Inference with Hamiltonian Sequential Monte Carlo Simulators

By Remi Daviet*

Draft: September 12, 2016

The paper proposes a new Monte-Carlo simulator that is robust to multimodality and complex shapes to use for inference in presence of difficult likelihoods or target functions.
JEL: C15

I. Introduction

Econometrics has traditionally relied on methods based on the optimization of a target function to perform inference. Such methods include the widely spread Maximum Likelihood Estimators (MLE), Generalized Method of Moments (GMM) or the Least Squares Estimators (LSE). These methods are especially effective when the target function is unimodal and is shown to converge rapidly to a point.

Recent development in economic modeling have lead to the emergence of more complicated target functions where multimodality, complex shapes and slow convergence make traditional inference approaches ineffective. Numerical maximization methods such as Nelder-Mead, Newton-Raphson or Expectation Maximization do not guarantee the convergence toward a global maximum [Gourieroux and Monfort, 1995]. In addition, these methods provide a point-estimate and do not give information about the shape of the target function.

* Daviet: University of Toronto, remi.daviet@mail.utoronto.ca

These more complicated target functions are found in almost every subfield of economics: Multimodal likelihoods for instance can be the result of using DSGE models in Macroeconomics [Herbst and Schorfheide, 2014], GARCH models in Finance [Doornik et al., 2000], BLP models in industrial organization [Bajari, 2003], Spatial linear models in Urban Economics [Mardia and Watkins, 1989]. In fact, with many non-linear models, likelihood functions are non-smooth and multimodal [Koop and Potter, 1999]. A similar observation can be made for models with structural breaks and outliers Koop and Potter [2000]. Various attempts have been made to mitigate the problem, generally relying on Bayesians or Quasi-Bayesian Methods and Markov-Chain Monte Carlo (MCMC) or Sequential Monte Carlo (SMC) simulations.

MCMC simulation has the advantage of being able to recover more information about a target function than optimization algorithms by providing a distribution of points matching the shape of the target function acting as a measure [Chernozhukov and Hong, 2003]. The only restriction is that the target function has to be positive. In the case of a non-positive target function, the exponential of the function can be taken. This exponentiated function can be treated as a quasi-likelihood. The draws from the simulation provide a Monte Carlo approximation of the distribution of interest, facilitating inference. Application of this method to economics can be found in several papers across all subfields. For instance, Herbst has proposed several Monte Carlo methods to solve the estimation problems with Macroeconomic DSGE models [Herbst and Schorfheide, 2014]. We find similar approaches in industrial organization for demand estimation models [Jiang et al., 2009], or in time series analysis Burda [2015].

The key problem with traditional MCMC simulators using Metropolis-Hastings or Gibbs Sampling methods is that they do not behave necessarily well under multimodality, concentrated mass or complex shapes. The Markov Chain used to simulate the distribution can get trapped in one of the modes if it is not close enough to the other modes and there is no mass between them. When facing

complex concentrated shapes, these simulators are not performing well exploring the space of interest due to a high level of rejections when trying to move away from the current point in the chain. Some specific simulators have shown interesting properties concerning these problems. The family of SMC simulators solves the problem of multimodality by not having a unique Markov Chain exploring the parameter space [Durham and Geweke, 2013]. The sequential approach of the SMC also partly solves the problem of concentrated mass by allowing the target function to progressively converge toward its final form [Chopin, 2002]. The MCMC simulators using Hamiltonian dynamics, sometimes refereed to as Hamiltonian Monte Carlo (HMC) simulators, have shown to be very effective in exploring the parameter space when the target function has a complex elongated shape or isolated concentrated mass.

Our contribution is to provide a method that is robust to multimodality and complex shapes by combining the advantages of the SMC and the Hamiltonian dynamics. Moreover, while most econometrics paper describing MCMC or SMC methods are put in a Bayesian framework, we keep the description in a general framework. Using our method, empirical economists should be able to perform inference without having to worry about the shape or complexity of their target function.

