sequence $Y \equiv Y_{1:N}$ of random variables defined over a measurable space $(\Phi, \mathcal{D})$ where $\mathcal{D}$ is a σ-field of subsets of Φ. Denote the joint density of Y implied by an economic model by $f(Y, \theta)$ where $\theta \in \Theta$ is a Euclidean parameter. Further denote by G_0 the prior distribution of θ over a measurable space $(\Theta, \mathcal{B})$ with $\mathcal{B}$ being a σ-field of subsets of Θ, and denote by g_0 the density associated with G_0.

In a parametric Bayesian model, the joint density of Y and θ is defined as

$$p(Y, \theta; G_0) = f(Y, \theta) g_0, \tag{10}$$

Conditioning on observed realizations y of Y turns $f(Y, \theta)$ into the likelihood function $p(\theta|y)$ and $p(Y, \theta; G_0)$ into the posterior density $\pi(\theta|G_0, y)$.

In the class of nonparametric Bayesian mixture models[8] considered here, the joint density of Y and θ is defined as a mixture

$$p(Y, \theta, G) = \int f(Y;\theta) dG(\theta),$$

where G is the mixing distribution over θ. The distribution G is now random which leads to a flexibility of the resulting mixture model. The model parameters θ are no longer restricted to follow any given pre-specified distribution as was stipulated by the fixed G_0 in the parametric case. The parameter space now also includes the random infinite-dimensional G with the additional need for a prior distribution for G. The Dirichlet Process prior is a popular alternative due to its numerous desirable properties; we proceed with its description in the next section.

5.1 Dirichlet Process Prior

In a seminal paper, Fergusson (1973) introduced the Dirichlet process (DP) prior for random measures whose support is large enough to span the space of probability distribution functions and that leads to analytically manageable posterior distributions. Antoniak (1974) further elaborated on using the DP as the prior for the mixing proportions of a simple distribution.

[8] A commonly used technical definition of nonparametric Bayesian models are probability models with infinitely many parameters (Bernardo and Smith, 1994).

A DP prior for G is determined by two parameters: a distribution G_0 that defines the "location" of the DP prior, and a positive scalar precision parameter α. The distribution G_0 may be viewed as a baseline prior that would be used in a typical parametric analysis. The flexibility of the DP prior model environment stems from allowing G – the actual prior on the model parameters – to stochastically deviate from G_0. The precision parameter α determines the concentration of the prior for G around the DP prior location G_0 and thus measures the strength of belief in G_0. For large values of α, a sampled G is very likely to be close to G_0, and vice versa.

More specifically, let $\mathcal{M}(\Psi)$ be a collection of all probability measures on Ψ endowed with the topology of weak convergence. The space $\mathcal{M}(\mathcal{M}(\Psi))$ is then the collection of all probability measures (i.e. priors) on $\mathcal{M}(\Psi)$ together with the topology of weak convergence derived from $\mathcal{M}(\Psi)$. Let $G_0 \in \mathcal{M}(\Psi)$ and let α be a positive real number. Following Fergusson (1973), a *Dirichlet Process* on $(\Psi, \mathcal{B})$ with a base measure G_0 and a concentration parameter α, denoted by $DP(G_0, \alpha) \in \mathcal{M}(\mathcal{M}(\Psi))$, is a distribution of a random probability measure $G \in \mathcal{M}(\Psi)$ over $(\Psi, \mathcal{B})$ such that, for any finite measurable partition $\{\Psi_i\}_{i=1}^{J}$ of the sample space Φ, the random vector $(G(\Psi_1), ..., G(\Psi_J))$ is distributed as $(G(\Psi_1), ..., G(\Psi_J)) \sim Dir(\alpha G_0(\Psi_1), ..., \alpha G_0(\Psi_J))$ where $Dir(\cdot)$ denotes the Dirichlet distribution. We write $G \sim DP(G_0, \alpha)$ if G is distributed according to the Dirichlet process $DP(G_0, \alpha)$.

5.2 Dirichlet Process Mixture Model

Bayesian nonparametric mixture models have been widely applied to solving problems such as clustering, density estimation and topic modeling. These models make relatively very weak assumptions about the underlying process that generated the observed data. When more data are collected, the complexity of these models can change accordingly. In the Bayesian mixture modeling framework it is possible to infer the number of components to model the data and therefore it is unnecessary to explicitly restrict the number of components a-priori (Görür and Rasmussen, 2010).

For a nonparametric continuous density estimation the discrete Dirichlet process is typically convolved with a continuous kernel. There are many various ways of doing so. We follow the approach laid out by Ghosal and van der Vaart (2017, section 5.1) based on previous literature on Bayesian nonparametrics cited therein. For each $\theta \in \Theta \subset \mathbb{R}^d$, let $f(Y|\theta)$ be a probability density function of Y, where Y is an observable random variable. The density (10) where G is endowed with the Dirichlet process prior, is known as a Dirichlet

process mixture (DPM). Realizations of the DP are discrete with probability one and hence a DPM can be viewed as a countably infinite mixture (Ghosal and van der Vaart, 2017).

For a sample size N, let Y_i with $i = 1, \ldots, N$ be distributed according to the density kernel

$$p_{i,G} = \int f_i(Y_i|\theta) dG(\theta),$$

where $G \sim DP(G_0, \alpha)$. The resulting model can be equivalently written in terms of N latent variables θ_i as

$$\begin{aligned} Y_i|\theta_i, G &\sim f_i(Y_i|\theta_i), \\ \theta_i|G &\sim G, \\ G &\sim DP(G_0, \alpha). \end{aligned}$$

The model can also be represented in terms of allocation variables $s_1, ..., s_N$ that link the observations to the components of the mixture model:

$$Y_i|s_i^* = k \sim f_i(Y_i|\theta_k^*),\ i = 1, \ldots, N,$$

where, conditional on G, s_i^* and θ_k^* are the distinct values of s_i and θ_k, respectively. MCMC posterior inference for DPM models has been detailed in and number of studies, including Neal (2011) and Ghosal and van der Vaart (2017).

6 Hamiltonian Sequential Monte Carlo

Here we first provide the details of an SMC algorithm suited for our context and then propose its extension to form HSMC. SMC generally consists of three phases, as described above: (i) *correction*, (ii) *selection*, and (iii) *mutation*. The state-of-the-art procedure for the *correction* phase for a Bayesian static nonparametric model with a Dirichlet Process (DP) prior and a non-conjugate likelihood is Algorithm 2 of Griffin (2017), which we use for particle *correction* in both SMC and HSMC:

Correction phase:
Let $m_{i,k}$ denote the number of s_i associated with the mixture component k, let m_0 denote the prior value for $m_{N,k}$ for all k, and let K_i denote the number of mixture components for $s_{1:i}$. For $i = 1, ..., N$:

Step 1: For all particles $j = 1, ..., H$, perform steps 1a and 1b.

Step 1a: Sample $\theta_{new} \sim G_0$, and $s_i^{*(j)}$ conditional on $y_{1:i}$ and $s_{1:(i-1)}^{*(j)}$ from

$$q(k) \propto \begin{cases} m_{k,i-1}^{(j)} f_i(y_i|\theta_k^{*(j)}) \text{ if } 1 \leq k \leq K_{i-1}^{(j)} \\ m_0 f_i(y_i|\theta_{new}) \text{ if } k = K_{i-1}^{(j)} + 1. \end{cases}$$

Step 1b: Calculate the unnormalized weight

$$\xi_i^{(j)} = m_0 f_i(y_i|\theta_{new}) + \sum_{k=1}^{K_{i-1}^{(j)}} m_{k,i-1}^{(j)} f_i(y_i|\theta_k^{*(j)})$$

Step 2: Reweight the particles according to the weights

$$w_i^{(j)} = \frac{\xi_i^{(j)}}{\sum_{j=1}^{N_j} \xi_i^{(j)}}.$$

Although a number of alternative *selection* schemes have been proposed in the literature, we utilize the popular Residual Resampling as described in Chopin (2004).

Selection phase:
Reproduce each particle $int\{Hw_i^{(j)}\}$ times, where $int\{x\}$ stands for the integer part of x. Complete the particle vector by $H^r = H - \sum_j int\{Hw_i^{(j)}\}$ independent draws from the multinomial distribution which reproduces the jth particle with probability $(Hw_i^{(j)} - int\{Hw_i^{(j)}\})/H^r$.

