BAYESIAN INFERENCE FOR STOCHASTIC CUSP CATASTROPHE MODEL

Haipeng Gao

A dissertation submitted to the faculty of University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the Department of Statistics and Operations Research

Chapel Hill
2019

Approved by:

Nilay Tanik Argon

Ding-Geng Chen

Chuanshu Ji

Feng Shi

Serhan Ziya

ABSTRACT

HAIPENG GAO: Bayesian Inference for Stochastic Cusp Catastrophe Model
(Under the direction of Chuanshu Ji)

In modern financial econometrics, diffusion processes have been broadly used to model the stochastic behavior of economic variables such as stock prices, interest rates, and exchange rates. Well-known models such as Black-Scholes, Vasicek, and Cox-Ingersoll-Ross (CIR mdoel), all assume that the underlying state variables follow diffusion processes. If one believes that the observed time-series are generated according to some parametric specification, developing rigorous statistical methods to calibrate the underlying model to measured observations has become a considerable subject of the field.

The thesis considers cusp model, one of the elementary catastrophe models studied in catastrophe theory. The research problem of this thesis is to develop an accurate and computationally feasible parameter estimation algorithm based on Bayesian principle that can be implemented in absence of an exact transition distribution for cusp model using discretely sampled observations. The problem can be further specified as parameter estimations using complete observations and using partial observations. Accuracy and efficiency of the approach are demonstrated and examined in a series of simulation-based studies that consist of both trajectory simulations and parameter estimations. We extend the developed algorithm and apply it to Bayesian hierarchical modeling and cusp model with time-varying parameters.

To Grace

ACKNOWLEDGEMENTS

I'm deeply indebted to my advisor Prof. Chuanshu Ji who guided me through my doctorate level research. I'm extremely grateful for his motivation, patience, encouragement, and continuous support.

I'm grateful to Prof. Ding-Geng Chen for introducing the research topic, extended discussions and insightful suggestions which have contributed greatly to the improvement of the thesis. I must also thank Dr. Bill (Feng) Shi for his unwavering support and invaluable suggestions. I would also wish to extend my sincere thanks to Prof. Nilay Tanik Argon and Prof. Serhan Ziya for serving on my PhD committee. The thesis has also benefited from comments and suggestions made by Tony (Ruito) Fan who inspired me in many aspects.

I never take the opportunity to study for granted, and I believe not everyone is as lucky as I am who had the opportunity to explore the field that one is genuinely interested in. I appreciate Prof. Vidyadhar Kulkarni who admitted me to the PhD program for giving me the opportunity to purse doctorate study at UNC Chapel Hill.

The achievement would not be possible without the support from my family and the extended family. Especially, I would express my gratitude to Yan Sheng, my mother, for her love, support, and sacrifice, as well as to Kyungeun, my wife and my best friend, for believing in me at the very beginning as well as her continuous support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

MLE	Maximum likelihood estimation
PDF	Probability density function
SDE	Stochastic differential equation
$\phi(z)$	PDF of the standard normal distribution $\mathcal{N}(0,1)$

CHAPTER 1
Introduction

In modern financial econometrics, diffusion processes have been broadly applied to model the stochastic behavior of economic variables such as stock prices, interest rates, and foreign exchange rates. Well-known models such as Black-Scholes, Vasicek, and Cox-Ingersoll-Ross (CIR mdoel), all assume the underlying state variables follow diffusion processes. The thesis considers stochastic cusp model, one of the elementary catastrophe models studied in catastrophe theory.

In this chapter, we give a brief overview of cusp model including its development and applications in economics. While doing so, we highlight unique characteristics which make cusp model appealing and valuable in economics and financial econometrics.

1.1 Catastrophe theory

Catastrophe theory is commonly regarded as a branch of bifurcation theory in the study of dynamic systems in mathematics. A *bifurcation* occurs when a change, usually small and smooth, made to the system's parameter values engenders a sudden "qualitative" change in its behavior. Catastrophe theory studies the mathematical characteristics of bifurcation phenomena and reveals that such bifurcations tend to occur as part of well-defined geometrical structures [Ivancevic and Ivancevic, 2007].

To better illustrate, imagine we have an object inside a system expressed by the simplest cubic polynomial $f(x) = x^3$ residing at its equilibrium $x = 0$. When the system is perturbed vaguely by adding an extra linear term x, $x = 0$ will no longer be an equilibrium under the new but perturbed system $f(x) = x^3 + x$ (Figure 1.1). In fact, such small perturbation experienced by particular parameters of some non-linear system could cause equilibria to appear or to disappear, or to change from attracting to repelling, and vice versa, that eventually leads to sudden "qualitative" changes of the behavior of the system [Costantino et al., 2005].

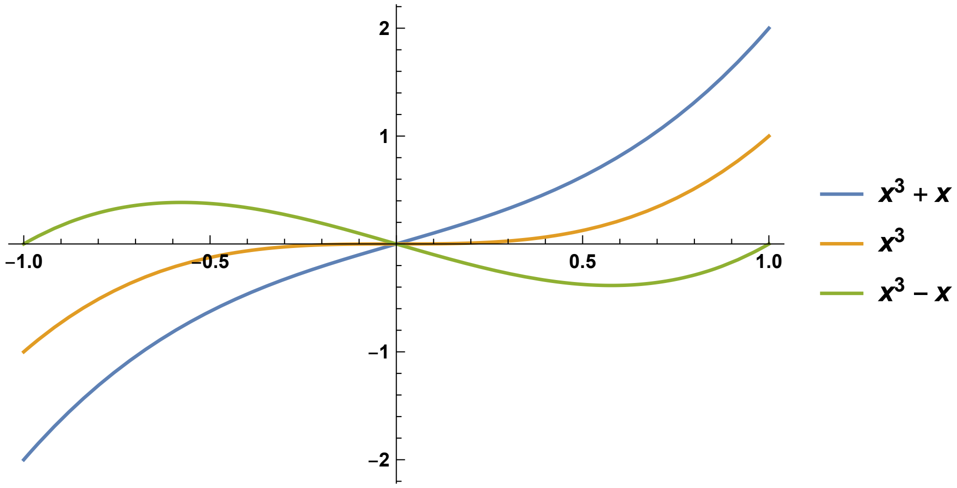


Figure 1.1: Example of degenerate singularity: function $f(x) = x^3$ at $x = 0$

1.1.1 Deterministic dynamics

The dynamics of a catastrophe model is often expressed in terms of a potential function $V(x)$, where $V(x)$ is commonly approximated by polynomials. The deterministic dynamics is then governed by the ordinary differential equation

$$\frac{dx}{dt} = -\frac{dV(x;\theta)}{dx}, \quad x \in \mathbb{R} \text{ and } \theta \in \mathbb{R}^p. \tag{1.1}$$

where x and θ denote location and system parameters respectively.

x_0 is an *equilibrium* point if $\frac{dV(x)}{dx}|_{x=x_0} = 0$. An equilibrium point x_0 is said to be unstable if $\frac{d^2V(x)}{dx^2}|_{x=x_0} < 0$, and x_0 is a local maximum; x_0 is called a *stable* equilibrium if $\frac{d^2V(x)}{dx^2}|_{x=x_0} > 0$, and in this case a local minimum. An object inside the system tends to move toward the point of lowest potential. Furthermore, an equilibrium point x_0 is *degenerate* if $\frac{d^2V(x)}{dx^2}|_{x=x_0} = 0$, and in this case x_0 is neither a local minimum nor a local maximum.

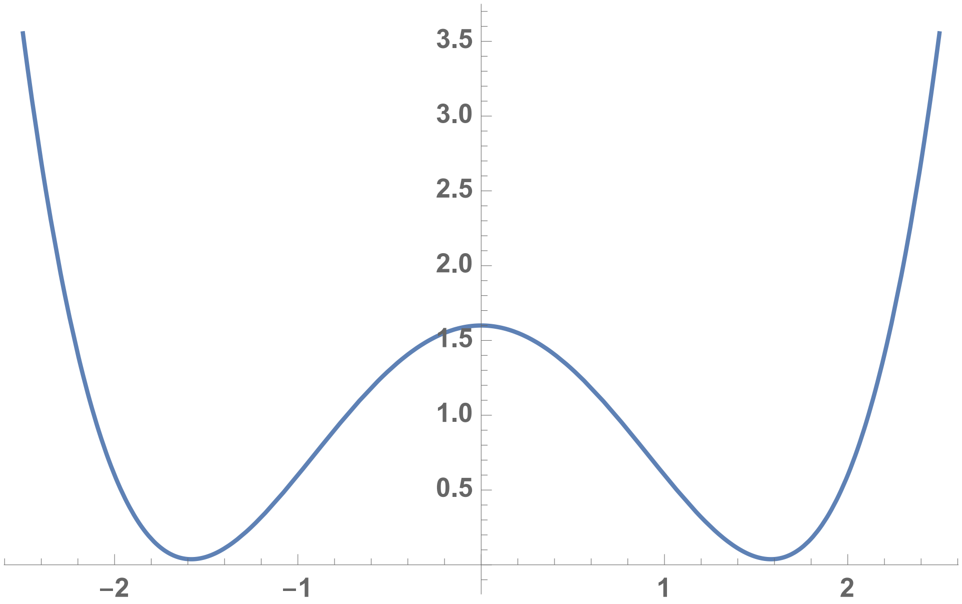


Figure 1.2: Example of a quartic polynomial potential function

1.1.2 Stochastic dynamics

The stochastic version of dynamics of catastrophe model can be obtained by adding a diffusion term to Equation 1.1. The corresponding stochastic differential equation is expressed as

$$dX_t = -\frac{dV\left(X_t;\theta\right)}{dx}dt + \sqrt{\varepsilon}dW_t. \tag{1.2}$$

One common approach to gain information about Equation 1.2 is to study the transition density that completely characterizes the stochastic dynamics. To better illustrate, suppose we have a diffusion process X_t descried by the SDE

$$dX_t = \mu\left(X_t, t; \theta\right)dt + \sigma\left(X_t, t, \theta\right)dW_t, \tag{1.3}$$

with drift μ and diffusion σ.

Let $p(x,t)$ to be the transition probability density function governed by Equation 1.3, i.e.

$$p(x,t) \equiv \frac{d}{du}\,\text{Prob}\{X_t < u | X_0 = x_0\}.$$

$p(x,t)$ is known to satisfy the *Fokker-Planck equation*, which is also commonly known as the Kolmogorov forward equation

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[\sigma^2(x,t)p(x,t)]. \tag{1.4}$$

Fokker-Planck equation describes the time evolution of the transition density $p(x,t)$ governed by Equation 1.3.

Unfortunately, unlike few those SDEs with simple expressions for both drift and diffusion terms, the transition density $p(x,t)$ of Equation 1.2 does not have analytic solution in most cases. Nonetheless, a less ambitious goal that is to obtain the stationary distribution could be achieved straightforwardly. The stationary density π is

$$\pi(x) = Ne^{-\frac{2V(x)}{\varepsilon}}, \tag{1.5}$$

where N is the normalizing constant [Cobb, 1981] and proof is given in Appendix A.

1.2 Cusp catastrophe model

The thesis studies *cusp model*, one of the elementary catastrophe models developed within the framework of catastrophe theory. *Cusp model* considers the case when the potential is quartic polynomial, for example,

$$V(x;\alpha,\beta) = \frac{1}{4}x^4 - \frac{1}{2}\beta x^2 - \alpha x.$$

According to Equation 1.1, the deterministic cusp dynamics therefore has the form

$$\frac{dx}{dt} = \alpha + \beta x - x^3. \tag{1.6}$$

An object that obeys Equation 1.6 has equilibria when

$$\frac{dx}{dt} = \alpha + \beta x - x^3 = 0,$$

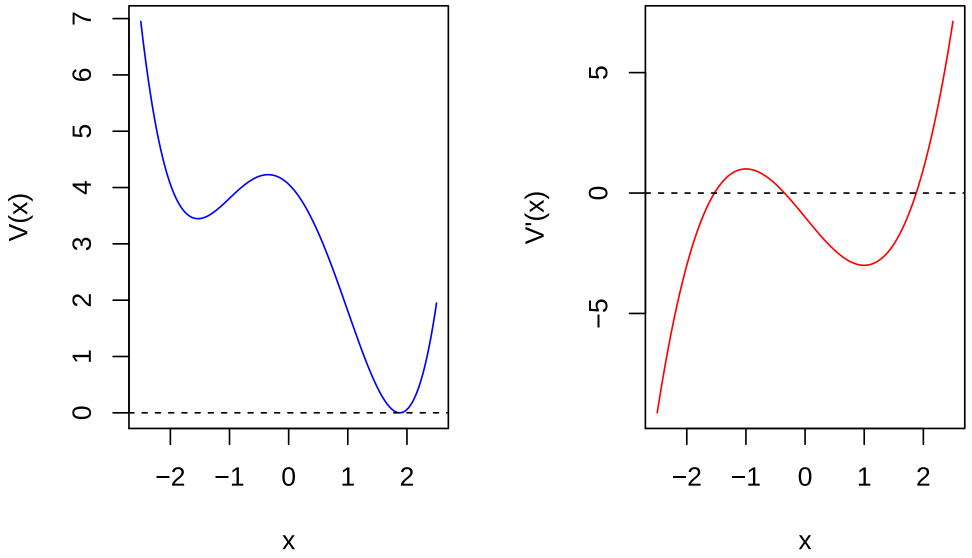


Figure 1.3: Example of potential function $V(x)$: $\alpha = 1, \beta = 3$

hence finding the equilibria is equivalent to finding the roots of the cubic function

$$f(x) = \alpha + \beta x - x^3. \tag{1.7}$$

Discriminant, defined as $\Delta^{\text{disc}} = 27\alpha^2 - 4\beta^3$ is often used as an aid to its classification of solution. In particular,

1. If $\Delta^{\text{Disc}} < 0$, Equation 1.7 will have three distinct real roots (Figure 1.3) ;
2. If $\Delta^{\text{Disc}} > 0$, Equation 1.7 will have only one real root (Figure 1.4);
3. If $\Delta^{\text{Disc}} = 0$, Equation 1.7 will have two distinct roots with one of the two being a double root (Figure 1.5).