The advantages of recovering the full shape of the target function are multiple. Multimodality can easily be identified, concentration can be measured and counter-factual checking or prediction can be done by integration over the parameter space. Moreover, the Monte Carlo approach allows for easy variable transformation without the need of derivation of a complicated Jacobian.

Our method can also be used for maximization of complicated functions by simulated annealing: the target quasi-likelihood is of the form $\exp(f(x))^{\gamma}$. The quasi-likelihood will become concentrated on the maximizers of $f(x)$ as $\gamma \to \infty$ [Hwang, 1980]. The use of SMC methods for optimization via simulated annealing has already been detailed and proven effective [Zhou and Chen, 2013].

We will first describe the method and its properties and then present toy examples as applications.

II. Method

A. *Sequence of distributions*

Our simulator belongs to the general class of Sequential Monte Carlo simulators. As such, it possesses the main characteristics and statistical properties of every SMC [Del Moral et al., 2006]. We want a sample $\{\theta_n\}_{n=1}^N$ from a sequence of distributions with densities $f_1(\theta_n), \ldots, f_T(\theta_n)$. We also require an initial distribution with density $f_0(\theta_n)$ that is easy to sample from. The HSMC simluator provides us for each step $t = 0, \cdots, T$ with a set of values $\{\theta_n\}_{n=1}^N$ that approximately follow the distribution $f_t(\theta_n)$. The $\{\theta_n\}_{n=1}^N$ are often called particles. As most SMC methods, the simulator we propose works best when target densities are smoother at the beginning and progressively converging toward the final sharper target density.

A sequence of distributions can be found in many applications relevant to economics. In frequentist econometrics, the sequence can be the likelihood or quasi-likelihood for data collected until time t, e.g. $f_t(\theta_n) \propto L(\theta_n; y_1, \ldots, y_t)$. The quasi-likelihood can be built from any estimator maximizing or minimizing a target function such as the Generalized Method of Moments or a Least Squarred Error estimator Chernozhukov and Hong [2003]. The counterpart in the Bayesian framework would be the posterior distribution of the parameter θ given the data until time T. Comparatively to a standard MCMC approach which requires an evaluation of the target function with every observation at each step, the SMC approach is less computationally intensive. Moreover, adding the observations one or a few at a time creates a desirable tempering effect [Chopin, 2002]. This approach is particularly efficient in large datasets where new observations come regularly as updating the estimator can be done in one step.

A second application is kernel density estimation when observations are added

a few at a time and bandwidth is progressively shrunk. An application of this approach can be found later in this paper.

Another application proposed by Neal [2001] shows the benefits of moving progressively from a tractable distribution $f_1(\theta_n)$ to a target distribution $f(\theta_n)$ by geometrically reweighting them : $f_t(\theta_n) \propto f(\theta_n)^{\phi_t} f_1(\theta_n)^{1-\phi_t}$ with $0 \leq \phi_1 < \ldots < \phi_T = 1$.

Finally, our simulator can be used for maximization using a simulated annealing approach. Our sequence of distributions will then become $f_t(\theta_n) \propto f(\theta_n)^{\gamma_t}$ with γ_t increasing to high values as t increases.

B. Algorithm

1) Initialization: Draw N particles $\{\theta_n^{(0)}\}_{n=1}^N$ from $f_0(\theta_n)$

2) Repeat for $t = 1, \ldots, T$

   a) Correction: assign weight $w_n^{(t)} = f_t(\theta_n)/f_{t-1}(\theta_n)$ to each of the particles $\{\theta_n^{(t-1)}\}_{n=1}^N$

   b) Selection: draw N new particles $\{\hat{\theta}_n^{(t)}\}_{n=1}^N$ with replacement from the current sample of particles using weights $w_n^{(t)}$. Give the new particles a weight of 1.

   c) Mutation: For each particle, perform a Hamiltonian step as described in section II.C to obtain a new sample of particles $\{\theta_n^{(t)}\}_{n=1}^N$.