Mutation phase:
Propagate each parameter $\theta_k^{*(j)}$ for $j = 1, ..., R_j$ and $k = K_i^{(j)}$ according to Hamiltonian transition dynamics as described in section 4.

In contrast, SMC uses the random walk transition kernel in the mutation phase. To the best of our knowledge, HSMC has not been proposed in the previous literature.

In the simulation study below and the application we demonstrate that HSMC attains superior mixing and convergence properties over SMC with parameter dimensionality of up to about 20 parameters. Performance of either sampler when the parameter dimensionality is in the order of 150-200 or more is currently an open question and we will seek to address it in future research.

7 Convergence Diagnostics

For the assessment of the performance a posterior sampler, it is standard practice to rely on convergence diagnostics obtained by examining the sampling output (Cowles and Carlin, 1996). A typical MCMC diagnostic starts several Markov chains at overdispersed initial values, and monitors convergence by comparing between-chain and within-chain variances for selected scalar quantities (Plummer et al., 2006). However, the bulk of such diagnostics is not applicable to particle-based samplers, including SMC and HSMC, as a significant proportion of sample chain paths are discontinued during the implementation during the resampling phase. In the absence of chains of parameter draws of full equal length the chains cannot be compared in terms of their sampling behavior.

For the purpose of comparing the convergence properties of HSMC and SMC we will use the *Partition-based approximation for convergence evaluation* (PACE) diagnostic procedure developed by VanDerwerken and Schmidler (2017) to address the limitations of other diagnostics discussed above. The PACE statistic involves initializing J sets of particles at overdispersed locations in the state space. At a given iteration, the trajectories of all particle draws are pooled together. The distance between the sample distributions of each particle set is quantified by comparing the within-set and across-set probabilities over a partition of the parameter space. When the particle sets are stationary, the proportion of within-set draws belonging to a given partition element will be approximately equal across the particle sets. Heuristically, when each particle set results in approximately the same posterior probability for a given partition element of the parameter space then the sets can be regarded as having converged. In contrast, when different particle sets imply different posterior probabilities for a given parameter space partition element then none of the sets can be guaranteed to have converged.

The PACE statistic is based on a comparison of approximate posterior probabilities over a parameter space partition. The posterior probabilities can be quantified using any number of particles without requiring that the chains of individual particle draws be of full equal length. The particles are thus free to be resampled in the SMC selection phase. Furthermore, the parameter space can be either Euclidean or a function space which renders PACE suitable for nonparametric estimation problems.

VanDerwerken and Schmidler (2017) proposed the PACE statistic using an adaptive parameter space partition whereby pooled sampler draws are suitably clustered in order to construct the partition. We used a

simpler version in which the partition is constructed over a fixed equidistant grid over the parameter space in a non-adaptive manner. This saves on computation time substantially and avoids introducing an ad-hoc clustering procedure which may act differently in HSMC and SMC obscuring the differences stemming from these two procedures alone. Correspondingly, we used a mean-absolute deviation (MAD) measure as the PACE distance function.

8 Simulation Study

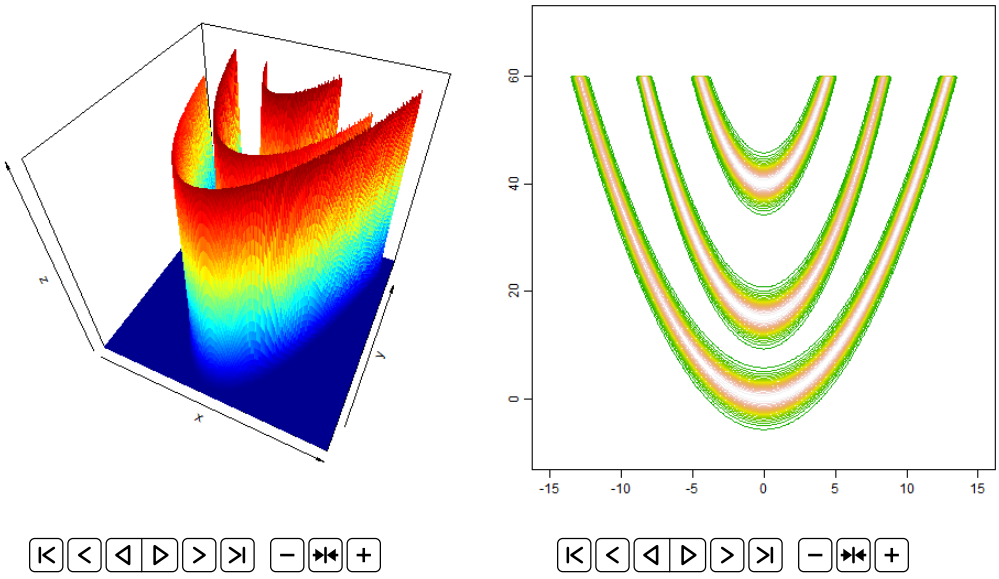
In this section we demonstrate the behaviour of HSMC and compare it with SMC and other alternative approaches. Figure 1 shows the posterior density kernel from which we would like to obtain a vector of parameter draws. The posterior is multi-modal, with each mode consisting of a re-scaled negative Rosenbrock function (Rosenbrock, 1960). Each mode is skewed with a well-defined modal value and an elongated bending shape from which it is relatively challenging to sample. The modes are separated by a relatively low probability regions, which makes it difficult to transition among the modes; indeed this is one of the drawbacks of HMC. All modes share a hard parameter constraint at y-value of 60.

The animations[9] of Figure 1 show posterior convergence. Starting from a diffuse prior, as more *iid* random variables are accumulated in the likelihood, the posterior becomes sharper. In Sequential Monte Carlo methods the particles explore the posterior from its early relatively diffuse stages, gradually discovering all modal regions. This is the key to the benefits of the sequential nature SMC and HSMC in both cross-sections and time series models. In contrast, if information were not added sequentially but instead all at once, the sampler would be faced directly with the sharp target posterior upon initialization. The initial particle vector might miss some of the modes, especially in higher dimensions, and never transition into them during the Markov chain run as this can require passing through low probability regions.

The performance of both SMC and HSMC may depend on the order in which the data are added into the algorithm. A new significant mode contained in the data at the end of the sample may not be imminently discovered by either approach. It is therefore advisable to run either algorithm several times with the dataset randomly re-ordered in the cross-sectional dimension and check the results for robustness of the modal features upon convergence.

[9] All animations were created using the R package *animation* (Xie, 2013) and imported to LaTeXusing the package *animate* (Grahn, 2021).

Figure 1: Posterior convergence
3 dimensions animation (left) and 2 dimensions contour animation (right)

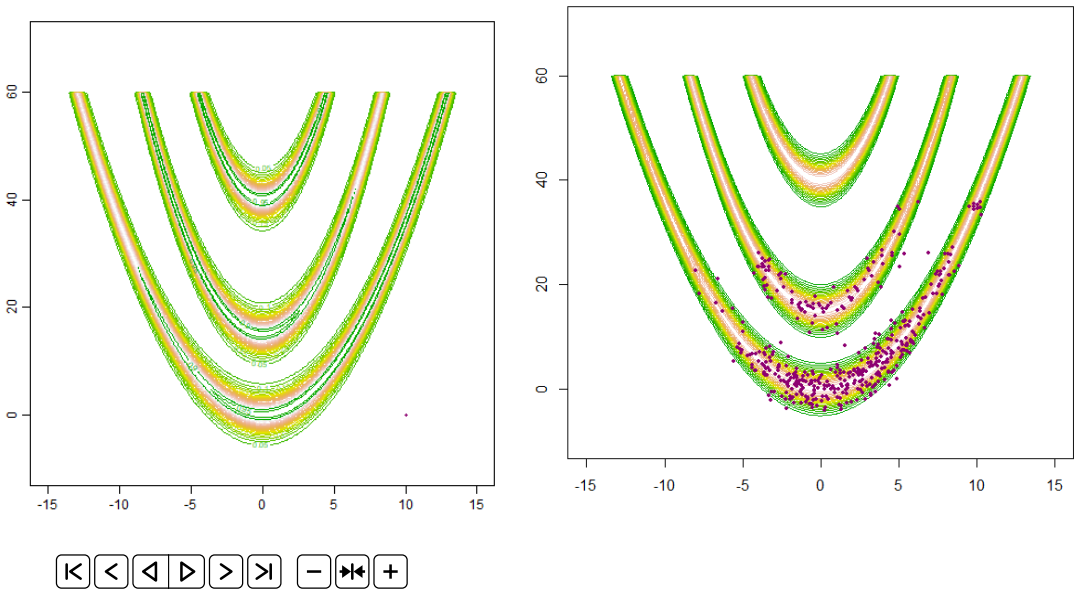


Both graphics are animated. In Adobe Acrobat or Adobe Reader click on the "play" triangle button to start the animations. In a web browser run the (double-blind) animation at this link.