Furthermore, the stochastic version to Equation 1.6, according to Equation 1.2 is

$$dX_t = \left(\alpha + \beta X_t - \frac{1}{4}X_t^3\right) dt + \sqrt{\varepsilon} dW_t, \tag{1.8}$$

where α, β are often referred to as asymmetry and bifurcation parameters respectively. Parameters α, β along with Δ^{Disc} differentiate the cusp model from other linear models.

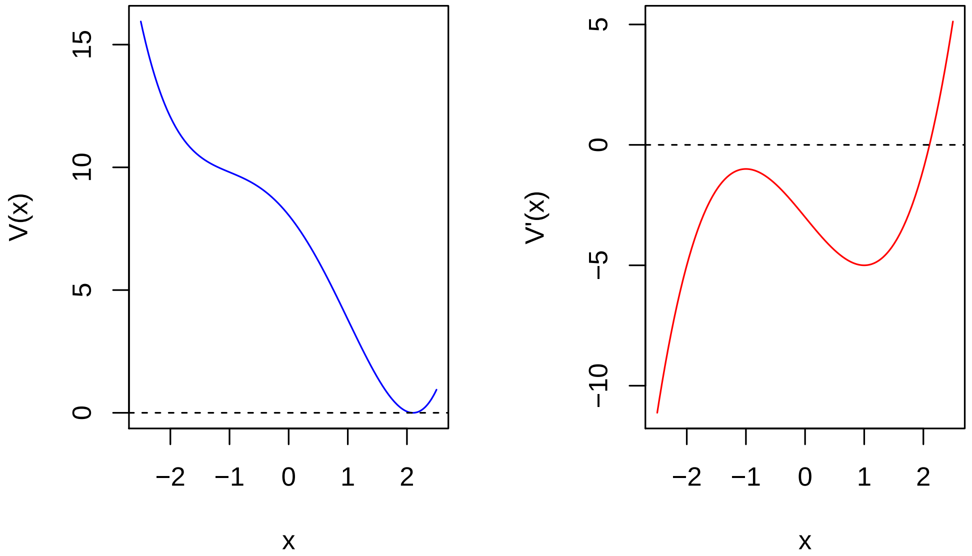


Figure 1.4: Example of potential function $V(x)$: $\alpha = 3, \beta = 3$

1.2.1 Cusp stationary density

According to Equation 1.5, the stationary density of the cusp stochastic differential equation has the form

$$\pi(x) = N \exp\left\{\frac{2}{\varepsilon}\ \left(\alpha x + \frac{1}{2}\beta x^2 - \frac{1}{4}x^4\right)\right\} \tag{1.9}$$

The cusp stationary density is characterized by two parameters α and β via $\Delta^{\text{disc}} = 27\alpha^2 - 4\beta^3$. In particular,

1. If $\Delta^{\text{disc}} > 0$, the stationary density distribution is unimodal. The asymmetric factor α and bifurcation factor β measure *skewness* and *kurtosis* respectively (Figure 1.6).

2. If $\Delta^{\text{disc}} < 0$, the stationary density distribution is *bimodal*. In this case α represent the relative height of the two modes, and β determines the separation of the two modes (Figure 1.8).

3. Moreover, the modes of stationary density (in either case) correspond to the stable equilibria of a differential equation.

One of most distinctive characteristics of the cusp stationary density is the flexibility to allow for bimodality, along with kewness and kurtosis. In fact cusp stationary density is a bimodal

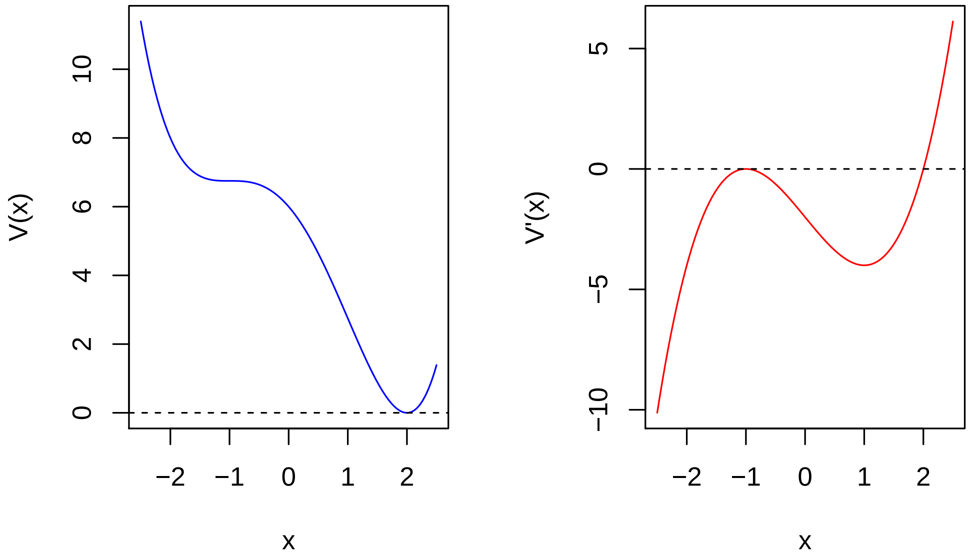


Figure 1.5: Example of potential function $V(x)$: $\alpha = 2, \beta = 3$

generalization of the Gaussian, gamma, inverse gamma, and beta distributions, and belongs to the exponential family [Cobb et al., 1983]. Such flexibility makes cusp stationary density an exceedingly appealing statistical model; for example it requires fewer parameters than corresponding mixture models (e.g. Gaussian mixture model) since it only needs four parameters. As pointed out by Chen et al. [2016], cusp stationary distribution is a complement to traditional approaches such as linear regression and non-parametric regression because of its capacity to simultaneously handle complex linear and nonlinear cases and the ability to capture sudden qualitative changes in dependent variables.

Cobb contributed to the development of stochastic cusp model by establishing an integrated structure for cusp stationary distribution, including parameter estimations of cusp stationary probability densities using method of moments and the maximum likelihood [Cobb, 1978] [Cobb, 1981].

1.2.2 Cusp model in economics

Zeeman pioneered works on applying cusp model in economics. Zeeman [1973] attempted to explain crashes on stock market by incorporating two types of major market players into cusp model, and they are *fundamentalists* and *chartists*. Their impact was reflected by incorporating fundamentalist and chartists into α and β respectively.

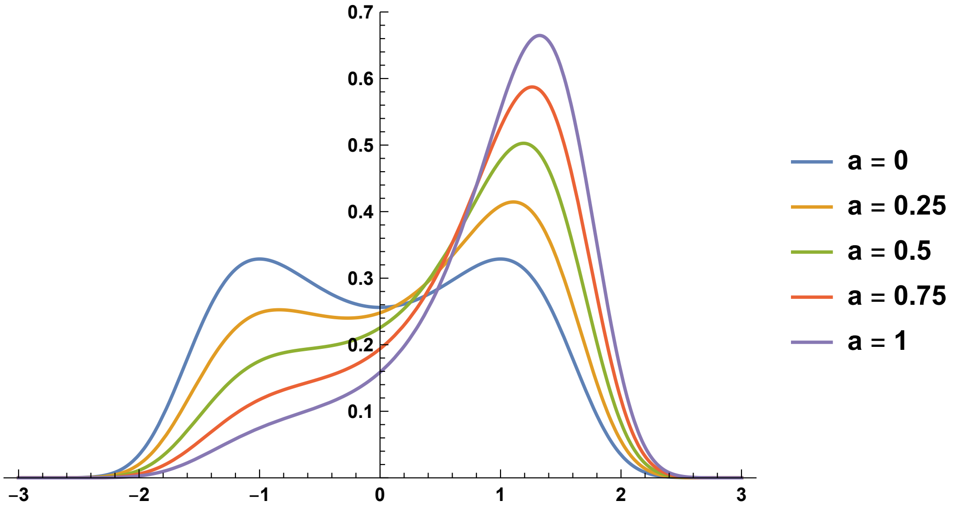


Figure 1.6: Cusp stationary density plots with varying asymmetry parameter α

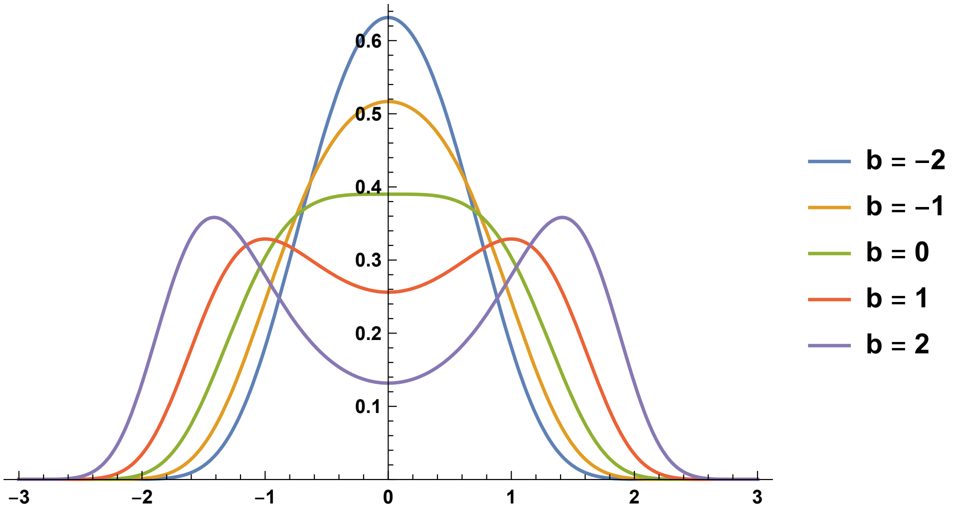


Figure 1.7: Cusp stationary density plots with varying bifurcation parameter β

Creedy made claim, based on the fundamental economics concept *law of supply and demand*, a cubic drift appears naturally in the stochastic processes governing the dynamics of asset market prices, and further explained that this is due to the characteristic of cubic equations that have the flexibility to allow one single root that corresponds to a unique equilibrium price or three roots corresponding to multiple equilibria [Creedy and Martin, 1993]. As shown in Figure 1.9, the stable market equilibrium could be impelled to an unstable one, leading to a movement in price that exhibits a large “sudden jump ” from the initial stable point. Creedy et al. [1996] applied the cusp stationary density to US/UK rate over the period 1973 and performed model parameter estimating using MLE methods.

Fernandes [2006] provided a density matching approach for time-varying parameters with exogenous variables setting, then applied it to investigate the interest rate differential during the Swedish twin crisis empirically.

Baruník and Vosvrda [2009] fitted cusp (regression) model to US stock market data and they showed that cusp (regression) model provides a better goodness-of-fit when fitting crash of stock market data than other classic regression models such as linear regression and logistic regression.

1.3 Key research problem

Although statistical properties of cusp *stationary* distribution have been well studied, unified estimating framework has established, and many empirical studies in economics or other subjects has been conducted, to best of our knowledge, little work that focuses on the actual stochastic dynamics of cusp model has been done.