In the initialization phase we obtain a sample of particles distributed according to $f_0(\theta_n)$ distribution. For the HSMC method to perform well we need a distribution that covers well the whole Θ space and has mass where the other distributions $f_t(\theta_n)$ in the sequence have mass too.

In the loop, before the correction phase, we have particles all weighted to 1 that provides a Monte-Carlo simulation of the distribution $f_{t-1}(\theta_n)$. In the correction phase, we reweight them to obtain an approximation of the distribution $f_t(\theta_n)$ by importance sampling.

In the selection phase, we perform sampling importance resampling to obtain an approximation of the distribution $f_t(\theta_n)$ using particles $\{\hat{\theta}_n^{(t)}\}_{n=1}^N$ with equal weights. Note that at the end of the selection phase, we can expect to have several particles sharing the same value.

Finally, in the mutation phase, we explore the Θ space by moving the particles using a Hamiltonian Monte Carlo (HMC) approach. Since $f_t(\theta_n)$ is the stationary distribution of our HMC, the particles both before and after the HMC step should approximate $f_t(\theta_n)$. However, in the case $\{\hat{\theta}_n^{(t)}\}_{n=1}^N$ do not follow exactly the distribution $f_t(\theta_n)$, performing a HMC step should improve the approximation by the convergence properties of HMC [Neal, 2011].

The main contribution of the paper lies in the use of Hamiltonian dynamics for the mutation phase in a SMC method. In the literature, SMC algorithms are generally found to use a standard Metropolis-Hastings step in their mutation phase. Conversely, HMC methods using multiple particles do not have the resampling phases 2a/2b.

C. Hamiltonian Monte Carlo step

The purpose of the Hamiltonian Monte Carlo method is to formulate a Markov chain on the parameter space for which, under certain conditions, $f_t(\theta_n)$ is the stationary distribution. It relies on Hamiltonian dynamics to move a particle $\hat{\theta}_n$ to a new point θ_n^* in the Θ space. This new point, called proposal will then be accepted or rejected as the new value for θ_n following a Metropolis-Hastings rejection method. The movement in the Θ space can be constrained and we then refer to the method as a Constrained Hamiltonian Monte Carlo (CHMC).

To describe the HMC step, we need to define the function $U_t(\theta_n)$ with gradient denoted $\nabla U_t(\theta_n)$:

$$U_t(\theta_n) = -\log(f_t(\theta_n))$$

We also need to define an auxiliary vector p of dimension $\dim(\theta)$ that will represent the momentum of the particle when moving following Hamiltonian dynamics.

Intuitively, the particle will move on a surface where the altitude is represented by $U_t(\theta_n)$. When the particle goes up a slope, it will slow down or even turn back. This will keep the particle where the target function $f_t(\theta_n)$ has mass. The continuous movement is approximated by a series of L steps of size ϵ. At the end of the last step, the momentum is reversed to make the proposal symmetric. If we start at the final position with the final reversed momentum, we will find the particle to go back to the original position after L steps. This ensures reversibility and facilitate the computation of the Metropolis-Hastings acceptance probability.

The HMC step proceeds as follow:

1) define the starting position of the proposal $\theta_n^* = \hat{\theta}_n$
2) draw an initial momentum vector p from a multivariate Gaussian distribution: $p \sim \mathcal{N}(0, M)$
3) update the momentum vector by half a step taking the gradient into account: $p^* = p - \frac{\epsilon}{2} \cdot \nabla U_t(\hat{\theta}_n)$
4) Repeat for $l = 1, \ldots, L$
    a) update the position by a full step: $\theta_n^* = \theta_n^* + \epsilon \cdot p^*$
    b) update the momentum by a full step, except at the end of the trajectory: if $(l \neq L)$, then $p^* = p^* - \epsilon \cdot \nabla U_t(\hat{\theta}_n)$
5) update the momentum vector by half a step: $p^* = p^* - \frac{\epsilon}{2} \cdot \nabla U_t(\hat{\theta}_n)$
6) negate the momentum vector: $p^* = -p^*$
7) compute the Metropolis-Hastings acceptance probability: $a = \min\left(1, \exp\left[U_t(\hat{\theta}_n) - U_t(\theta_n^*) + \frac{\sum p^2}{2} - \frac{\sum p^{*2}}{2}\right]\right)$
8) set $\theta_n = \theta_n^*$ with probability a, and $\theta_n = \hat{\theta}_n$ otherwise