In all simulation Figures below, the animation on the left shows the first 75 RW-MH draws, while the graph on the right displays all subsequent 1,000 draws, referred to as "the full run". All samplers were tuned to a typical configuration. Samplers with RW proposals use independent multivariate normal density draws with variance scaled to achieve about 30% acceptance rate and samplers with HMC proposals contain 20 proposal steps with ε scaled to achieve about 50% acceptance rate.

The first benchmark sampler is the random walk Metropolis-Hastings (RW-MH), with output shown in Figure 2. The sampler tends to stay for a relatively long time in a single node before randomly jumping to the next one. During the full run it has not explored the node tails and missed the third node completely.

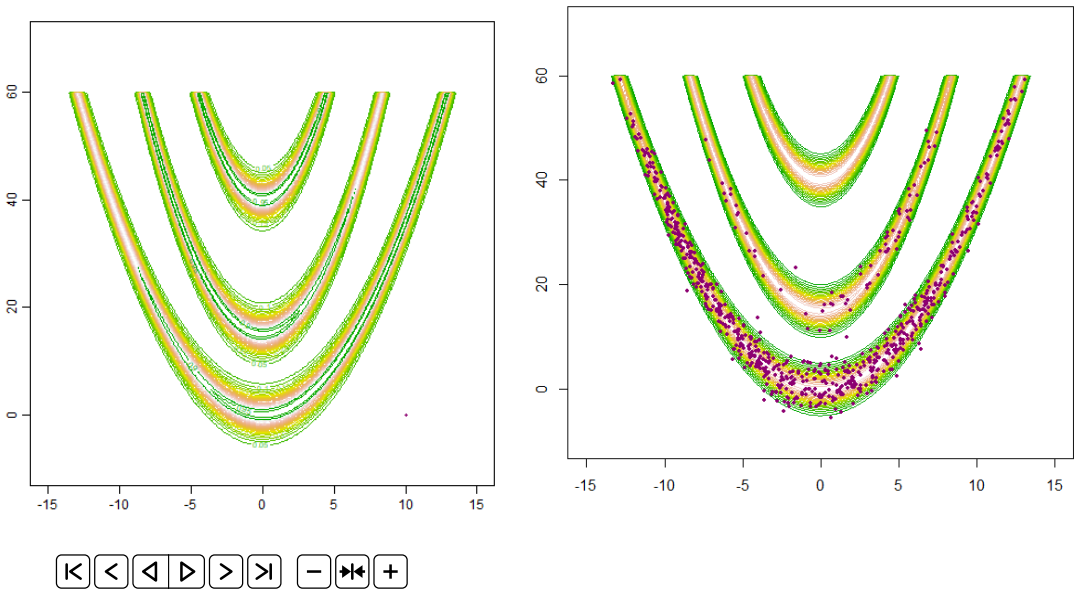
Figure 2: RW-MH
2 dimensions animation (left) and particles after completed run (right)



The left graphics is animated. In Adobe Acrobat or Adobe Reader click on the "play" triangle button to start the animation. In a web browser run the (double-blind) animation at this link.

The second benchmark sampler is Hamiltonian Monte Carlo (HMC). Figure 3 shows the first 75 HMC draws in the animation on the left, and all the subsequent draws up to 1,000 in the graph on the right. A well-known advantage of HMC over RW-MH is that it explores posterior tails efficiently, with superior mixing properties; this is evident from the full run graph on the right as compared to the RW-MH graph. However, HMC does not transition well among multiple modes through low probability regions. While the sampler found its way between the two lower nodes it missed the top node completely during the full simulation run.

Figure 3: HMC
2 dimensions animation (left) and particles after completed run (right)



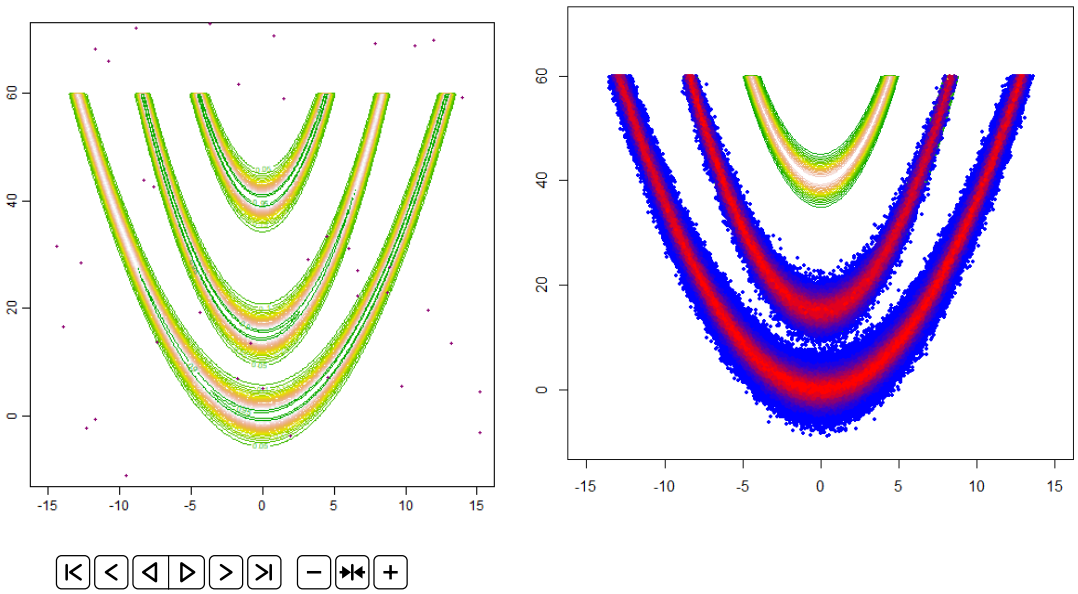
The left graphics is animated. In Adobe Acrobat or Adobe Reader click on the "play" triangle button to start the animation. In a web browser run the (double-blind) animation at this link.

The next benchmark sampler is what we will call Parallel Hamiltonial Monte Carlo (PHMC). It is a special case of our proposed HSMC where information is *not added sequentially* but rather all at once. Thus, instead of one parameter vector draw in a single Monte Carlo iteration, PHMC goes through all SMC stages (correction, selection, mutation) where the latter is performed by HMC. We include PHMC here to highlight the benefits of sequential accumulation of information in HSMC and SMC. The sampler was tuned with a typical configuration: it contains 20 proposal steps with the proposal ε scaled to achieve about 50% acceptance rate. The output of the particle-based methods is color coded on the red-blue scale. A particle with lower weight is colored towards the blue end of the scale while a particle with higher weight is colored towards the red end.

When PHMC is initialized and a particle happens to land in a modal region of the parameter space it will attain relatively high weight in the correction stage, generate more copies of itself in the selection stage, and typically stay in the same modal region in the mutation stage. If no particle is initialized in a modal region

that region will likely be missed by the sampler altogether as other particles are unlikely to transverse to it by crossing through low probability areas. Figure 4 shows precisely this scenario. The first frame of the animation on the left shows particle initialization. In this simulation 500 particles were spread randomly over the area $(-100, 100) \times (-100, 100)$ as could be expected of an analyst a-priori unaware of the location of the modes. No particle has landed in the top mode while some particles happened to hit the middle and bottom nodes. As the animated run shows, all particles are quickly distributed over the latter nodes and HMC mutation moves ensure efficient exploration of their tail areas. However, the no particles have crossed over to the top node during the full run. As apparent from the graph on the right, the top node was missed completely. This was typical situation from the many simulation runs that we have conducted.