One of the motivation for considering the actual stochastic cusp dynamics (Equation 1.8) comes from the application of diffusion processes in modeling stochastic behavior of economics variables in financial econometrics; it is also driven by the nature of the data one collects, in particular time-

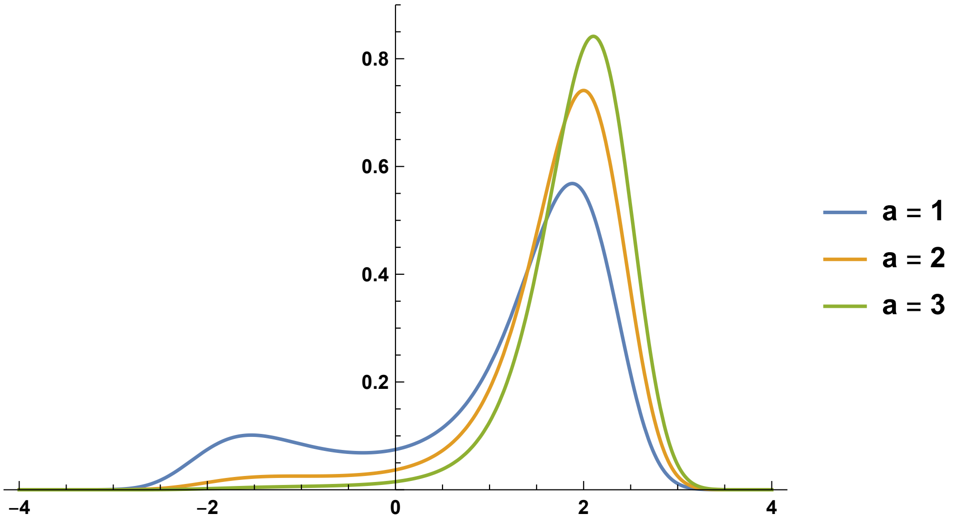


Figure 1.8: Cusp stationary density plots with fixed $\beta = 3$

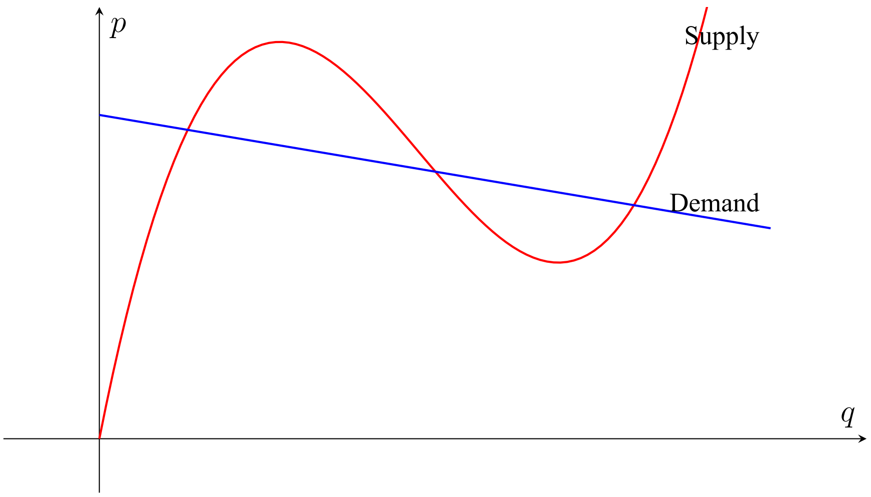


Figure 1.9: Example of stable and unstable market equilibria

series data. Time series are often collected sequentially in time. If one believes that the observed time-series are generated according to some parametric specification, developing rigorous statistical methods to calibrate the underlying model to measured observations has become an important research topic.

In reality, there's always some discrepancy between continuous-time model and discrete-time observations because the underlying model is assumed to be a diffusion process hence written in continuous time, while the available observations are often sampled discretely in time, and may not be *continuous* enough. This model-observation discrepancy should not be neglected, because otherwise it could lead to inconsistent estimators [Melino, 1996], [Jones, 1998], [Ait-Sahalia et al., 2008].

Furthermore, based on the *quality* of the data, inference problem can be specified as parameter estimation from complete observations and from partial observations. One major difference between the two is that, sample observations in partial observation scenario are considerably sparse than that in complete observation scenario. In particular, if one is able to obtain a *completely* observed data set , where this thesis simply call this case the complete observations scenario, then this model-observation discrepancy may not be so considerable. According to our simulation study, this is the relatively easier case to obtain satisfying estimation results. However, this discrepancy issue is not negligible if the sampled data is *partially* observed, and let's call it the partial observation scenario. It is not negligible because for example, the actual transition density of the process is often approximated in some way due to intractability, a larger Δ^{obs} due to missing values would lead to a larger deviation from the true but intractable transition density. Parameter estimation for cusp SDE with partially observed data points is a much harder case to handle.

Transition probability density

That being said, if one believes the observed values are generated according to the underlying parametric specification, it is naturally concerned with parameter inference for the underlying diffusion process X_t. Imagine data points are observed at discrete time points $t_i = i\Delta^{obs}$ with $i = 0, \cdots, n$, and observed measurements being $x = (x_0, \cdots, x_n)$.

With the assumption that the transition density $p(\Delta, x|x_0; \theta)$, the conditional density of $X_{t+\Delta} = x$ given $X_t = x_0$ is available in the explicit form, by Markov property, the likelihood function is

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta\left(\Delta, x_i|x_{i-1}\right) p_\theta\left(x_0\right). \tag{1.10}$$

Let $\ell_n(\theta) = \log L_n(\theta)$ be the log-likelihood function, we obtain

$$\ell n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log p_\theta\left(\Delta, x_i|x_{i-1}\right) + \log\left(p_\theta\left(x_0\right)\right). \tag{1.11}$$

The thesis assumes $p_\theta(X_0) = 1$ for simplicity, since the weight of $p_\theta(X_0)$ in the likelihood $L_n(\theta)$ becomes negligible as n increases.

Cusp transition density $p(x, t)$ is analytically intractable though it is known to satisfy the Fokker-Planck equation (Equation 1.4). The research problem of this thesis is to develop an accurate and computationally feasible parameter estimation algorithm based on Bayesian principle that can be implemented in the absence of an analytically exact transition distribution for cusp model using discretely and perhaps partially sampled observations.

1.4 Major contributions of the thesis

The dissertation claims following major and original contributions to the topic:

1. Under the complete observations scenario, an accurate and computationally feasible parameter estimation algorithm based on Bayesian principle and implemented by *Hamiltonian Monte Carlo* was developed. Intensive simulation study was conducted and the results support the claim made on the algorithm.

2. Under the partial observations scenario, three parameter estimation methods were developed and compared, they are Euler approximation, closed-from approximation using Hermite polynomials by Ait-Sahalia, and Euler approximation with data augmentation. We compare the performances of three different methods by running intensive simulation studies. The result shows that the Euler approximation with data augmentation outperforms the other two in the demonstrated cases.

3. Motivated by real-world problems when only sparsely sampled observations are available (for example, longitudinal-type of data), we also investigated how the number of augmented data points would affect the parameter estimation result by simulation studies.

4. The parameter estimation algorithm was extended to more complex settings, namely the Bayesian hierarchical modeling and cusp model with time-varying parameters. The accuracy of the algorithm is supported by simulation study whose result is also presented in the thesis.

1.5 Dissertation structure

chapter 2 reviews basics of numeric method for SDE trajectory simulation. Derivation of the Taylor-Ito expansion which plays the same essential role as Taylor expansion does in numerical analysis is given. We introduce the Euler-Maruyama method for implementing the cusp SDE trajectory simulation.

chapter 3 reviews the basics of Bayesian inference as well as Markov Chain Monte Carlo as a means to sample intractable posterior distribution. We introduce Metropolis-Hasting algorithm and Hamiltonian Monte Carlo from a practitioner's perspective focusing on their motivation and implementation. We also highlight the ability of Hamiltonian Monte Carlo that reduces correlation between successive draws by utilizing Hamiltonian dynamics.

chapter 4 reviews two approaches to approximate the intractable cusp transition density, they are (1) closed-form approximation using finite Hermite polynomials by Ait-Sahalia [2002], Ait-Sahalia et al. [2008], and (2) the Euler approximation. We compare two different approximations by examining their plots with different time increments.

chapter 5 reviews MCMC convergence diagnostics. Regardless of whether the parameter estimation results are satisfying or not, it is necessary to make sure the obtained samples were indeed coming from the stationary distribution (hence the target posterior distribution) of the constructed Markov chain as we expect. Several commonly used MCMC convergence diagnostics are introduced and implemented to spot anything undesired.

chapter 6 contains an intensive simulation study under the complete observations scenario. The results support the claim that the proposed parameter estimation algorithm is accurate and computationally feasible.

chapter 7 considers the partial observation scenario. We introduce the idea behind formulating partial observation as missing value problem and attempt to improve the estimation accuracy with data augmentation. Simply saying, data augmentation in Bayes treats those unobserved or missing data points between two consecutive observed as unknown parameters in addition to the unknown model parameters. In particular, our simulation study shows that data augmentation outperforms both closed-form approximation by Hermite polynomials and Euler approximation in demonstrated cases. We also investigated how number of augmented data points would affect the parameter estimation result by running simulation studies.

chapter 8 showed one great advantage of using Bayesian inference by considering two complex model settings, they are the Bayesian hierarchical modeling and the time-varying parameter with exogenous processes setting. Accuracy of the algorithm is supported by simulation study whose result is also presented in this chapter.

CHAPTER 2

Numerical Methods

Cusp SDE, like most of SDEs, does not have explicit or analytic solution. Consequently, trajectory simulation is needed in order to gain information about it. By choosing and implementing appropriate discretization schemes, numerical simulations give approximations to the continuous solutions to the underlying cusp SDE. In this chapter, we review the basics of numerical methods for SDE.

2.1 Convergence criteria

Methods of approximation scheme are often classified based on the task objective. If one is interested in the whole trajectory, then it requires *strong* convergence; for other cases that one only needs approximation to some functional properties such as moment and distribution, *weak* convergence is often adequate.

In practice, the choice of the convergence criterion is determined by the type of the problem one is to investigate, and is often specified before constructing a numerical method and optimizing its efficiency with respect to the chosen convergence criterion [Platen, 1999].

Definition 1. A time discretization $\Pi_N([0,T])$ over the time interval $[0,T]$ contains points

$$0 = \tau_0 < \tau_1 < \cdots < \tau_n < \cdots < \tau_N = T,$$

and the step-size is commonly chosen to be $\Delta = T/N$.

2.1.1 Strong convergence

Definition 2. A time-discretized time approximation

$$Y_n := Y_{\tau_n}, \quad n \in \{0, \cdots, N\}$$

of a continuous-time process $X_t, \;\; t \in [0, T]$ governed by an SDE converges in the *strong sense* with order $\gamma \in (0, \infty]$ if for any fixed time horizon T it holds true that

$$E\left|Y_N - X_T\right| \leq K\Delta^{\gamma} \tag{2.1}$$

for all step-sizes $\Delta \in (0, 1)$ and K is a constant not depending on Δ.

Strong convergence criterion should be used when the task involving trajectory simulations directly. For example, simulation of a stock price usually requires the simulated sample trajectory to be close to the solution of geometric Brownian motion, and similarly simulation of a short-term rate usually requires the simulated sample trajectory be close to the solution Ornstein–Uhlenbeck process. In these cases, numerical methods are classified according to their strong order γ of convergence, using the absolute error

$$E\left|Y_N - X_T\right|$$

at the terminal time T.

2.1.2 Weak convergence

Definition 3. A discrete time approximation Y of a solution X of an SDE converges in the *weak* sense with order $\beta \in (0, \infty]$ if, for any polynomial g, there exists a constant $K_g < \infty$ such that

$$\left|E\left(g\left(Y_N\right)\right) - E\left(g\left(X_T\right)\right)\right| \leq K_g\Delta^{\beta} \tag{2.2}$$

for all step-sizes $\Delta \in (0, 1)$, provided that these functionals exist.

If one is only interested in computing some functional such as probability distributions or moments that does not require us to approximate the entire trajectory of X, strong convergence criterion that requires an almost exact replica of the sample path of the solution of the underling SDE may not be necessary. One typical example is the Monte Carlo simulation of option prices at a terminal time T, where option prices can be approximated by simply random walk instead of Brownian motion due to the fact that its first two moments (mean and variance) match the ones for Brownian motion correspondingly. In such cases, approximations of probability distribution that corresponds to X is often sufficient, and consequently only a much weaker type of convergence is needed.

Theorem 1. $\beta \geq \alpha$ (Weak order $\geq$ Strong order)

Proof. Suppose $|f'| \leq K$, then by mean value theorem $\left|\int g - h\right| \leq \int |g-h|$

$$\begin{aligned} |\mathbb{E}f\left(Y_N\right) - \mathbb{E}f\left(X_T\right)| &\leq \mathbb{E}\left|f\left(Y_N\right) - f\left(X_T\right)\right| \\ &\leq K\mathbb{E}\left|Y_N - X_T\right| \\ &\leq K\left(\mathbb{E}\left|Y_N - X_T\right|^2\right)^{1/2}. \end{aligned}$$

□

2.2 Ito-Taylor expansion

Taylor series expansion plays an essential role in numerical analysis. For a sufficiently smooth function $f(x)$ in a neighborhood of some given point x_0, one could obtain approximations to any desired order of accuracy by truncating the Taylor series up to certain term.

We now briefly review the derivation of Ito-Taylor expansion, which plays the role of Taylor series in the stochastic setting.

Given a diffusion process $X(t)$ that obeys

$$dX(t) = \mu\left(X\left(t\right)\right)dt + \sigma\left(X\left(t\right)\right)dW(t), \tag{2.3}$$

we make a further assumption that $\mu = \mu[X(t)], \mu = \mu[X(t)]$ (i.e. they do not depend on time explicitly).