The HMC mutation step needs to be tuned by choosing appropriately the quantities (M, L, ϵ). To learn more about how to choose these quantities, consult the review paper by Neal [2011].

When we want to put constraints on some of the dimensions of θ_n, we can modify the HSMC method to have the particles to bounce on the constraints as if they were walls. Most of the time, the constraints are of the form $\theta_{dn} \leq u_d$ or $\theta_{dn} \geq l_d$ for some dimensions d of θ_n. The position updating step 4a is then replaced by:

1) for each dimension d of θ_n^*:

   a) update position in dimension d: $\theta_{dn}^* = \theta_{dn}^* + \epsilon \cdot p_d^*$

   b) if θ_{dn}^* is constrained, repeat the following until θ_{dn}^* satisfies all constraints:

      i) if $(\theta_{dn}^* > u_d)$, then $\theta_{dn}^* = u_d - (\theta_{dn}^* - u_d)$ and $p_d^* = -p_d^*$

      ii) if $(\theta_{dn}^* < l_d)$, then $\theta_{dn}^* = l_d + (l_d - \theta_{dn}^*)$ and $p_d^* = -p_d^*$

With this modified approach, if the particle passes a constraint "wall" during a position update, the symmetric of the particle θ_n^* relative to the wall in the Θ space is taken, and the momentum in the constrained dimension is reversed. This approach simulates the particle bouncing on the wall and preserves reversibility. A similar approach can be used for more complex constraints of the form $G(\theta_n) \geq 0$.

D. Algorithm's properties

Our HSMC method fits into the SMC framework described in Chopin [2004] and his central limit theorem can be applied to our simulator. Provided that our Hamiltonian mutation step preserves the distribution $f_t(\theta_n)$, the following convergences hold almost surely as $N \to \infty$ for any measurable function ϕ such that the expectations below exists:

$$N^{-1} \sum_{n=1}^{N} \phi(\theta_n^{(t)}) \to \mathbb{E}_{f_t}[\phi(\theta_n)]$$

$$\frac{\sum_{n=1}^{N} w_n^{(t)} \phi(\theta_n^{(t-1)})}{\sum_{n=1}^{N} w_n^{(t)}} \to \mathbb{E}_{f_t}[\phi(\theta_n)]$$

$$N^{-1}\sum_{n=1}^{N}\phi(\hat{\theta}_n^{(t)}) \rightarrow \mathbb{E}_{f_t}[\phi(\theta_n)]$$

The proof that our Hamiltonian transition kernel satisfies the conditions to have $f_t(\theta_n)$ as a stationary distribution can be found in the review paper on Hamiltonian Monte Carlo by Neal [2011].

III. Method variations

While the Hamiltonian step is designed to be applied on continuous distributions, it is easy to extend our method for spaces with discrete dimensions. We can split our space Θ in two blocs $\{\Theta_c, \Theta_d\}$ where Θ_c includes the dimensions where Θ is continuous and Θ_d includes the dimensions where Θ is discrete. From this separation into blocs, a standard Metropolis within Gibbs [Gilks et al., 1995] step can be used with the Hamiltonian step being used in the continuous block.

Another variation of the method consists in running the algorithm in parallel for J groups of N particles. At the end, the sample properties of the J groups can be compared. If the properties differ too much from each other, we can suspect a convergence problem. This is the approach taken by Durham and Geweke [2013] in their adaptive SMC simulator.