Figure 4: PHMC
2 dimensions animation (left) and particles after completed run (right)

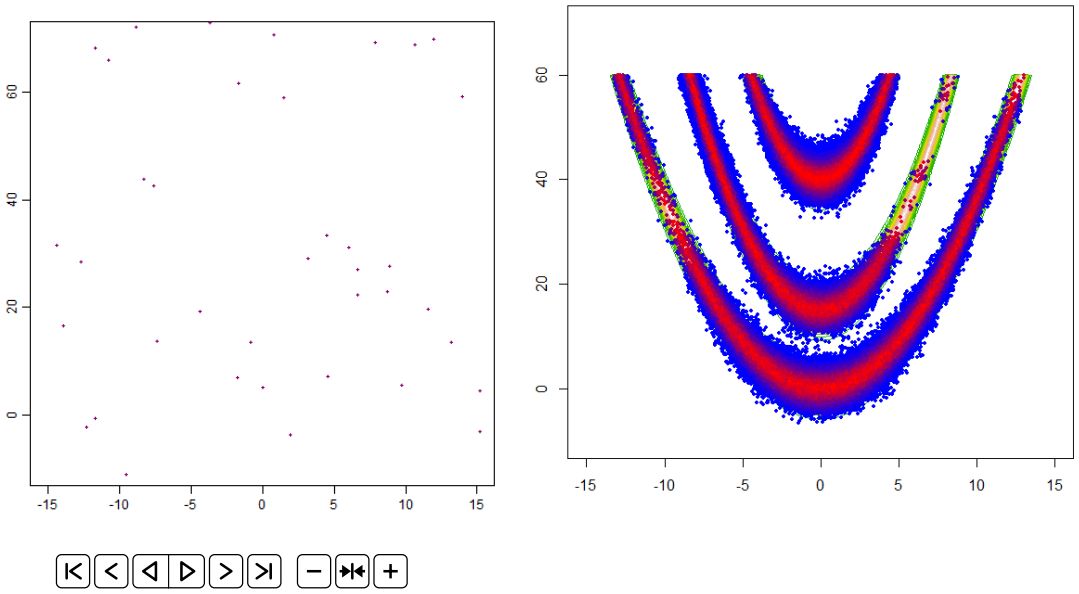


The left graphics is animated. In Adobe Acrobat or Adobe Reader click on the "play" triangle button to start the animation. In a web browser run the (double-blind) animation at this link.

In SMC information is accumulated gradually, by adding one or a relatively small batch of random variables into the likelihood. Starting from a diffuse prior, the posterior surface thus gets sharper, as shown above in Figure 1 in our simulation, while the particles explore its surface. During the accumulation phase particles can cross over into the forming modal regions with relatively high probability. This progression is shown in

the left animation of Figure 5, where all modes are discovered, in contrast to PHMC. Nonetheless, as SMC uses the RW-MH sampler for each particle transition in the mutation phase, the same drawback of RW-MH as shown above in Figure 3 transpires also in SMC: the tails of the posterior are not explored efficiently, as is apparent from the full run graph of Figure 5 on the right.

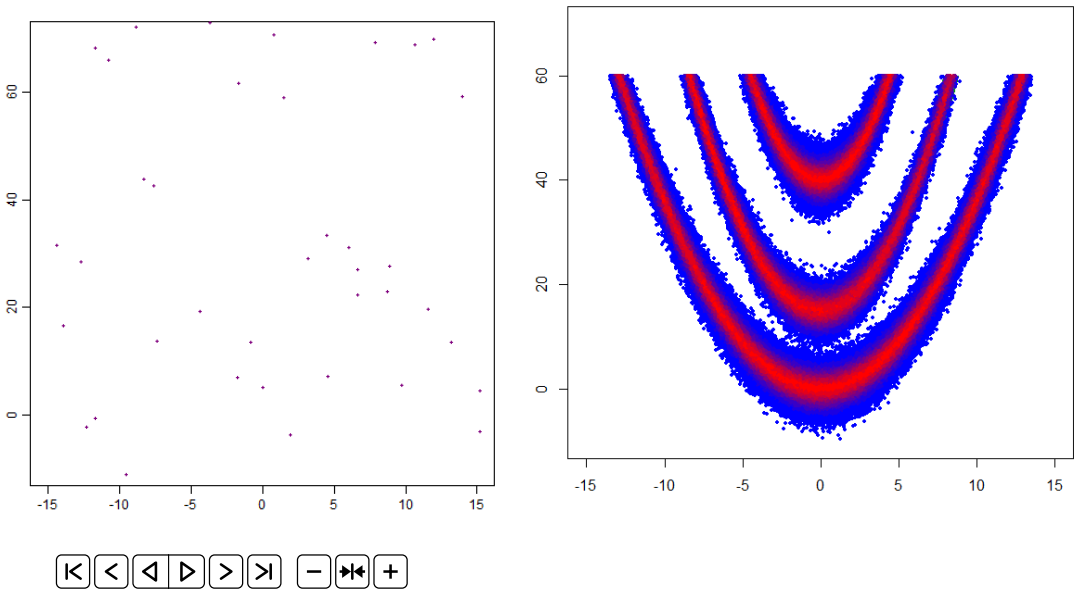
Figure 5: SMC
2 dimensions animation (left) and particles after completed run (right)



The left graphics is animated. In Adobe Acrobat or Adobe Reader click on the "play" triangle button to start the animation. In a web browser run the (double-blind) animation at this link.

The drawback of SMC is remedied in HSMC by mutation moves using HMC transition dynamics. The sequential nature of the sampler enables the particles to discover all modal regions, unlike PHMC. In contrast with SMC, HSMC's Hamiltonian transitions ensure efficient exploration and superior mixing of the posterior tails, as shown in Figure 6.

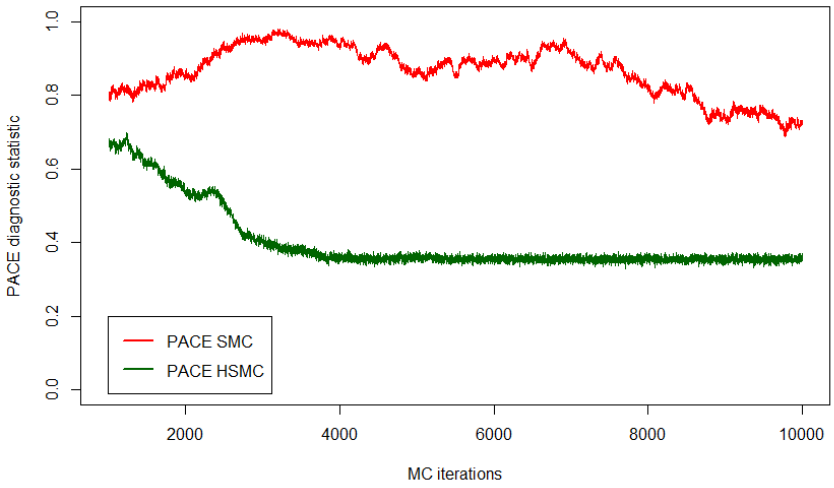
Figure 6: HSMC
2 dimensions animation (left) and particles after completed run (right)



The left graphics is animated. In Adobe Acrobat or Adobe Reader click on the "play" triangle button to start the animation. In a web browser run the (double-blind) animation at this link.

The PACE convergence statistics for SMC and HSMC are plotted in Figure 7 as functions of Monte Carlo iterations. For the diagnostics we used 20 runs with 500 particles each. The PACE statistic then effectively compares their coverage of the posterior. While HSMC has converged well within about the first quarter of the run, SMC has not reached convergence. In SMC each set of particles appears to cover different areas of the modal regions of the posterior, as seen in Figure 3, which is then reflected in the PACE statistic. In contrast, upon relatively fast convergence in HSMC each set of particles covers and efficiently mixes over the entire posterior.

Figure 7: PACE



9 Application: The Nonparametric Mixed Logit Model

The mixed logit can approximate any random utility model (McFadden and Train, 2000) and remains popular among practitioners for its analytical tractability. Mixed logit models can be obtained under different behavioral specifications, and each derivation provides a particular interpretation of the model fundamentals. Any behavioral specification whose choice probabilities take its particular form is called a mixed logit model (Train, 2009). With traditional MCMC, the nonparametric mixed logit has been difficult to implement in more than three dimensions without imposing further restrictive assumptions (Burda et al., 2008). Here HSMC enables us to increase the number of nonparametric dimensions to close to twenty.