Applying Ito lemma to any twice differentiable function f would lead to

$$df[X(t)] = \left\{ \mu \frac{\partial}{\partial X} f[X(t)] + \frac{1}{2}\sigma^2[X(t)] \frac{\partial^2}{\partial X^2} f[X(t)] \right\} dt + \sigma[X(t)] \frac{\partial}{\partial X} f[X(t)] dW(t). \quad (2.4)$$

To simplify the expression, we define the following two operators:

$$\mathcal{L}^0 \equiv \mu \frac{\partial}{\partial X} + \frac{1}{2}\sigma^2[X] \frac{\partial^2}{\partial X^2} \quad \text{and} \quad \mathcal{L}^1 \equiv \sigma[X] \frac{\partial}{\partial X}.$$

Hence Equation 2.4 can be notation-wise simplified to

$$df[X(t)] = \mathcal{L}^0 f[X(t)] dt + \mathcal{L}^1 f[X(t)] dW(t). \quad (2.5)$$

An equivalent expression in integral form gives

$$f[X(t)] = f\left[X\left(t_0\right)\right] + \int_{t_0}^{t} \mathcal{L}^0 f[X(s)] ds + \int_{t_0}^{t} \mathcal{L}^1 f[X(s)] dW(s). \quad (2.6)$$

Since Ito lemma holds for any twice differentialble function f, if we specify our choice of function $f(x)$ to be $f(x) = x$, Equation 2.6 gives

$$X(t) = X\left(t_0\right) + \int_{t_0}^{t} \mu[X(s)] ds + \int_{t_0}^{t} \sigma[X(s)] dW(s). \quad (2.7)$$

Similarly, by choosing $f(x) = \mu(x)$, and $f(x) = \sigma(x)$, and apply Ito lemma, Equation 2.6 gives

$$\mu[X(t)] = \mu\left[X\left(t_0\right)\right] + \int_{t_0}^{t} \mathcal{L}^0 \mu[X(s)] ds + \int_{t_0}^{t} \mathcal{L}^1 \mu[X(s)] dW(s), \quad (2.8)$$

and

$$\sigma[X(t)] = \sigma\left[X\left(t_0\right)\right] + \int_{t_0}^{t} \mathcal{L}^0 \sigma[X(s)] ds + \int_{t_0}^{t} \mathcal{L}^1 \sigma[X(s)] dW(s). \quad (2.9)$$

Substituting Equation 2.8 and Equation 2.9 into Equation 2.7 leads to

$$\begin{aligned}X(t) =&X\left(t_0\right)+\int_{t_0}^t\left\{\mu\left[X\left(t_0\right)\right]+\int_{t_0}^{s_1} \mathcal{L}^0 \mu\left[X\left(s_2\right)\right] d s_2+\int_{t_0}^{s_1} \mathcal{L}^1 \mu\left[X\left(s_2\right)\right] d W\left(s_2\right)\right\} d s_1 \\ &+\int_{t_0}^t\left\{\sigma\left[X\left(t_0\right)\right]+\int_{t_0}^{s_1} \mathcal{L}^0 \sigma\left[X\left(s_2\right)\right] d s_2+\int_{t_0}^{s_1} \mathcal{L}^1 \sigma\left[X\left(s_2\right)\right] d W\left(s_2\right)\right\} d W\left(s_1\right).\end{aligned} \tag{2.10}$$

Note that, in the above equation, $\mu[X(t_0)]$ and $\sigma(t_0)]$ are inside integrand, however both remains constant in time. Therefore, we separate the two terms out from the remaining terms, which leads to

$$X(t)=X\left(t_0\right)+\mu\left[X\left(t_0\right)\right] \int_{t_0}^t d s_1+\sigma\left[X\left(t_0\right)\right] \int_{t_0}^t d W\left(s_1\right)+R, \tag{2.11}$$

with reminder term R being

$$\begin{aligned}R \equiv & \int_{t_0}^t \int_{t_0}^{s_1} \mathcal{L}^0 \mu\left[X\left(s_2\right)\right] d s_2 d s_1+\int_{t_0}^t \int_{t_0}^{s_1} \mathcal{L}^1 \mu\left[X\left(s_2\right)\right] d W\left(s_2\right) d s_1 \\ & +\int_{t_0}^t \int_{t_0}^{s_1} \mathcal{L}^0 \sigma\left[X\left(s_2\right)\right] d s_2 d W\left(s_1\right)+\int_{t_0}^t \int_{t_0}^{s_1} \mathcal{L}^1 \sigma\left[X\left(s_2\right)\right] d W\left(s_2\right) d W\left(s_1\right).\end{aligned} \tag{2.12}$$

Ito-Taylor expansion with higher order terms

Higher order accuracy could be achieved If one uses substitution repeatedly to obtain constant integrands in higher order terms. To better illustrate how substitution works, we continue working on the remainder term R Equation 2.12. In particular, the very last term in R, namely

$$\int_{t_0}^t \int_{t_0}^{s_1} \mathcal{L}^1 \sigma\left[X\left(s_2\right)\right] d W\left(s_2\right) d W\left(s_1\right) \tag{2.13}$$

is of the lowest order in Δt in R. This is a simple rule according to the *box calculus*:

$$dt \cdot dt = 0, \quad dt \cdot dW_t = 0, \quad \text{and} \quad dW_t \cdot dW_t = dt.$$

Applying Ito's lemma to $f(x) = \mathcal{L}^1 b(x)$ gives

$$\mathcal{L}^1\sigma\left[X\left(s_2\right)\right] = \mathcal{L}^1\sigma\left[X\left(t_0\right)\right] + \int_{t_0}^{s_2} \mathcal{L}^0\mathcal{L}^1\sigma\left[X\left(s_3\right)\right] ds_3 + \int_{t_0}^{s_2} \mathcal{L}^1\sigma\left[X\left(s_3\right)\right] dW\left(s_3\right). \tag{2.14}$$

Substituting Equation 2.14 to Equation 2.13 leads to

$$\begin{aligned} &\int_{t_0}^{t}\int_{t_0}^{s_1} \mathcal{L}^1\sigma\left[X\left(s_2\right)\right] dW\left(s_2\right) dW\left(s_1\right) \\ &= \int_{t_0}^{t}\int_{t_0}^{s_1} \left\{\mathcal{L}^1\sigma\left[X\left(t_0\right)\right] + \int_{t_0}^{s_2} \mathcal{L}^0\mathcal{L}^1\sigma\left[X\left(s_3\right)\right] ds_3\right. \\ &\left. + \int_{t_0}^{s_2} \mathcal{L}^1\sigma\left[X\left(s_3\right)\right] dW\left(s_3\right)\right\} dW\left(s_2\right) dW\left(s_1\right). \end{aligned} \tag{2.15}$$

Clearly $\mathcal{L}^1\sigma[X(t_0)]$ is the constant in time, hence we separate it out from the remaining terms, which in turns leads to

$$\int_{t_0}^{t}\int_{t_0}^{s_1} \mathcal{L}^1\sigma\left[X\left(t_0\right)\right] dW\left(s_2\right) dW\left(s_1\right). \tag{2.16}$$

Note that

$$\mathcal{L}^1\sigma = \sigma\sigma',$$

therefore Equation 2.16 becomes

$$\int_{t_0}^{t}\int_{t_0}^{s_1} \mathcal{L}^1\sigma\left[X\left(t_0\right)\right] dW\left(s_2\right) dW\left(s_1\right) = \sigma\left[X\left(t_0\right)\right]\sigma'\left[X\left(t_0\right)\right]\int_{t_0}^{t}\int_{t_0}^{s_1} dW\left(s_2\right) dW\left(s_1\right).$$

Consequently, Equation 2.11 becomes

$$\begin{aligned} X(t) =& X\left(t_0\right) + \mu\left[X\left(t_0\right)\right]\int_{t_0}^{t} ds_1 + \sigma\left[X\left(t_0\right)\right]\int_{t_0}^{t} dW\left(s_1\right) \\ &+ \sigma\left[X\left(t_0\right)\right]\sigma'\left[X\left(t_0\right)\right]\int_{t_0}^{t}\int_{t_0}^{s_1} dW\left(s_2\right) dW\left(s_1\right) + \tilde{R}, \end{aligned} \tag{2.17}$$

where $\tilde{R}$ being a new remainder.

By applying Ito' lemma again, the double Ito's integral gives

$$\int_{t_0}^{t} \int_{t_0}^{s_1} dW\left(s_2\right) dW\left(s_1\right) = \frac{1}{2}\left[W(t) - W\left(t_0\right)\right]^2 - \frac{1}{2}\left(t - t_0\right)$$

Consequently,

$$\begin{aligned} X(t) = & X\left(t_0\right) + \mu\left[X\left(t_0\right)\right] \int_{t_0}^{t} ds_1 + \sigma\left[X\left(t_0\right)\right] \int_{t_0}^{t} dW\left(s_1\right) \\ & + \sigma\left[X\left(t_0\right)\right] \sigma'\left[X\left(t_0\right)\right] \left\{\frac{1}{2}\left[W(t) - W\left(t_0\right)\right]^2 - \frac{1}{2}\left(t - t_0\right)\right\} + \widetilde{\mathcal{R}} \end{aligned} \tag{2.18}$$

2.3 Euler–Maruyama method

Euler–Maruyama method is a numerical method for approximating the solution of a stochastic differential equation (SDE), and not surprisingly, it's generalizes Euler method in ordinary differential equation case. Euler-Maruyama method is obtained by truncating Equation 2.18 at the first order terms (or simply look at Equation 2.11).

Let $\{X_t, 0 \leq t \leq T\}$ be a solution to the stochastic differential equation

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t$$

with initial deterministic value X_{τ_0} over a time window $[0, T]$. The Euler-Maruyama approximation of X_t is a continuous stochastic process $Y_n := Y_{\tau_n}$ satisfying the following iterative relation

$$Y_{i+1} = Y_i + \mu\left(Y_i\right)\left(t_{i+1} - t_i\right) + \sigma\left(Y_i\right)\left(W_{i+1} - W_i\right), \tag{2.19}$$

for $i = 0, 1, \cdots, N-1$ with $Y_0 = X_0$ and $W_{i+1} - W_i \sim \mathcal{N}(0, \sqrt{\Delta})$.

Below is a simple implementation of Euler-Maruyama method when simulating sample trajectory of cusp SDE

```
x[1] = 0.5
for (i in 2:N){
  x[i] = x[i-1] + (alpha + beta * x[i-1] - x[i-1]^3) * Dt + s * rnorm(1,0,sqrt(Dt))
}
```

Listing 2.1: Euler's method

2.4 Milstein method

By including the second-order term in Equation 2.18, one obtains the Milstein method which has a higher accuracy of the approximation compared to Euler method.

The Milstein method is in the following form,

$$\begin{aligned} Y_{i+1} =& Y_i + \mu(t_i, Y_i)(t_{i+1} - t_i) + \sigma(t_i, Y_i)(W_{i+1} - W_i) \\ &+ \frac{1}{2}\sigma(t_i, Y_i)\sigma_x(t_i, Y_i)\{(W_{i+1} - W_t)^2 - (t_{i+1} - t_i)\}. \end{aligned} \tag{2.20}$$

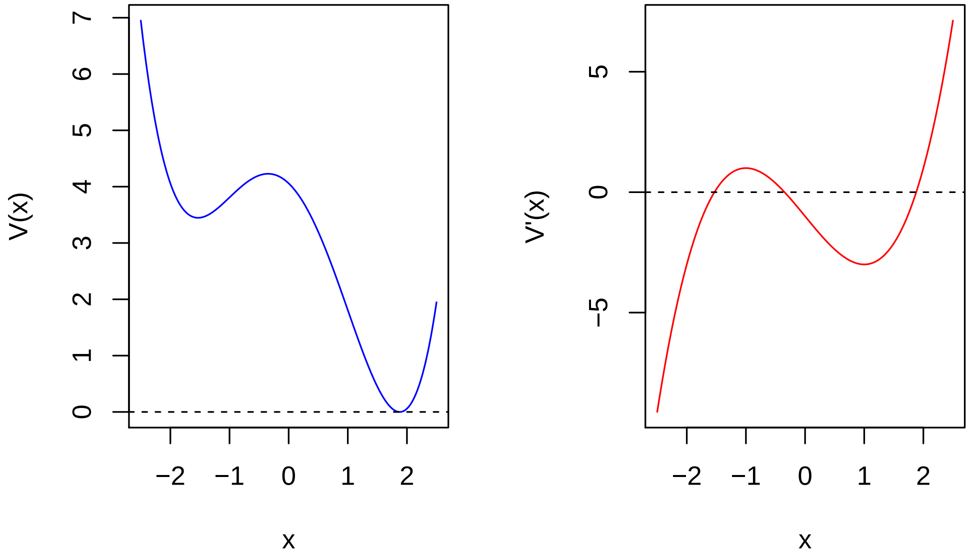
The Milstein scheme has both weak and strong order of convergence Δ, which is, not surprisingly, superior to the Euler–Maruyama method who has the same weak order of convergence, Δ, but inferior strong order of convergence, $\sqrt{\Delta}$. However in our case, since the diffusion term is constant hence does not depend on X_t, Milstein method is in fact equivalent to the Euler-Maruyama method.