Finally, the Hamiltonian step preserving the distribution of interest, multiple steps can be made at each mutation phase. This solution increase the performance of the method when particles are not exploring the Θ space fast enough.

IV. Examples

We used the HSMC method on kernel density estimates of two functions known to challenge classical optimizers and MCMC simulators. The first function is created for this paper and called the smiley function. It is a mixture of 3 Rosenbrock smile functions often found as an example to show the limitations of MCMC algorithms. The second function is a dropwave function. Both functions present multimodality and follow complex shapes. For ease of visualization, we kept the

functions bidimentional.

We chose to simulate Gaussian kernel density estimates as they can mimic the progressive convergence of several target function when data are added progressively. In our case, data are added by blocks of 100 points and the bandwidth of the kernel density estimate is $n^{-1/5}$. This rate has been chosen as it is the order of the bandwidth reduction rate when using optimal bandwidth for most distributions [Yatchew, 1998].

A. Smiley kernel density estimate

We generated a sample of 2048 data points with coordinates (x, y) using a density proportional to the following function:

$$\begin{aligned} g(x,y) = &\exp\left(\frac{1}{5}\left(-6\left(-(2.5-x)^2 - 1.5y + 38\right)^2 - (2.5-x)^2\right)\right) \\ &+ \exp\left(\frac{1}{5}\left(-6\left(-(x+2.5)^2 - 1.5y + 38\right)^2 - (x+2.5)^2\right)\right) \\ &+ \exp\left(\frac{1}{5}\left(-5\left(y - x^2\right)^2 - x^2\right)\right) \end{aligned}$$

The contour plot and 3D plot of the function used can be found in figure 1. We can easily see the multimodaly and the elongated shapes on the function plots. The data generated match the shape of the smiley function and have been represented in figure 2.

The target function we want to simulate is a kernel density estimate using these data points. We used 4 independent groups of 512 particles to simulate the kernel density for a total of 2048 particles. The data points are partitioned in 20 blocks of 100 points and 1 block of 48 points, for a total of 21 blocks. Consequently, the sequence will include $T = 21$ target functions with the addition of the initial density. The initial density $f_0(\theta_n)$ used is a bivariate normal distribution with parameters $\{\mu_1 = 0, \mu_2 = 10, \sigma_1 = 10, \sigma_2 = 20, \rho = 0\}$. The HMC tunning parameters (M, L, ϵ) used are $(I_2, 20, 0.05)$, where I_2 is the 2×2 identity matrix. The

FIGURE 1. CONTOUR PLOT AND 3D PLOT OF THE SMILEY FUNCTION

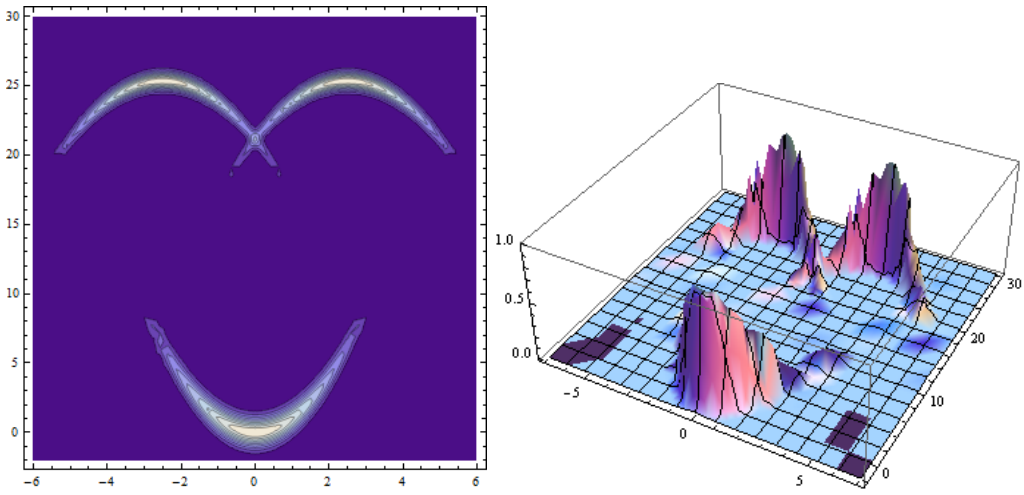
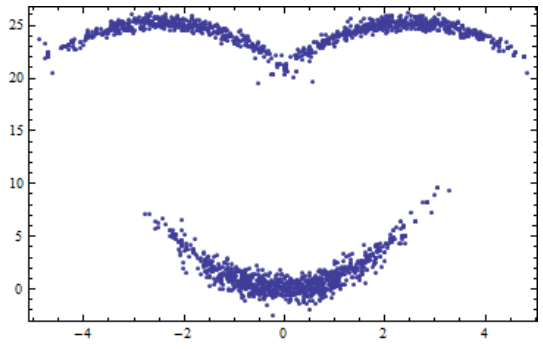