9.1 Model Environment

There are N individuals, $i = 1, \ldots, N$, choosing in each of T time periods, $t = 1, \ldots, T$, one out of J alternatives, $j = 1, \ldots J$. Let y_{it} denote the choice of individual i at time t. The latent utility of individual i at time t of choice j is given by

$$u_{itj} = \beta_i' \mathbf{x}_{itj} + \varepsilon_{itj},$$

with latent iid residual $\varepsilon_{itj} \sim F_\varepsilon$ where F_ε is the Extreme Value Type 1 distribution. The first element of β_i is normalized to zero for identification purposes. Then, conditional on the vector of covariates $x_{it} =$

$(x_{it1}, \ldots, x_{itJ})'$ and the vector of parameters β_i, the probability of choosing y_{it} at time t is given by

$$L_{it}(y_{it}|\beta_i, \mathbf{x}_{it}) = \frac{\exp(\beta_i' \mathbf{x}_{ity_{it}})}{\sum_{j=1}^{J} \exp(\beta_i' \mathbf{x}_{itj})},$$

and the probability of choosing the vector $y_i = (y_{i1}, \ldots, y_{iT})'$ is given by

$$K(y_i|\beta_i, \mathbf{x}_i) = \prod_{t=1}^{T} L_{it}(y_{it}|\beta_i, \mathbf{x}_{it}).$$

The mixed logit model specification is obtained by expressing the choice probabilities in the form

$$P_i(y_i|\mathbf{x}_i) = \int K(y_i|\beta_i, \mathbf{x}_i) f(\beta_i) d\beta_i \ .$$

The mixed logit probability is a weighted average of the logit formula evaluated at different values of β_i, where the weights are given by the density $f(\beta_i)$.

Under the Bayesian nonparametric mixture model approach, we specify the model for the distribution of β_i as follows:

$$\begin{aligned} y_i|\beta_i &\sim K(y_i|\beta_i, \mathbf{x}_{it}), \\ \beta_i|G &\sim G, \\ G &\sim DP(G_0, \alpha). \end{aligned}$$

with G_0 a standard Normal distribution and α obtained implicitly by setting $m_0 = 1$.

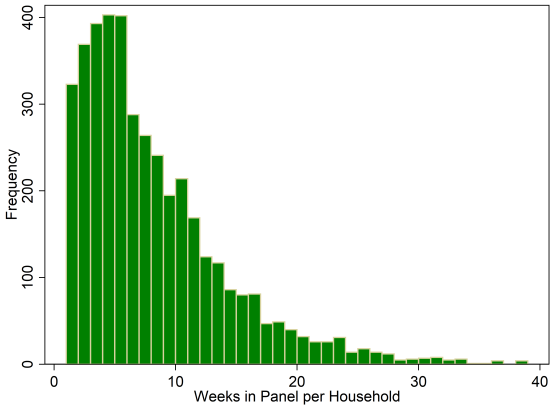
Fox et al. (2012) showed that the mixed logit model is nonparametrically identified. Fox and Gandhi (2016) analyze a nonparametric estimator for the case when the observable random variables have a discrete support. Fox et al. (2016) propose a computationally attractive projection-based estimator of the joint distribution of random coefficients over a fixed support grid in structural models including the mixed logit. The Bayesian framework allows for continuous support of observable random variables and does not impose the fixed support grid restriction on the parameter space.

In Bayesian multinomial choice modeling, MCMC has so far been the dominant approach to inference (Kim et al., 2004; Burda et al., 2008; Keane and Wasi, 2013; Li and Ansari, 2014). Although SMC has been utilized in analysis of generic Bayesian DPM models, we are not aware of its application to Bayesian multinomial discrete choice model. In the sequel, we will estimate the distribution of β_i by both HSMC and SMC in a real-world application. We will then evaluate and compare the convergence properties of both methods.

9.2 Data

Our empirical analysis is based on the IRI Academic Dataset (Bronnenberg et al., 2008), containing panel data of grocery stores purchases in two U.S. cities. We chose to focus on the purchases of mayonnaise since this product category is composed of relatively few well defined homogenous items. In our data set, two dominating brands cover 87% of the market: Hellman's (46%) and Kraft (41%). The remainder of the market is served by "private label" (8%), Cains (3%), and "other" (2%). We use the time period from June 2010 through December 2012, totalling 138 weeks. During this time period, the data contains a stable choice set without introducing new or discontinuing old products in the set of the choice alternatives. Each of the 2,684 households in our sample was recorded as making mayonnaise purchases on average for 7.86 weeks. The distribution of weeks observed in the sample for all households is shown in Figure 8.

Figure 8: Weeks Observed Making Mayonnaise Purchases



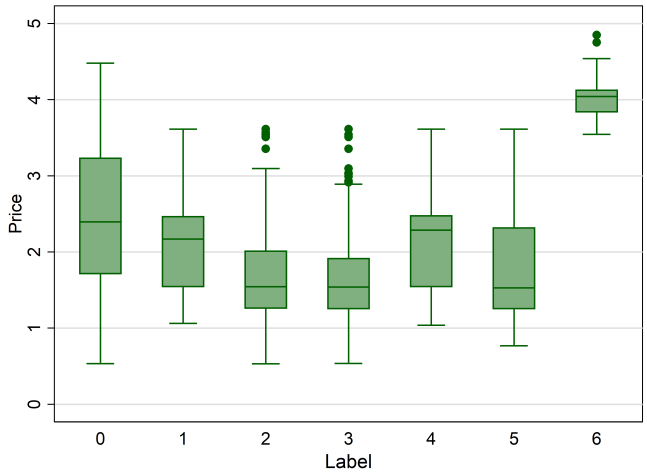
Similarly to Thomadsen (2016), we assume that consumers choose among a set of "top" alternatives, or else choose an outside option if they choose a product in the category that does not belong to a top alternative. The "top" alternatives are selected by ranking the alternatives by the number of purchases made by the panelists, include alternatives starting from the most popular ones and going in decreasing order of popularity until the set of included top alternatives covers all major ones. Thus, we consider the following six alternatives plus an outside option, as given in Table 1.

Table 1: Choice Set

Label	*Name*	*Frequency*	*%*	*Cum. %*
1	Hellmann's Real Mayonnaise	7,041	24.42	24.42
2	Kraft Miracle Whip	6,189	19.88	44.30
3	Kraft Miracle Whip Low Fat	3,031	11.35	55.65
4	Hellmann's Light Mayonnaise	1,752	5.26	60.91
5	Kraft Soybean Mayonnaise	1,015	4.43	65.34
6	Hellmann's Soybean Mayonnaise	786	2.05	67.38
0	outside option	12,533	32.62	100.00

The panel contains information about product attributes and consumer characteristics. Given the high degree of product homogeneity within any given category, in addition to brand we only included price among the product attributes. The price dispersion for each product in the choice set is shown in Figure 9. The Label of the alternatives corresponds to the product name code listed in Table 1. Although on a given choice occasion price is only observed for the selected alternative, we infer the prices of the remaining alternatives of the choice set from observations of other customers who selected such alternatives in any given store.

Figure 9: Choice Set Price Dispersion

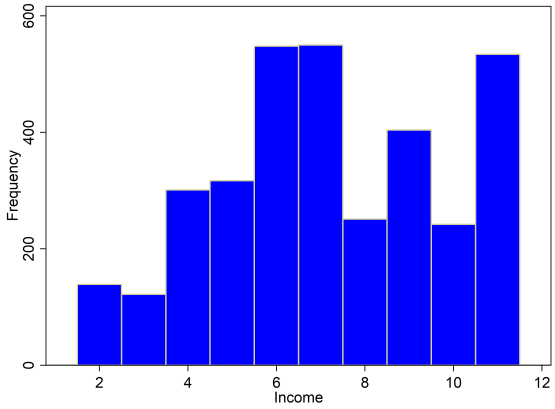


In order to keep dimensionality of the parameter vector low, we have selected income as a key consumer characteristic. Table 2 specifies the ordinal coding for the income ranges in our dataset and Figure 10 shows a histogram of the income data codes in our sample. We have dropped all households whose income data was missing.