2.5 Cusp SDE trajectory simulations

Recall in chapter one we discussed the sign of Δ^{Disc} determines the number of roots of a cubic function, hence equilibra of the cusp dynamics. In this section, we run trajectory simulations for three different cases, namely $\Delta^{\text{Disc}} < 0$, $\Delta^{\text{Disc}} > 0$, and $\Delta^{\text{Disc}} = 0$. For all the following trajectory simulations, we fix the diffusion coefficient σ to be 2 and treat it as known instead of unknown model parameter.

2.5.1 Three roots

When $\Delta^{\text{Disc}} < 0$, the cusp deterministic dynamic system has three equilibira: two stable attracting equilibria divided by one repelling unstable equilibrium and the cusp stationary density distribution is bimodal.



Example of potential function $V(x)$: $\Delta^{\text{Disc}} < 0$

One sample trajectory using Euler's method is given by Figure 2.1. In this case, we observe one appealing feature hence advantage to the stochastic cusp model over more traditionally used linear mean-reverting models in describing certain systems - that is a regime-switching type of behavior with regime being two stable equilibra. A similar behavior to this *bimodal* cusp SDE could be obtained by using a switching model involving two linear mean-reverting models.

2.5.2 One root

When $\Delta^{\text{Disc}} > 0$, the cusp deterministic dynamic system has only one equilibrium which is a stable one.

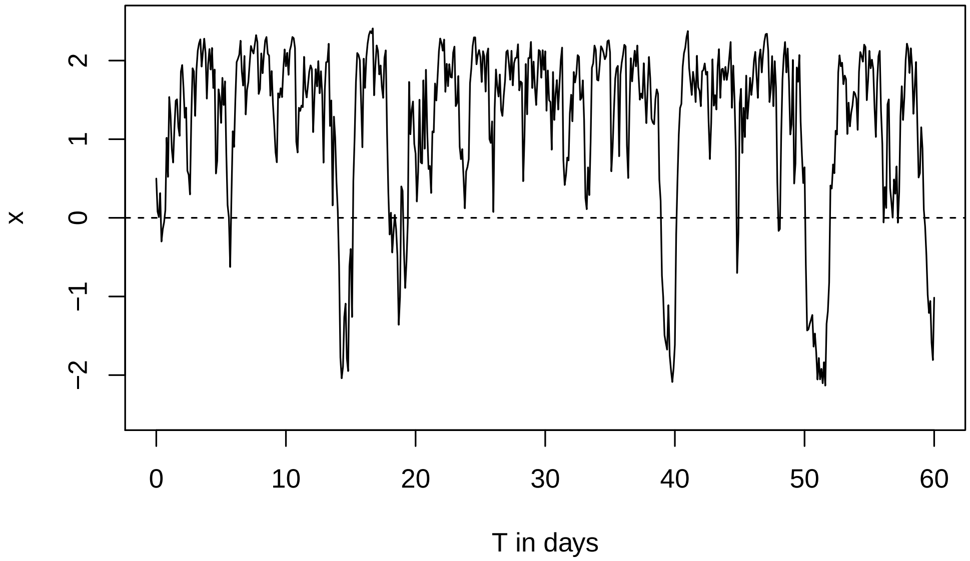


Figure 2.1: Sample trajectory: Cusp SDE with $\alpha = 1, \beta = 3$

In this case, the cusp stationary density distribution is unimodal. According to Figure 2.2 which compares cusp model with the famous Vasicek model,

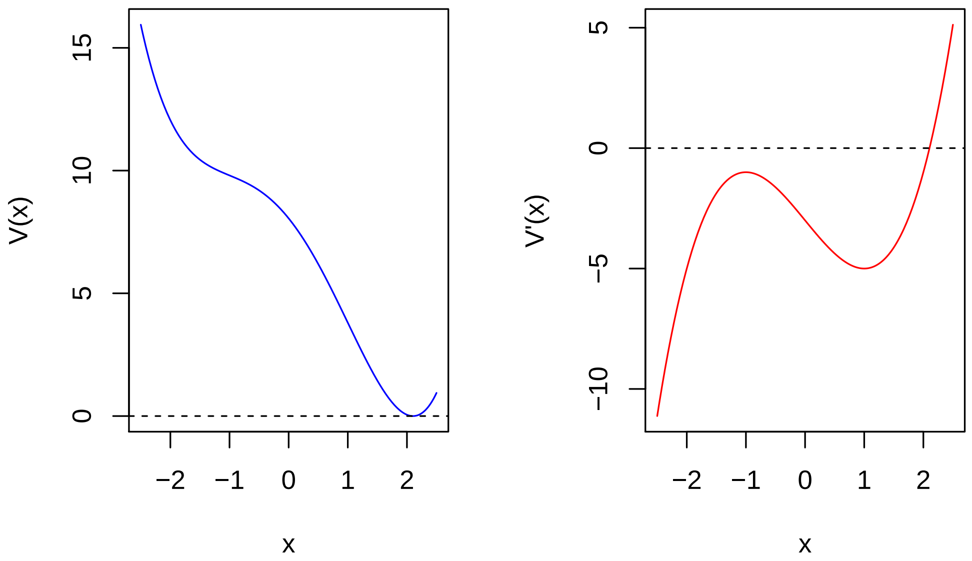
$$dX_t = \theta\left(\mu - X_t\right)dt + \sigma dW_t, \tag{2.21}$$

the sample trajectory exhibits stronger mean reversion than the mean-reverting models with Gaussian stationary densities (for example, Vasicek model in this case) due to the drift term contains a cubic term, which supports the claim by [Ait-Sahalia, 1996].

2.5.3 Two roots

When $\Delta^{\text{Disc}} = 0$, the cusp deterministic dynamic system has two equilibrium, one being stable and the other unstable. The unstable equilibrium corresponds to the double root of the cubic function.

In this case, the cusp stationary density distribution is also unimodal, and the behavior from the sample trajectory (Figure 2.3) is very similar to the one-root case (Figure 2.2).



Example of potential function $V(x)$: $\Delta^{\text{Disc}} > 0$

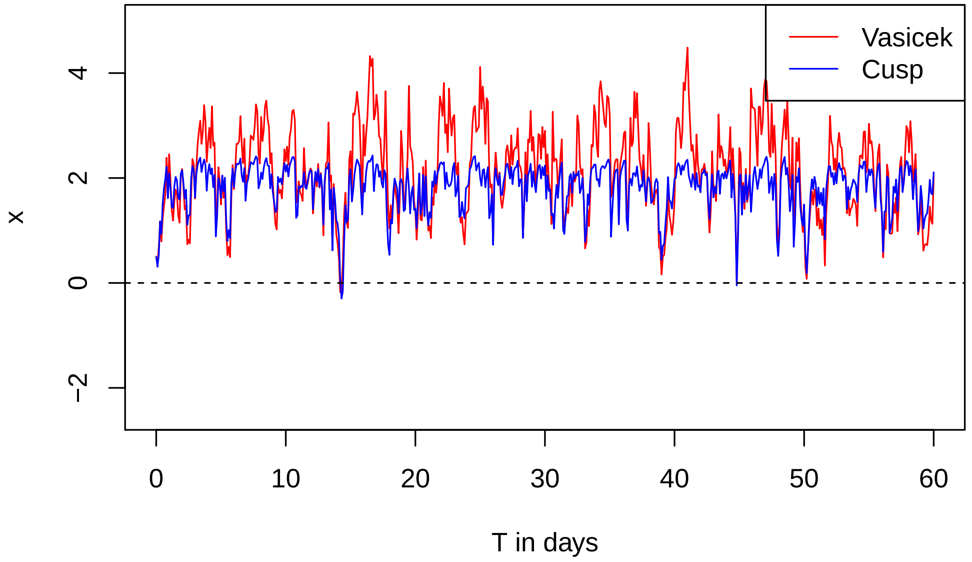
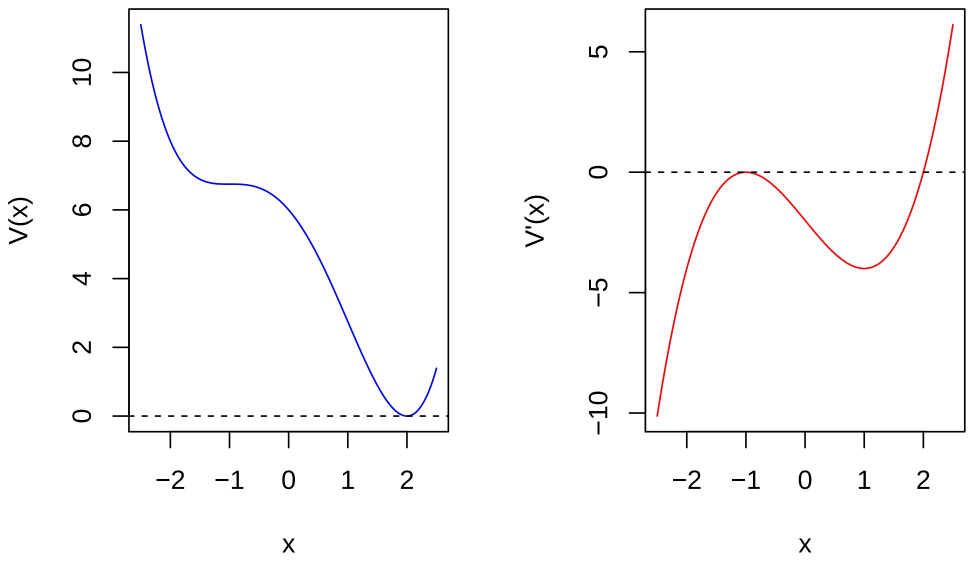


Figure 2.2: Sample trajectory: Cusp SDE with $\alpha = 3, \beta = 3$



Example of potential function $V(x)$: $\Delta^{\text{Disc}} = 0$

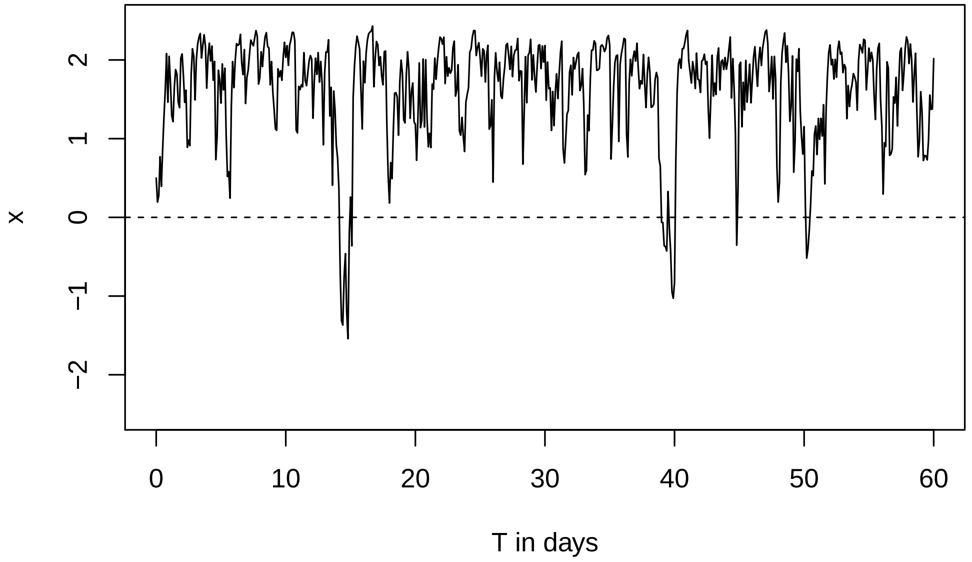


Figure 2.3: Sample trajectory: Cusp SDE with $\alpha = 2, \beta = 3$

CHAPTER 3

Bayesian Inference using Markov Chain Monte Carlo

Model parameters of cusp SDE, namely asymmetry parameter α and bifurcation parameter β, have well-defined geometric and statistical interpretations, and their values further determine the unique structural behaviors of cusp dynamics. This makes it particularly important that the model parameter values need to be estimated accurately and properly; therefore when doing statistical inference on model parameters, we would expect to include not only their most likely values or point estimations, but also the associated uncertainties, for example confidence interval estimations.

Bayesian inference is a method of statistical inference where Bayes' theorem is used to update one's prior belief on probability of unknown quantity based on observed data. The goal in carrying through Bayesian inference is to do parameter estimations for cusp model with discretely sampled data points. This chapter first reviews basis of Bayesian inference, followed by discussing MCMC as a means of sampling the intractable posterior distribution in our case.

3.1 Bayesian inference

When performing Bayesian inference for an unknown model parameter, one usually starts with some prior belief about it. As data comes in, one could compute and use the corresponding posterior distribution to draw conclusion about it. It is a conceptually straightforward application of Bayes's theorem:

$$P(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}, \tag{3.1}$$

where $p(\theta)$ and $P(\theta|x)$ are called the *prior* and *posterior distribution* respectively.