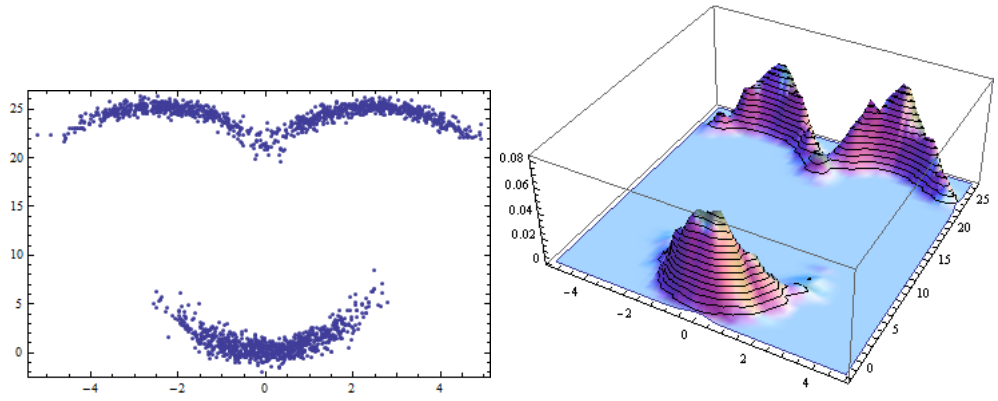
FIGURE 2. GENERATED DATA



results of the HSMC simulation as well as a smoothed histogram of the simulated points have been represented in figure 3. We can see that the HSMC method provided satisfying results and successfully converged with only 21 iterations. The lowest mutations acceptance rate we observed across several runs was 2043/2048.

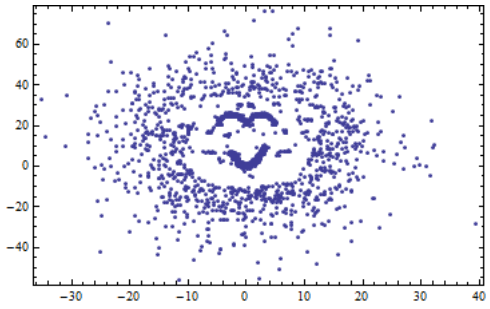
To illustrate the performance of our HSMC method, we also tried to simulate the same sequence of functions using a parallel HMC approach with 21 iterations on the same kernel density. To do so we performed the mutation phase 21 times on $f_2 1(\theta_n)$. The result represented in figure 4 show that several particles failed to

FIGURE 3. HSMC SIMULATION RESULTS



converge. Moreover, there is no guarantee that the particles around each mode are distributed according to the mass around these modes in the target function. Particles having a tendency to converge to the closest mode, it is possible to have a first mode with twice the mass of a second mode but only half of the particles around it.