Table 2: Income Codes

Code	*Household Income per Year*
1	\$00,000 to \$ 9,999
2	\$10,000 to \$11,999
3	\$12,000 to \$14,999
4	\$15,000 to \$19,999
5	\$20,000 to \$24,999
6	\$25,000 to \$34,999
7	\$35,000 to \$44,999
8	\$45,000 to \$54,999
9	\$55,000 to \$64,999
10	\$65,000 to \$74,999
11	\$75,000 to \$99,999

Figure 10: Household Income Distribution



The utility of the outside option has been normalized to zero for identification purposes. The model contains an individual-specific intercept for each of the choice alternatives, other than the outside option. With three random parameters (intercept, price, income) per each of the six choice alternatives, our model contains 18 parameters whose joint distribution we seek to estimate nonparametrically.

9.3 Implementation

In the implementation, we have run both HSMC and SMC for one hour of wallclock time. The implementation was run with a Coarray Fortran 2008 code using Intel 2016 compiler on 4 nodes of a 40-core 2.4 GHz Linux cluster (Loken et al., 2010; Ponce et al., 2019). We used 20 steps in constructing the Hamiltonian proposal in HSMC and tuned the step size to achieve transition acceptance rate of about 80%. Theoretical analysis of optimal step sizes and acceptance rates for HMC is provided in Beskos et al. (2010). We introduced the data in ten batches of equal size, one per 100 iterations. We tuned the SMC step size to achieve transition acceptance rates of about 30% (Roberts et al., 1997). Due to the Hamiltonian transition dynamics, HSMC takes somewhat longer than SMC to complete one full iteration but features superior mixing properties. During the run, HSMC completed about 8,800 iterations while SMC completed 10,650 iterations. Our model is nonparametric and in this context each particle represents a mixture of kernels whose count fluctuates during the implementation run. HSMC sampled on average 21 kernels per each particle mixture while SMC sampled on average 24 kernels. We used 3,200 particle mixtures and thus each procedure utilized on average over 67,000 parameter vectors, each of which had 18 dimensions.

9.4 Estimated Distribution of Coefficients

Here we also report the output on estimated distribution of the mixed multinomial logit coefficients. Table 3 provides summary statistics, mean and standard deviation of a benchmark parametric model Mlogit, SMC, and HSMC. The parametric model Mlogit was implemented by the command *mlogit*[10] in R (Croissant and Réunion, 2012).

Table 3: Summary of Estimated Coefficients

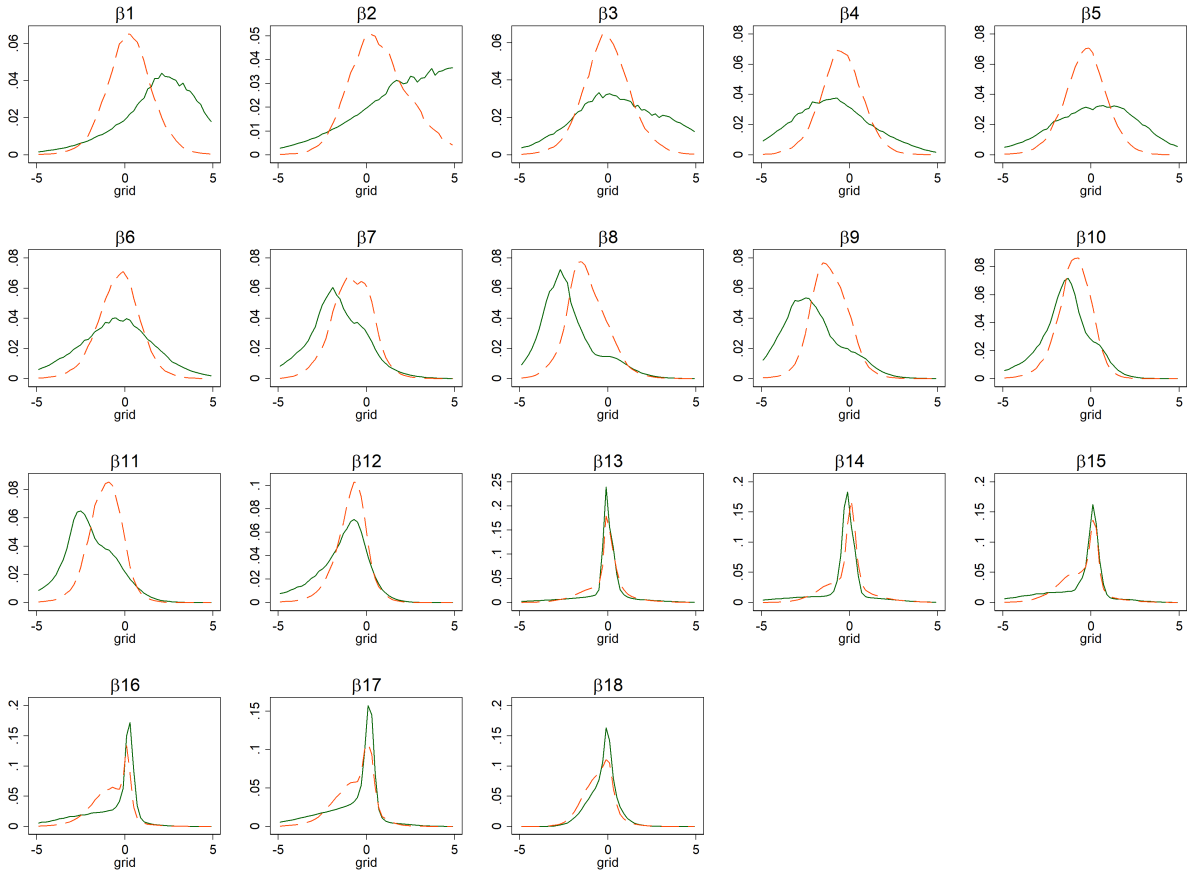
		Mlogit		SMC		HSMC	
Coefficient	Alternative	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
Intercept	1	2.428	0.121	0.000	0.024	0.0275	0.046
	2	2.817	0.109	0.005	0.026	0.0310	0.052
	3	1.201	0.142	-0.001	0.024	0.0107	0.036
	4	0.985	0.162	-0.006	0.025	-0.0118	0.035
	5	1.248	0.163	-0.005	0.025	0.0013	0.030
	6	3.516	0.216	-0.004	0.024	-0.0053	0.030
Price	1	-1.407	0.035	-0.011	0.028	-0.0288	0.048
	2	-1.407	0.035	-0.014	0.035	-0.0394	0.066
	3	-1.407	0.035	-0.014	0.032	-0.0355	0.053
	4	-1.407	0.035	-0.013	0.032	-0.0265	0.044
	5	-1.407	0.035	-0.015	0.033	-0.0321	0.050
	6	-1.407	0.035	-0.013	0.029	-0.0221	0.032
Income	1	-0.029	0.011	-0.002	0.016	-0.0004	0.015
	2	-0.088	0.011	-0.002	0.018	-0.0030	0.016
	3	0.015	0.025	-0.008	0.023	-0.0118	0.024
	4	0.042	0.026	-0.011	0.024	-0.0127	0.024
	5	-0.009	0.028	-0.010	0.027	-0.0128	0.023
	6	-0.018	0.020	-0.010	0.025	-0.0064	0.020

However, the summary statistics obscure important information about the shape of the estimated densities of the coefficients, which is undetectable in the parametric model. The estimated coefficient densities are presented in Figure 11. The probability mass of the HSMC densities are generally somewhat farther away from zero (prior mean) than SMC densities, suggesting that the former has explored the parameter space and updated the posterior more effectively than the latter. The overall pattern of the densities reveals that the income coefficients, $\beta_{i,13} - \beta_{i,18}$, are much closer to zero than the intercept or price coefficients. Nonetheless, several income coefficient densities feature a prominent left tail, suggesting a negative income effect. The intercept coefficients, $\beta_{i,1} - \beta_{i,6}$, tend to have more probability mass distributed on the positive side of the

[10]mlogit does not take into account the time dimension. With our dataset, mlogit failed to converge for alternative-specific price coefficients and hence we were only able to obtain output for a common price parameter across all choice alternatives.

real line, while the price coefficients, $\beta_{i,7} - \beta_{i,12}$, on the negative side. This pattern is in general agreement with the parametric benchmark Mlogit model estimates obtained in R.