Bayesian inference naturally associates unknown model parameters to some probability distributions and explores the entire distribution region rather than searching for optimal of a given function such as likelihood function. Consequently, the uncertainty over the range of model parameter values could be estimated naturally and directly.

3.1.1 Prior belief

Bayesian inference offers flexibility by incorporating prior belief into model inference. For example, if an expert (subjectively but reasonably) believes that some parameters might be more likely than others, that knowledge from the expert can be elicited and used as prior information in Bayesian inference. The combination of prior knowledge with likelihood of the observation will result in the posterior distribution. For other cases where little prior information is available, a flat prior, possibly improper, or other non-informative priors such as Jeffrey's prior would be a more proper choice to reflect that (objective) belief.

3.1.2 Connections between MLE and MAP

In statistical inference, from *Frequentists'* viewpoint, *optimal* model parameters are usually those which maximize the likelihood of the observations. Optimality is usually achieved by applying various hill-climbing type of optimization methods. The optimization result is therefore point estimate of parameter value; one could also construct confidence intervals utilizing the asymptotically Gaussian property of MLE.

In particular, Frequentists would seek for a vector of parameters, θ, that

$$\theta_{\text{MLE}} = \arg\max_{\theta} p(x|\theta),$$

or equivalently the log likelihood function

$$\theta_{\text{MLE}} = \arg\max_{\theta} \log p(x|\theta). \tag{3.2}$$

In Bayesian inference, a common choice for point estimation is the Maximum A Posteriori, or MAP. As the name suggests, the opimality is associated with posterior distribution, instead of the likelihood function:

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} p(\theta|x) \\ &= \arg\max_{\theta} \frac{p(x|\theta)p(\theta)}{p(x)} \\ &= \arg\max_{\theta} p(x|\theta)p(\theta),\end{aligned} \tag{3.3}$$

or similarly the log-likelihood function

$$\theta_{\text{MAP}} = \arg\max_{\theta} \log p(x|\theta)p(\theta). \tag{3.4}$$

When comparing MLE (Equation 3.2) and MAP (Equation 3.4), the only thing differs is the inclusion of prior $p(\theta)$ in MAP. It's not so surprising to see MLE is equivalent to MAP if one chooses to use the flat prior, perhaps the simplest prior. To better illustrate this,

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} \log p\left(x|\theta\right) p(\theta) \\ &= \arg\max_{\theta} \log p\left(x|\theta\right) \times \text{constant} \\ &= \arg\max_{\theta} \log p\left(x|\theta\right) \\ &= \theta_{\text{MLE}}.\end{aligned} \tag{3.5}$$

3.1.3 Bayesian data augmentation

As discussed earlier, there's a discrepancy between the continuous underlying model assumption and the discretely sampled data points in reality. The discrepancy is even more magnified when the observed data points are considerably sparse.

One approach to remedy the discrepancy is via data augmentation. Bayesian inference naturally supports this approach by treating any missing/unobserved data as additional parameters and to estimate them along with the unknown model parameters in the posterior [Gelman et al., 2013]. More detailed illustration and simulation study is given in chapter 6.

3.1.4 Motivation of MCMC in Bayesian inference

In Bayesian inference, once prior knowledge is incorporated into the statistical model or the likelihood function, one expects to perform Bayesian analysis based on the posterior distribution. Unfortunately, posterior distributions do not have explicit analytic forms in most cases. One way to resolve this is to utilize prior distribution that is conjugate with respect to the underlying statistical model. In this conjugate case, posterior distribution $p(\theta|x)$ is in the same probability distribution family as the prior distribution $p(\theta)$ so that one could easily obtain posterior distribution in analytic form. However this approach often fails when the specified prior is not conjugate.

Since analytic posterior distribution is difficult to obtain, a different approach would to obtain posterior distribution empirically. In particular, one would aim for taking a collection of samples that are draw from the posterior distribution, and we hope this collection of samples could be used to represent the true but intractable posterior distribution.

Vanilla Monte Carlo would not work in this case, since it still requires samples to be drawn from a target distribution (posterior distribution in this case) which as analytic form, which one does not have.

Markov Chain Monte Carlo (MCMC) provides an alternative route to the vanilla Monte Carlo. MCMC, as its name suggests, utilize a knowingly constructed Markov chain whose equilibrium distribution is the target distribution. In Bayesian inference, the target distribution is usually the posterior distribution. To better explain, let's assume one would like to draw samples from a target distribution $p(\theta|x)$ with a prior distribution $p(\theta)$. According to Bayes' theorem

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{\int p(\eta)p(x|\eta)d\eta}, \tag{3.6}$$

the denominator usually requires high dimensional integration hence analytically intractable.

Instead of draw i.i.d samples from $p(\theta|x)$ directly, which is exactly what vanilla Monte Carlo does, MCMC aims for construction of a large collection of θ values, so that the empirical distribution

$$\{\theta^{(1)}, \cdots, \theta^{(S)}\}$$

approximates $p(\theta|x)$.

In this section we introduce two MCMC methods, namely the Metropolis-Hasting algorithm and Hamiltonian Monte Carlo. Since there's really a huge literature on MCMC and HMC, we decide to take the practitioner's approach focusing on their implementations rather than the theoretical justification. This chapter greatly benefits from Brooks et al. [2011].

3.2 Metropolis-Hasting

That being said, MCMC involves construction of a Markov chain whose stationary distribution is the target distribution. To better illustrate, let's assume there already exists a collection of samples $\{\theta^{(1)}, \cdots, \theta^{(s)}\}$ where $\theta^{(s)}$ being the latest draw. Now we would like to expand the existing collection by adding some new value $\theta^{(s+1)}$ to it.

3.2.1 Construction of a Markov chain

Start with a proposal value θ^{*}, which often close to the latest draw $\theta^{(s)}$. The question is whether the proposed value θ^{*} should be included into the existing collection or not. Intuitively, for any two different values θ_a and θ_b, one should expect

$$\frac{\#\left\{\theta^{(s)}\text{ 's in the collection } = \theta_a\right\}}{\#\left\{\theta^{(s)}\text{ 's in the collection } = \theta_b\right\}} \approx \frac{p\left(\theta_a|y\right)}{p\left(\theta_b|y\right)} \tag{3.7}$$

in the collection [Hoff, 2009].

Our answer to the question should be positive, if $p(\theta^{*}|y) > p(\theta^{(s)}|y)$. This is because $\theta^{(s)}$ is already in the working collection and according to the (3.7), more θ^{*}'s are expected in the set than $\theta^{(s)}$'s. Therefore, θ^{*} is expected to be accepted as well. Conversely, if $p(\theta^{*}|y) < p(\theta^{(s)}|y)$, including θ^{*} or not will be determined by the ratio of $p(\theta^{*}|y)$ to $p(\theta^{(s)}|y)$ that represents their relative frequencies. The ratio comparison could be made without even computing $p(\theta|y)$ explicitly, because

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)} \frac{p(y)}{p(y|\theta^{(s)})p(\theta^{(s)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})} \tag{3.8}$$

One great advantage over the vanilla Monte Carlo is that the target distribution only needs to be proportional to the posterior distribution. This means evaluation the intractable marginal likelihood is not required, which is just a normalizing constant in parameters of interest.

3.2.2 Algorithm

The Metropolis algorithm produces a value $\theta^{(s+1)}$ as follows:

```
Start with a collection of samples $\{\theta^{(1)}, \cdots, \theta^{(s)}\}$ where $\theta^{(s)}$ being the latest draw;
Generate a proposal parameter value $\theta^*$ according to some proposal function $J(\theta|\theta^{(s)})$ -
  usually random-walk type;
Compute the acceptance ratio:
    $r = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})}$
Accept $\theta^*$ with following acceptance-rejection criterion:
    $\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(s)} & \text{with probability } 1 - \min(r, 1) \end{cases}$
```

Algorithm 1: Metropolis- Hasting algorithm

After some burn-in time, the Markov chain with accepted draws is expected to converge to the equilibrium distribution, regardless where the chain started initially. Samples after the burn-in time would be a good empirical approximation to the true target distribution. More detailed assessment on convergence of the chain is given in chapter 5.

3.3 Hamiltonian Monte Carlo

In practice, one often encounters the problem of large autocorrelation hence slow mixing when applying the (random-walk) Matropolis-Hasting algorithms to sample intractable posterior distribution. Inarguably, the inefficient proposal $J(\cdot|\cdot)$ is responsible for making Markov chain convergence to the target equilibrium distribution $\pi(x)$ slow. To reduce the correlation between successive sampled states, Hamiltonian Monte Carlo has been developed under the framework of MCMC and it differs from the Metropolis–Hastings algorithm by adopting a Hamiltonian dynamics between states to achieve the goal of reducing autocorrelation. By utilizing the gradient information rather than just the probability distribution alone, Hamiltonian Monte Carlo is able to explore the target distribution much more efficiently compared with metropolis-Hasting, resulting in faster convergence [Neal et al., 2011]. This section on Hamiltonian Monte Carlo benefited greatly from Neal et al. [2011]and Betancourt [2017].

3.3.1 Hamiltonian dynamics

For an object inside a dynamic system, the object's state or motion is governed by both location $x \in \mathbb{R}^n$ and momentum $p \in \mathbb{R}^n$. There is an associated potential energy, commonly denoted by $U(x)$ for each location the object takes; and similarly there's an associated kinetic energy commonly denoted by $K(p)$ for each momentum the object posses. We say the object obeys Hamiltonian dynamics if the total energy is conserved; in this case, the total energy is called the Hamiltonian, denoted by $H(x, p)$, which is defined as the sum of potential and kinetic energies for a given object:

$$H(x,p) = U(x) + K(p) \tag{3.9}$$

Hamiltonian dynamics expresses how kinetic energy and potential energy are converted to one or the other as an object inside a system moves in time, and it is mathamatically described by the Hamiltonian equations:

$$\begin{aligned} \frac{\partial x_i}{\partial t} &= \frac{\partial H}{\partial p_i} \\ \frac{\partial p_i}{\partial t} &= -\frac{\partial H}{\partial x_i} \end{aligned} \tag{3.10}$$

for given expressions for $\frac{\partial H}{\partial x_i}$ and $\frac{\partial H}{\partial p_i}$.

Once Equation 3.10 as well as an initial position x_0 and initial momentum p_0 at time t_0 are given, both the location and momentum of an object at some future time $t = t_0 + \Delta$ can be computed by simulating these dynamics for Δ unit of times.

3.3.2 The Leap Frog Method

The Hamiltonian equations (3.10) describe an object's motion in time, and the trajectory is continuous in time. It is necessary to approximate the Hamiltonian equations by discretizing over time, in order to numerically simulate the trajectory. This can be usually achieved by the *Leap Frog method*:

1. Take a half step forward in time $\delta/2$ to update the momentum variable while fixing position variable at t:
$$p_i(t + \delta/2) = p_i(t) - (\delta/2)\frac{\partial U}{\partial x_i}(t);$$

2. Take a full step forward in time to update the position variable while fixing momentum variable computed at time $t + \delta/2$ from previous step:
$$x_i(t + \delta) = x_i(t) + \delta\frac{\partial K}{\partial p_i}(t + \delta/2);$$

3. Take the remaining half step in time to finish updating the momentum variable while fixing position variable at time $t + \delta$:
$$p_i(t + \delta) = p_i(t + \delta/2) - (\delta/2)\frac{\partial U}{\partial x_i}(t + \delta).$$

It's common to run Leap Fog method L steps forward to simulate the Hamiltonian dynamics over $L \times \delta$ units of time.

3.3.3 The target distribution

By developing a Hamiltonian function $H(x, p)$, the resulting Hamiltonian dynamics could be used to efficiently explore the target distribution $\pi(x)$.

For any energy function $E(\theta)$ over a set of variables θ, the corresponding *Gibbs' canonical distribution* can be defined as:

$$\pi(\theta) = \frac{1}{Z} e^{-E(\theta)} \tag{3.11}$$

where Z is the normalizing constant. The Hamiltonian is the sum of potential and kinetic energies:

$$E(\theta) = H(x, p) = U(x) + K(p) \tag{3.12}$$

The canonical distribution for the Hamiltonian dynamics energy function is defined as

$$\pi(x, p) \propto e^{-H(x,p)} = e^{-[U(x)-K(p)]} = e^{-U(x)} e^{-K(p)} \propto \pi(x)\pi(p) \tag{3.13}$$

The fact that the joint distribution for x and p factorizes implies that the canonical distribution $\pi(x)$ is independent of the analogous distribution for the momentum $\pi(p)$. By introducing momentum as *auxiliary variables*, the Markov chain path could be facilitated based on Hamiltonian dynamics; it would not be possible without momentum. Due to the independence of the canonical distributions for x and p, theoretically any distribution for sampling momentum variables could be used. A zero-mean Normal distribution with unit variance is a often good choice for the momentum variables.

$$K(p) \propto \frac{p^T p}{2} \tag{3.14}$$

3.3.4 Algorithm

Hamiltonian dynamics is mainly used as an efficient proposal function for a Markov Chain aiming to improve the efficiency when exploring the target probability density $\pi(x)$ defined by $U(x)$, compared with a (random-walk) Metropolis-Hasting proposal probability distribution.