FIGURE 4. PARALLEL HMC SIMULATION RESULTS



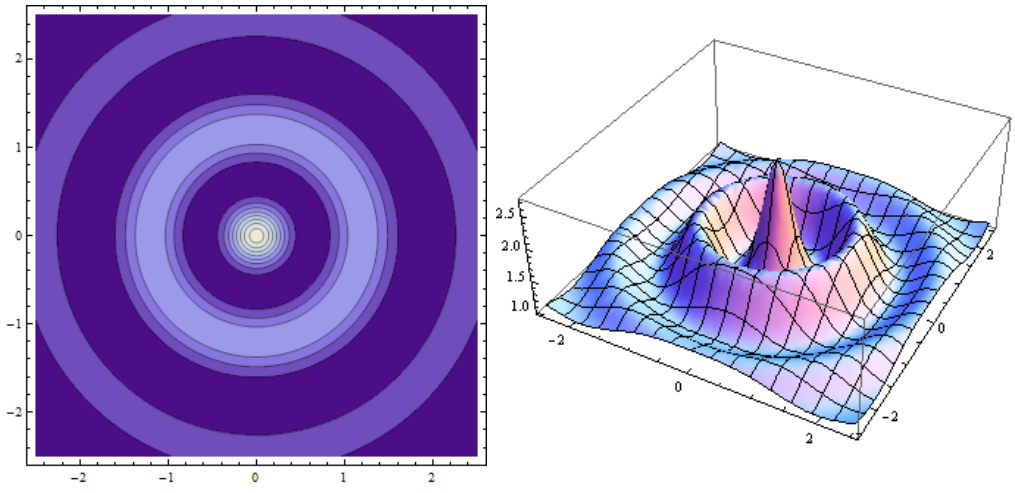
B. *Constrained dropwave kernel density estimate*

We generated a sample of 4096 data points with coordinates (x, y) using a density proportional to the following function defined on $[-2.5; 2.5] \times [-2.5; 2.5]$:

$$g(x, y) = \exp\left(\frac{\cos\left(5\sqrt{x^2 + y^2}\right) + 1}{x^2 + y^2 + 2}\right)$$

The contour plot and 3D plot of the function used can be found in figure 5.

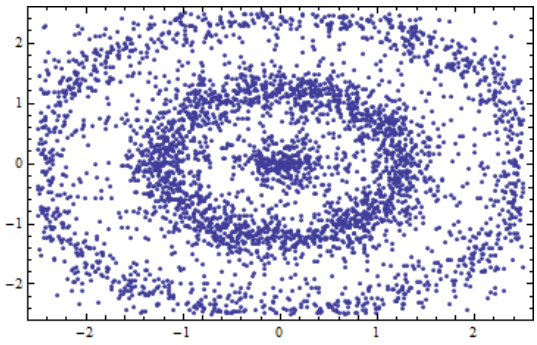
FIGURE 5. CONTOUR PLOT AND 3D PLOT OF THE DROPWAVE FUNCTION



The data generated match the shape of the dropwave function and have been represented in figure 6.

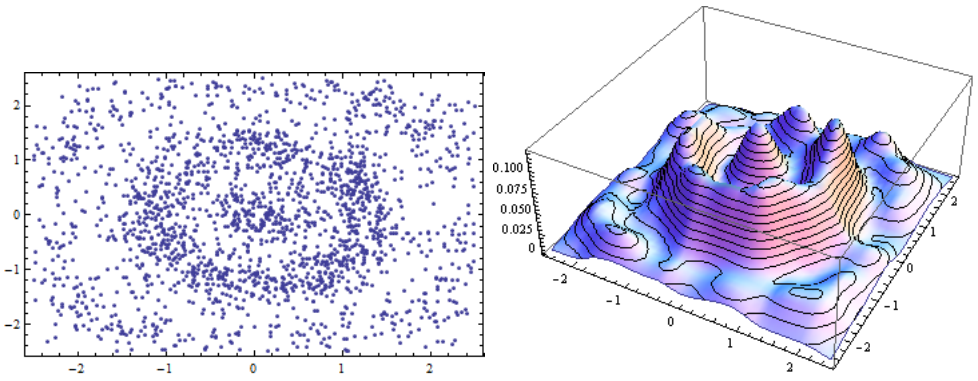
The target function we want to simulate is a constrained kernel density estimate using these data points. The Gaussian kernel is defined on $\mathbb{R}^2$ but we want to limit the domain to $[-2.5; 2.5] \times [-2.5; 2.5]$. To do so we keep the Gaussian kernel as is but use the constrained HMC method during our mutation phase. We used 4 independent groups of 512 particles to simulate the kernel density for a total of 2048 particles. The data points are partitioned in 40 blocks of 100 points and 1 block of 96 points, for a total of 41 blocks. Consequently, the sequence