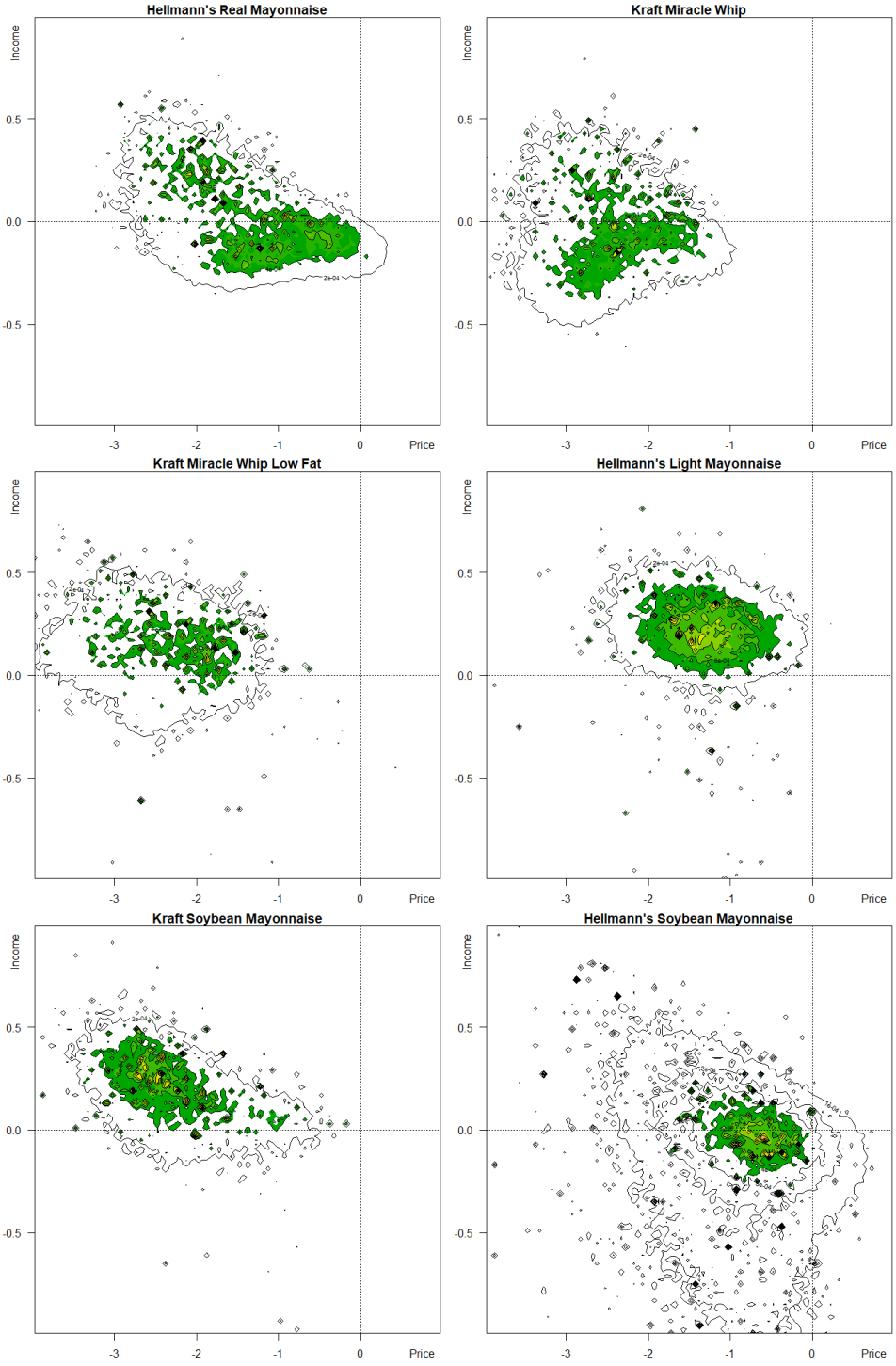
Figure 11: Estimated coefficient densities: HSMC (solid line) and SMC (dashed line)



Further insights into the price vs income effects on consumers can be obtained from bivariate posterior plots presented in Figure 12. As these are intended for economic analysis rather than HSMC vs SMC comparison we only provide output of HSMC.

Both most frequently purchased types of mayonnaise, Hellmann's Real Mayonnaise and Kraft Miracle Whip, feature a well pronounced modal region with both negative price and income effect that are positively correlated. For these two products lower prices and incomes thus imply higher purchase frequency. On the contrary, the low fat versions, Kraft Miracle Whip Low Fat and Hellmann's Light Mayonnaise, exhibit a dominant positive income effect that appears negatively correlated with the price effect. Individuals with

Figure 12: Estimated bivariate coefficient densities



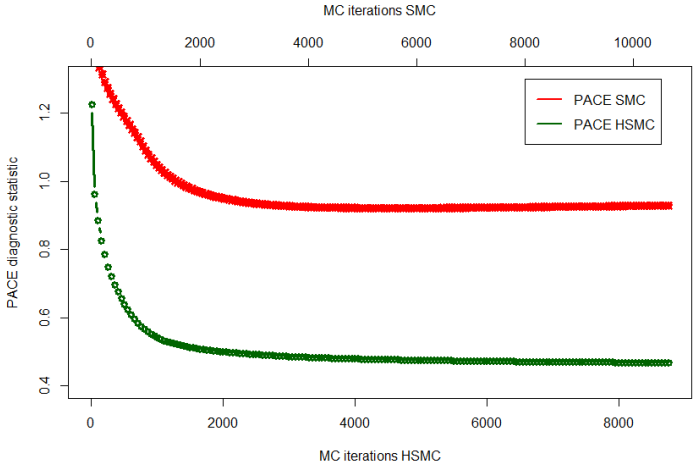
higher incomes are more likely to purchase these products, but income increases are traded off with lower prices. The soybean product varieties, Kraft Soybean Mayonnaise and Hellmann's Soybean Mayonnaise, can be considered niche products as each only accounts for less than 5% of mayonnaise purchases in our sample. While the former exhibits a long tail of a positive income effect for the latter this effect is insignificant.

An interesting feature of Hellmann's vs Kraft brand is that for all Hellmann's product categories the price effect is closer to zero than for all Kraft product categories. Consumers who would typically purchase a Kraft mayonnaise are thus more price sensitive than customers buying Hellmann's. This makes intuitive sense, since Hellmann's products are on average more expensive than Kraft products in our sample, as indicated by Figure 9, and as such are presumably more likely purchased by consumers with low price sensitivity and perhaps stronger brand preference.

9.5 PACE Convergence Diagnostics

In this application we estimate the density of β_i, which is 18 dimensional. Obtaining PACE in such relatively high-dimensional space turned out computationally prohibitive. We have implemented PACE for the bivariate distributions of all pairwise combinations of elements in β_i which reflect at least to some extent the information contained in the joint distribution of β_i beyond the univariate margins. We then calculated the average PACE statistic as a function of Monte Carlo iterations. The results are presented in Figure 13. Both methods seem to have converged well within the first half of the run. Throughout the run PACE of HSMC has dominated SMC by a substantial margin, attesting to the superior mixing properties of HSMC.

Figure 13: Application PACE



10 Conclusions

In this paper, we have proposed Hamiltonian Sequential Monte Carlo (HSMC), which uses Hamiltonian transition dynamics in particle mutation phase, in place of random walk transitions used in Sequential Monte Carlo (SMC), in the context of a Bayesian nonparametric mixture model. HSMC combines the advantages of SMC in terms of convenience of approximation of complex posterior shapes and parallelizability with the benefits of superior convergence properties stemming from Hamiltonian transition dynamics utilizing information about the first derivative of the likelihood function. We have contrasted the behavior of SMC and HSMC in a challenging simulation study, and showed favorable performance of HSMC. We have applied SMC and HSMC to a panel discrete choice model with a nonparametric distribution of unobserved individual heterogeneity, using the IRI panel data set.

References

Akhmatskaya, E., N. Bou-Rabee, and S. Reich (2009). A comparison of Generalized Hybrid Monte Carlo methods with and without momentum flip. *Journal of Computational Physics 228*(6), 2256–2265.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 1*, 1152–1174.

Bernardo, J. M. and A. F. M. Smith (1994). *Bayesian Theory*. New York: Wiley.