By achieving so, HMC consists of the two alternating steps: one is a stochastic step that performs random transition between energy levels, the other is a deterministic step that performs leapfrog method along a given energy level followed by determining whether or not accept the proposal based on the Metropolis acceptance-rejection criterion.

Let (x_{t-1}, p_{t-1}) be the latest draw in the chain. To illustrate how the two alternative steps work, let's suppose we are at an initial state $(x_0, p_0) = (x_{t-1}, p_{t-1})$, numerical simulation using the Leap Frog methods is performed according to Hamiltonian dynamics for a short time, which leads to a new state denoted by (x^*, p^*) at the end of the simulation, and further used as the *proposal*. Due to inevitable discretization error, Metropolis acceptance criterion is used again here to determine whether or not the proposed state is accepted. Specifically if the probability of the proposed state after Hamiltonian dynamics

$$\pi(x^*, p^*) \propto e^{-[U(x^*)+K(p^*)]} \tag{3.15}$$

is greater than probability of the state prior to the Hamiltonian dynamics

$$\pi(x_0, p_0) \propto e^{-[U(x^{(t-1)}),K(p^{(t-1)})]} \tag{3.16}$$

then the proposed state is accepted; otherwise, the proposed state is accepted randomly with a computed ratio.

For a given set of initial conditions in x and p, the Hamiltonian dynamics actually follows contours of constant energy in phase space. Randomly perturb to the dynamics is needed to explore all of $\pi(x)$, which can be achieved by simply drawing a random momentum from the corresponding canonical distribution $\pi(p)$ before running the dynamics prior to each sampling iteration.

```
Start with a collection of samples $\{x^{(1)}, \cdots, x^{(s)}\}$ where $x^{(s)}$ being the latest draw;
Let $x_0$ be $x^{(s)}$;
Sample a new initial momentum variable $p_0$ from $pi(p)$;
Run *Leap Frog algorithm* starting at $[x_0, p_0]$ for $L$ steps with step-size $\delta$ to obtain
    proposed states $x^*$ and $p^*$ ;
Compute the acceptance ratio:
    $r = \exp\left(-U\left(x^*\right) + U\left(x_0\right) - K\left(p^*\right) + K\left(p_0\right)\right)$
Accept $\theta^* := (\alpha^*, \beta^*)$ with following acceptance-rejection criterion:
    $\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{(s)} & \text{with probability } 1 - \min(r, 1) \end{cases}$
```

Algorithm 2: Hamiltonian Monte Carlo

CHAPTER 4
Transition Density Approximations

In previous chapter, we reviewed MCMC as a means to sample intractable posterior distribution in Bayesian inference, and we also showed posterior distribution requires an explicit likelihood function. However the transition density of cusp dynamics, hence the likelihood function is analytically intractable, approximation to the transition density is therefore needed. In this chapter, we review two different approaches to approximate the transition density of cusp SDE.

4.1 Motivation

Consider a diffusion process X_t governed by

$$dX_t = \mu\left(X_t;\theta\right)dt + \sigma\left(X_t;\theta\right)dW_t, \tag{4.1}$$

where μ and σ are known functions that both might depend on a vector of model parameters θ.

Let $p_X(\Delta, x|x_0;\theta)$ be the conditional density of $X_{t+\Delta} = x$ given $X_t = x_0$ induced by the model Equation 4.4, the transition probability density. Further assume data points are observed at discrete time points $t_i = i\Delta^{obs}$ with $i = 0,\cdots,n$, and observed measurements being $x = (x_0,\cdots,x_n)$. Due to Markovian property, the log-likelihood function has the simple form

$$\ell_n(\theta) \equiv \sum_{i=1}^{n} \ln\left\{p_X\left(\Delta, X_{i\Delta}|X_{(i-1)\Delta};\theta\right)\right\} \tag{4.2}$$

4.2 Closed-form approximation using Hermite polynomials

Ait-Sahalia [2002] develops two approaches to construct a sequence of closed-form expansions for the log-likelihood function that approximate the intractable (log) transition density of a diffusion process. One is based on finding the coefficients of a Hermite expansion for the transition density; the other takes the form of the Hermite series first, and computes its coefficients by solving the Fokker-Planck equation which characterize the transition function. The two approaches give the same final expression [Ait-Sahalia et al., 2008].

4.2.1 Hermite polynomials

The modified Hermite polynomial with degree n is denoted by

$$H_n(z) = e^{z^2/2}\frac{d^n}{dz^n}\left(e^{-z^2/2}\right), \quad n \geq 0. \tag{4.3}$$

If we let $\phi(z)$ be the pdf of the standard normal distribution, H_n has the property

$$\int_{-\infty}^{\infty} \phi(z)H_n(z)H_m(z)dz = \begin{cases} 0 & n \neq m \\ n! & n = m \end{cases}$$

4.2.2 Derivation

Consider a continuous-time parametric diffusion

$$dX_t = \mu\left(X_t;\theta\right)dt + \sigma\left(X_t;\theta\right)dW_t, \tag{4.4}$$

and the goal is to find a closed-from approximation to the transition density $p_X(\Delta, x|x_0;\theta)$.

First transformation

First, applying the Lamperti transformation $\gamma(\cdot)$ to X_t gives us a transformed new process Y_t such that

$$Y_t = \int^{X_t} \frac{du}{\sigma(u;\theta)}. \tag{4.5}$$

By Ito's lemma, the transformed process Y_t satisfies the following SDE

$$dY_t = \widehat{\mu}(Y_t;\theta)dt + dW_t, \tag{4.6}$$

with drift term being

$$\widehat{\mu}(Y_t;\theta) = \frac{\mu(X;\theta)}{\sigma(X_t;\theta)} - \frac{1}{2}\frac{\partial\sigma(X_t;\theta)}{\partial x},$$

and unit diffusion term.

Second transformation

Let Y_t be the value of Y corresponding to X_t, and we further transform Y_t by normalizing it in

$$Z = \frac{Y - Y_t}{\sqrt{\Delta}}. \tag{4.7}$$

Intuitively, the transition density of $p(Z_{t+\Delta}|Z_t = z_t)$ is well, or at least better approximated by Gaussian with mean $\widehat{\mu}\left(Y_k;\theta\right)\sqrt{\Delta}$ and unit variance. Hence, it suggests using Hermite series expansion to approximate the transition density $f(z,t)$

$$f(z,t) = \phi(z)\sum_{n=0}^{\infty}\eta_n(z,t)H_n(z), \tag{4.8}$$

where $\phi(z)$ is the probability density function of standard normal distribution.

Hermite series expansion

To approximate the transition density, Ait-Sahali proposed to take the form of the Hermite series and determines its coefficients by solving the Fokker-Planck equation which characterize the transition function. In particular, first rewrite Equation 4.8 as

$$f(y,t) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(y-Y_k)^2}{2t}\right) \times \psi(y,t) \exp\left(\int_{Y_k}^{y} \widehat{\mu}(u)du\right). \tag{4.9}$$

The right hand side of equation Equation 4.9 is the product of $\phi(z)$ expressed in terms of Y and two remaining terms that plays the role of the infinite Hermite sum in Equation 4.8.

The goal is to express $\psi(y,t)$ in terms of a convergent power series in t, where the coefficients of the series are expected to capture the contribution made by the entire family of Hermite polynomials at each order in t. In particular a desired ψ is in the form of a power series expansion in t,

$$\psi(y,t) = \sum_{n=0}^{\infty} \frac{c_n(y)t^n}{n!}, \tag{4.10}$$

and coefficients are to be determined.

To achieve so, the unit diffusion process Y_t is known to satisfy the Fokker-Planck equation, which gives

$$\frac{\partial f}{\partial t} = -\widehat{\mu}\frac{\partial f}{\partial y} - f\frac{\partial \widehat{\mu}}{\partial y} + \frac{1}{2}\frac{\partial^2 f}{\partial y^2}. \tag{4.11}$$

Solution of equation Equation 4.11 may be represented by expression Equation 4.9 provided $\psi(y,t)$ satisfies

$$\frac{\partial \psi}{\partial t} = \frac{1}{2}\frac{\partial^2 \psi}{\partial y^2} - \frac{y-Y_k}{t}\frac{\partial \psi}{\partial y} + \lambda\psi, \quad \lambda = -\frac{1}{2}\left(\widehat{\mu}^2 + \frac{\partial \widehat{\mu}}{\partial y}\right). \tag{4.12}$$

Coefficient functions $c_0(y), c_1(y), \cdots$ can be determined by matching the coefficients of powers of t. In particular, by substituting Equation 4.11 into Equation 4.12 leads to

$$\sum_{n=1}^{\infty} \frac{c_n(y)t^{n-1}}{(n-1)!} = \frac{1}{2}\sum_{n=0}^{\infty} \frac{d^2 c_n(y)}{dy^2}\frac{t^n}{n!} - \frac{y-Y_k}{t}\sum_{n=0}^{\infty} \frac{dc_n(y)}{dy}\frac{t^n}{n!} + \sum_{n=0}^{\infty} \frac{\lambda(y)c_n(y)t^n}{n!}. \tag{4.13}$$

If we re-index summation and re-arrange some of the terms in Equation 4.13, we obtain

$$\sum_{n=0}^{\infty} \frac{t^n}{n!}\left[c_{n+1}(y) + \frac{(y-Y_k)}{n+1}\frac{dc_{n+1}(y)}{dy} - \left(\frac{1}{2}\frac{d^2 c_n(y)}{dy^2} + \lambda(y)c_n(y)\right)\right] = -\frac{y-Y_k}{t}\frac{dc_0(y)}{dy}, \tag{4.14}$$

from which it follows immediately that

$$\frac{dc_0(y)}{dy} = 0 \tag{4.15}$$

$$c_{n+1}(y) + \frac{(y-Y_k)}{(n+1)}\frac{dc_{n+1}(y)}{dy} = \frac{1}{2}\frac{d^2 c_n(y)}{dy^2} + \lambda(y)c_n(y). \tag{4.16}$$

The first condition (equation 4.15) implies that $c_0(y)$ is a constant in y; furthermore, this constant function must be $c_0(y) = 1$ in order to maintain the correctness of short time asymptotic expression for the transition density.

The condition that $\psi(y,t)$ being finite at $y = Y_k$ for all $t > 0$ requires the solution of equation 4.16 to be

$$c_{n+1}(y) = \frac{n+1}{(y-Y_k)^{n+1}}\int_{Y_k}^{y} (u-Y_k)^n \left(\frac{1}{2}\frac{d^2 c_n(u)}{du^2} + \lambda(u)c_n(u)\right) du, \quad n \geq 0 \tag{4.17}$$

More on the recursive formula

By adapting the compact notation given in Ait-Sahalia et al. [2008], the approximation to the log-transition density for Y_t up to order K has the form:

$$\ln p_Y^{(K)}(y|y_0) = -\frac{1}{2}\ln(2\pi\Delta) - \frac{1}{2}\ln\left(\sigma^2(x)\right) + \frac{C_Y^{(-1)}(y|y_0)}{\Delta} + \sum_{k=0}^{K} C_Y^{(k)}(y|y_0)\frac{\Delta^k}{k!}. \tag{4.18}$$

The closed-form expansion can be obtained by the following recursive relations:

$$\begin{aligned} C_Y^{(-1)}(y|y_0) &= -\frac{1}{2}\left(y-y_0\right)^2 \\ C_Y^{(0)}(y|y_0) &= (y-y_0)\int_0^1 \mu_Y\left(y_0+u\left(y-y_0\right)\right)du. \end{aligned} \tag{4.19}$$

For $k \geq 1$,

$$C_Y^{(k)}(y|y_0) = k\int_0^1 G_Y^{(k)}\left(y_0+u\left(y-y_0\right)|y_0\right)u^{k-1}du. \tag{4.20}$$

And the functions $G_Y^{(k)}$ are given by

$$G_Y^{(1)}(y|y_0) = -\frac{\partial \mu_Y(y)}{\partial y} - \mu_Y(y)\frac{\partial C_Y^{(0)}(y|y_0)}{\partial y} + \frac{1}{2}\frac{\partial^2 C_Y^{(0)}(y|y_0)}{\partial y^2} + \frac{1}{2}\left(\frac{\partial C_Y^{(0)}(y|y_0)}{\partial y}\right)^2. \tag{4.21}$$

For $k \geq 2$

$$\begin{aligned} G_Y^{(k)}(y|y_0) = &- \mu_Y(y)\frac{\partial C_Y^{(k-1)}(y|y_0)}{\partial y} + \frac{1}{2}\frac{\partial^2 C_Y^{(k-1)}(y|y_0)}{\partial y^2} \\ &+ \frac{1}{2}\sum_{h=0}^{k-1}\left(\begin{array}{c} k-1 \\ h \end{array}\right)\frac{\partial C_Y^{(h)}(y|y_0)}{\partial y}\frac{\partial C_Y^{(k-1-h)}(y|y_0)}{\partial y} \end{aligned} \tag{4.22}$$

The desired approximation can be determined recursively. Mathematica or some other symbolic tool packages could be used to obtain the coefficients precisely and without any error.