FIGURE 6. GENERATED DATA



will include $T = 41$ target functions with the addition of the initial density. The initial density $f_0(\theta_n)$ used is a bivariate normal distribution with parameters $\{\mu_1 = 0, \mu_2 = 0, \sigma_1 = 10, \sigma_2 = 10, \rho = 0\}$. The HMC tunning parameters (M, L, ϵ) used are $(I_2, 20, 0.05)$, where I_2 is the 2×2 identity matrix. The results of the HSMC simulation as well as a smoothed histogram of the simulated points have been represented in figure 7. We can see that the constrained HSMC method also provided satisfying results and successfully converged with 41 iterations. The lowest mutations acceptance rate we observed across several runs was 2023/2048.

FIGURE 7. CHSMC SIMULATION RESULTS



V. Future work and conclusion

The encouraging performance of the algorithm finds direct potential applications in empirical work where standard MCMC methods have shown limitations. One of this potential application is the use of optimal instruments for non-linear BLP models in industrial organization.

REFERENCES

P. Bajari. Discussion of allenby, chen and yang. *Quantitative Marketing and Economics*, 1(3):277–283, 2003.

M. Burda. Constrained hamiltonian monte carlo in bekk garch with targeting. *Journal of Time Series Econometrics*, 7(1):95–113, 2015.

V. Chernozhukov and H. Hong. An mcmc approach to classical estimation. *Journal of Econometrics*, 115(2):293–346, 2003.

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

N. Chopin. Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *Annals of statistics*, pages 2385–2411, 2004.

P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

J. A. Doornik, M. Ooms, et al. Multimodality and the garch likelihood. In *World Congress of the Econometric Society, Seattle, August*, 2000.

G. Durham and J. Geweke. Adaptive sequential posterior simulators for massively parallel computing environments. *Bayesian Model Comparison (Advances in Econometrics); Jeliazkov, I., Poirier, DJ, Eds*, pages 1–44, 2013.

W. R. Gilks, N. Best, and K. Tan. Adaptive rejection metropolis sampling within gibbs sampling. *Applied Statistics*, pages 455–472, 1995.

C. Gourieroux and A. Monfort. *Statistics and econometric models*, volume 1. Cambridge University Press, 1995.

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

C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182, 1980.

R. Jiang, P. Manchanda, and P. E. Rossi. Bayesian analysis of random coefficient logit models using aggregate data. *Journal of Econometrics*, 149(2):136–148, 2009.

G. Koop and S. M. Potter. Bayes factors and nonlinearity: evidence from economic time series. *Journal of Econometrics*, 88(2):251–281, 1999.

G. Koop and S. M. Potter. Nonlinearity, structural breaks or outliers in economic time series. *Nonlinear Econometric Modeling in Time Series Analysis*, pages 61–78, 2000.

K. Mardia and A. Watkins. On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76(2):289–295, 1989.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2): 125–139, 2001.

R. M. Neal. Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

A. Yatchew. Nonparametric regression techniques in economics. *Journal of Economic Literature*, 36(2):669–721, 1998.

E. Zhou and X. Chen. Sequential monte carlo simulated annealing. *Journal of Global Optimization*, 55(1):101–124, 2013.