Beskos, A., N. S. Pillai, G. O. Roberts, J. M. Sanz-Serna, and A. M. Stuart (2010). The acceptance probability of the Hybrid Monte Carlo method in high-dimensional problems. *AIP Conference Proceedings 1281*(1), 23 – 27.

Blevins, J. R. (2016). Sequential monte carlo methods for estimating dynamic microeconomic models. *Journal of Applied Econometrics 31*(5), 773–804.

Bouchard-Côté, A., A. Doucet, and A. Roth (2017, January). Particle gibbs split-merge sampling for bayesian inference in mixture models. *J. Mach. Learn. Res. 18*(1), 868–906.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela (2008). Database paper: The iri marketing data set. *Marketing Science 27*(4), 745–748.

Burda, M., M. C. Harding, and J. A. Hausman (2008). A bayesian mixed logit-probit model for multinomial choice. *Journal of Econometrics 147*(2), 232–246.

Carvalho, C. M., H. F. Lopes, N. G. Polson, and M. A. Taddy (2010, 12). Particle learning for general mixtures. *Bayesian Analysis 5*(4), 709–740.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician 49*(4), 327–335.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika 89*(3), 539–552.

Chopin, N. (2004, 12). Central limit theorem for sequential Monte Carlo methods and its application to bayesian inference. *The Annals of Statistics 32*(6), 2385–2411.

Cowles, M. and B. Carlin (1996). Markov Chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association 91*, 883–904.

Creal, D. (2012). A survey of Sequential Monte Carlo methods for economics and finance. *Econometric Reviews 31*(3), 245–296.

Croissant, Y. and U. D. L. Réunion (2012). Estimation of multinomial logit models in r: The mlogit packages.

Doucet, A., A. Smith, N. de Freitas, and N. Gordon (2001). *Sequential Monte Carlo Methods in Practice*. Information Science and Statistics. Springer New York.

Duane, S., A. Kennedy, B. Pendleton, and D. Roweth (1987). Hybrid Monte Carlo. *Physics Letters B 195*(2), 216–222.

Durham, G. and J. Geweke (2014). Adaptive sequential posterior simulators for massively parallel computing environments. *Advances in Econometrics 34*, 1–44.

Fearnhead, P. (2004, Jan). Particle filters for mixture models with an unknown number of components. *Statistics and Computing 14*(1), 11–21.

Fergusson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*, 209–230.

Fernández-Villaverde, J. and J. F. Rubio-Ramírez (2007). Estimating macroeconomic models: A likelihood approach. *The Review of Economic Studies 74*(4), 1059–1087.

Fox, J., K. i. Kim, and C. Yang (2016). A simple nonparametric approach to estimating the distribution of random coefficients in structural models. *Journal of Econometrics 195*(2), 236–254.

Fox, J. T. and A. Gandhi (2016). Nonparametric identification and estimation of random coefficients in multinomial choice models. *The RAND Journal of Economics 47*(1), 118–139.

Fox, J. T., K. Kim, S. Ryan, and P. Bajari (2012). The random coefficients logit model is identified. *Journal of Econometrics 166*(2), 204–212.

Ghosal, S. and A. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Görür, D. and C. Rasmussen (2010, July). Dirichlet process gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology 25*(4), 653–664.

Grahn, A. (2021). The animate package.

Griffin, J. E. (2017, 11). Sequential Monte Carlo methods for mixtures with normalized random measures with independent increments priors. *Statistics and Computing 27*(1), 131–145.

Gupta, R., G. Kilcup, and S. Sharpe (1988). Tuning the Hybrid Monte Carlo algorithm. *Physical Review D 38*(4), 1278–1287.

Herbst, E. and F. Schorfheide (2014). Sequential monte carlo sampling for DSGE models. *Journal of Applied Econometrics 29*(7), 1073–1098.

Herbst, E. P. and F. Schorfheide (2016). *Bayesian Estimation of DSGE Models*. Princeton University Press.

Ishwaran, H. (1999). Applications of Hybrid Monte Carlo to generalized linear models: Quasicomplete separation and neural networks. *Journal of Computational and Graphical Statistics 8*, 779–799.

Keane, M. and N. Wasi (2013). Comparing alternative models of heterogeneity in consumer choice behavior. *Journal of Applied Econometrics 28*(6), 1018–1045.

Kim, J. G., U. Menzefricke, and F. M. Feinberg (2004). Assessing heterogeneity in discrete choice models using a dirichlet process prior. *Review of Marketing Science 2*(1), 1–39.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: Likelihood inference and comparison with arch models. *The Review of Economic Studies 65*(3), 361–393.

Leimkuhler, B. and S. Reich (2004). *Simulating Hamiltonian Dynamics*. Cambridge University Press.

Li, Y. and A. Ansari (2014). A bayesian semiparametric approach for endogeneity and heterogeneity in choice models. *Management Science 60*(5), 1161–1179.

Liu, J. S. (2004). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics.

Loken, C., D. Gruner, L. Groer, R. Peltier, N. Bunn, M. Craig, T. Henriques, J. Dempsey, C.-H. Yu, J. Chen, L. J. Dursi, J. Chong, S. Northrup, J. Pinto, N. Knecht, and R. V. Zon (2010). Scinet: Lessons learned from building a power-efficient top-20 system and data centre. *Journal of Physics: Conference Series 256*(1), 012026.

Lopes, H. F. and C. M. Carvalho (2013). Online Bayesian learning in dynamic models: An illustrative introduction to particle methods. In *Bayesian Theory and Applications*. Oxford University Press.

McFadden, D. L. and K. Train (2000). Mixed MNL models for discrete response. *Journal of Applied Econometrics 15*(5), 447–270.

Neal, R. M. (1993). Probabilistic inference using Markov Chain Monte Carlo methods. Technical report crg-tr-93-1, Dept. of Computer Science, University of Toronto.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman & Hall / CRC Press.

Orbanz, P. and Y. W. Teh (2010). Bayesian nonparametric models. In *Encyclopedia of Machine Learning*. Springer.

Plummer, M., N. Best, K. Cowles, and K. Vines (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News 6*(1), 7–11.

Ponce, M., R. van Zon, S. Northrup, D. Gruner, J. Chen, F. Ertinaz, A. Fedoseev, L. Groer, F. Mao, B. C. Mundim, M. Nolta, J. Pinto, M. Saldarriaga, V. Slavnic, E. Spence, C.-H. Yu, and W. R. Peltier (2019). Deploying a top-100 supercomputer for large parallel workloads: The niagara supercomputer. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*, PEARC '19, New York, NY, USA. Association for Computing Machinery.

Rasmussen, C. E. (2003). Gaussian processes to speed up Hybrid Monte Carlo for expensive Bayesian integrals. *Bayesian Statistics 7*, 651–659.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). New York: Springer.

Roberts, G. O., A. Gelman, and W. R. Gilks (1997, 02). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab. 7*(1), 110–120.

Rosenbrock, H. H. (1960, 01). An Automatic Method for Finding the Greatest or Least Value of a Function. *The Computer Journal 3*(3), 175–184.

Rupp, K. (2018). 42 years of microprocessor trend data. `https://www.karlrupp.net/2018/02/42-years-of-microprocessor-trend-data/`. Accessed: 2018-06-13.

Thomadsen, R. (2016). The impact of switching stores on state dependence in brand choice. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2759868`. Working Paper, Olin Business School, Washington University in St. Louis.

Train, K. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press.

Tuckerman, M., B. Berne, G. Martyna, and M. Klein (1993). Efficient molecular dynamics and Hybrid Monte Carlo algorithms for path integrals. *The Journal of Chemical Physics 99*(4), 2796–2808.

Ulker, Y., B. Günsel, and T. Cemgil (2010, 13–15 May). Sequential monte carlo samplers for dirichlet process mixtures. In Y. W. Teh and M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Volume 9 of *Proceedings of Machine Learning Research*, Chia Laguna Resort, Sardinia, Italy, pp. 876–883. PMLR.

VanDerwerken, D. and S. C. Schmidler (2017). Monitoring joint convergence of MCMC samplers. *Journal of Computational and Graphical Statistics 26*(3), 558–568.

Xie, Y. (2013). animation: An R package for creating animations and demonstrating statistical methods. *Journal of Statistical Software 53*(1), 1–27.