4.2.3 Asymptotic properties

The log-likelihood function $\ell_n(\theta)$ is therefore approximated by

$$\ell^{(K)}(\theta) = \sum_{i=1}^{n} \ln\{p_X^{(K)}\left(\Delta, X_{i\Delta}|X_{(i-1)\Delta;\theta}\right)\},$$

and the approximate maximum likelihood estimator is defined as

$$\widehat{\theta}_K = \arg\max_{\theta} \ell_n^{(K)}(\theta).$$

Moreover, it has been proved that if K tends to infinity [Ait-Sahalia, 1996], [Ait-Sahalia et al., 2008], then

$$\widehat{\theta}_K \to \widehat{\theta}_{MLE}.$$

4.3 Euler approximation

Instead of approximating the transition density directly, a different route to obtain some workable likelihood function is to approximate or to discretize the path of the process. The approach is commonly known as the pseudo-likelihood approximation or locally Gaussian approximation.

To better illustrate the idea behind this approach, we again start with a diffusion process governed by the general SDE

$$dX_t = \mu\left(X_t;\theta\right) dt + \sigma\left(X_t;\theta\right) dW_t. \tag{4.23}$$

We may assume the coefficients μ and σ of the above SDE remain constant over time intervals $[t, t+\Delta)$, then Euler scheme would give the discretization

$$X_{t+\Delta} - X_t = \mu\left(X_t;\theta\right)\Delta + \sigma\left(X_t;\theta\right)\left(W_{t+\Delta} - W_t\right), \tag{4.24}$$

with $X_{t+\Delta} - X_t$ being Gaussian with mean $\mu\left(X_t;\theta\right)\Delta$ and standard deviation $\sigma\left(X_t;\theta\right)\sqrt{\Delta}$.

The approximated transition density of the process therefore can be written as

$$p_\theta(t, x_t|x) = \frac{1}{\sqrt{2\pi t\sigma^2(x;\theta)}} \exp\left\{-\frac{1}{2}\frac{(x_t - x - \mu(x;\theta)t)^2}{t\sigma^2(x,\theta)}\right\}. \tag{4.25}$$

Definition 4. We say the drift term of a SDE satisfy the polynomial growth condition. if there exist $L > 0$ and $m > 0$ (independent of θ such that

$$|\mu(x;\theta)| \leq L\left(1 + |x|^m\right), \quad \theta \in \Theta.$$

By assuming the polynomial growth condition, the log-likelihood of the discretized process is

$$\ell_n(\theta) = -\frac{1}{2}\left\{\sum_{i=1}^{n} \frac{\left(X_i - X_{i-1} - b\left(X_{i-1},\theta\right)\Delta\right)^2}{\sigma^2\Delta} + n\log\left(2\pi\sigma^2\Delta\right)\right\}. \tag{4.26}$$

Equation 4.26 is commonly referred as the Euler approximation. This approximation often works well when discretization Δ is small; however in other case that Δ being not small enough, bias maybe introduced and considerable [Iacus, 2009].

Florens-Zmirou [1989] claimed that pseudo-likelihood estimators based methods including Euler approximation are inconsistent for fixed Δ. In practice however, this inconsistency does not discourage researchers from using it, and in fact this (locally Gaussian approximation) is still a very convenient choice due to its simplicity and computational efficient, especially for situations where the Δ is small hence small bias [Durham and Gallant, 2002]. For others, this estimator is often used as initial values for other more complete parameter estimation methods even when the bias introduced by the discretization cannot be disregarded [Jimenez et al., 2005].

4.4 Cusp transition density approximations

Given discretely sampled observation $\{x_0, x_1, \cdots, x_n\}$, the likelihood function is in the form

$$L_n(\theta) = \prod_{i=1}^{n} p_\theta\left(\Delta, X_i|X_{i-1}\right) p_\theta\left(X_0\right), \tag{4.27}$$

with log-likelihood function being

$$\begin{aligned}\ell n(\theta) = \log L_n(\theta) &= \sum_{i=1}^{n} \log p_\theta\left(\Delta, X_i|X_{i-1}\right) + \log\left(p_\theta\left(X_0\right)\right) \\ &= \sum_{i=1}^{n} l_i(\theta) + \log\left(p_\theta\left(X_0\right)\right).\end{aligned} \tag{4.28}$$

In order for us to perform Bayesian inference (or MLE), we need an explicit or closed-form transition density that is workable. In this section, we introduced two different approaches to obtain some workable likelihood function, namely the closed-form approximation by Hermite polynomials by Ait-Sahalia (denoted by HPE) and the locally Gaussian approximation (denoted by Euler). In this section, we compare two by examining their plots with different discretization step-size Δ. The explicit approximation using Hermite polynomials by Ait-Sahalia is given in **??**.

We fix parameter values to be $\alpha = 1$ and $\beta = 3$, hence $\Delta^{\text{disc}} < 0$. In this case, the cusp stationary density is known to be bimodal.

Euler differs slightly to HPE when Δ_t is small, suggested by Figure 4.1 and Figure 4.2. However, a considerable difference appears when Δ_t is large, for example shown in Figure 4.3. More importantly and perhaps more interestingly, with large Δ_t, approximation by FHE is actually able to capture and retain its distinctive bimodality feature associated to cusp stationary distribution model.

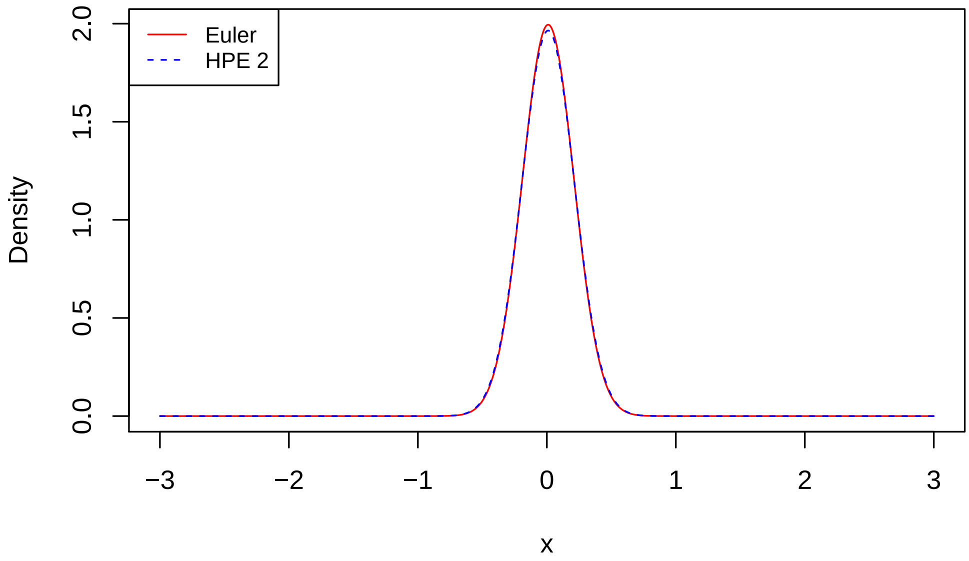


Figure 4.1: Comparison of Euler and HPE: $\alpha = 1, \beta = 3, x_0 = 0$ and $\Delta = 0.01$

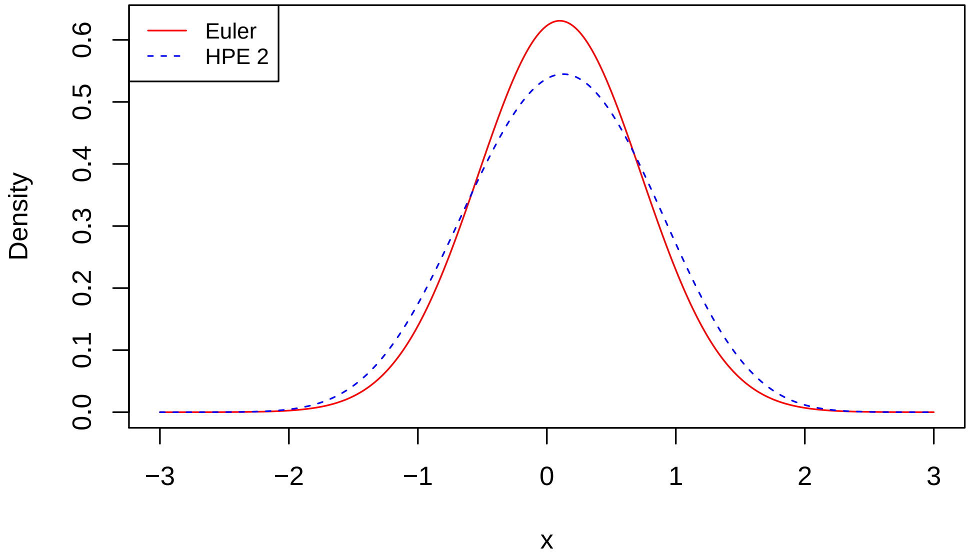


Figure 4.2: Comparison of Euler and HPE: $\alpha = 1, \beta = 3, x_0 = 0$ and $\Delta = 0.10$

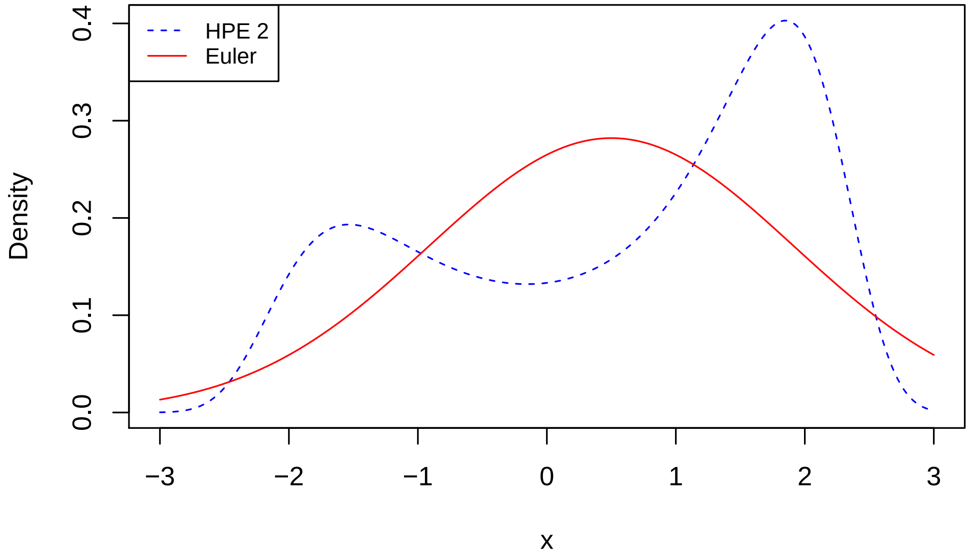


Figure 4.3: Comparison of Euler and HPE: $\alpha = 1, \beta = 3, x_0 = 0$ and $\Delta = 0.50$

CHAPTER 5
MCMC Convergence Diagnostics

In chapter 3, we reviewed the basics on Bayesian inference and MCMC as means to sample intractable posterior distribution. In particular, MCMC is constructed in a way such that the desired or target distribution is its equilibrium distribution. Regardless of its actual result, we first must make sure the samples drawn by a MCMC indeed represent the actual posterior distributions. In this chapter, we review MCMC convergence diagnostics from a practitioner's perspective.

5.1 Motivation

That being discussed, samples obtained by running an MCMC is expected to be a good representation of the true but intractable target distribution, which further requires the obtain samples to be coming from its equilibrium distribution.

MCMC simulation often starts at some arbitrary point in the parameter space. Inevitably, the arbitrary starting point may not be even close to the actual high probability region of the posterior distribution, due to the lack of prior information. Consequently, in most cases samples drawn from the early stage in fact has not enter the stationary phase yet, hence not a good representation of the target distribution. From a more global-optimization perspective, samples from MCMC chain in the early stages of the MCMC runs, particularly those with "far" starting point are unlikely to occur in samples from the true distribution. The early stage of a MCMC are in fact commonly referred as the "transient" phase, in contrast to those desired ones drawn from "stationary" phase .

If one were able to correctly identify and separate samples from the transient phase with those coming from stationary phase as desired, we can therefore confidently claim samples that are good representations of the posterior distribution by discarding the former. However, it is definitely not trivial to determine when the transient phase ends in a chain. Furthermore there is no such universal cut-off value and the answer varies from problem to problem. In the following sections, we review several commonly used MCMC diagnostic criterion, some of them can be used to help identify the transient phase, while others help us assess the convergence of MCMC in general.

Preparation

In the next chapter, chapter 6, we are going to perform a series of intensive simulation studies. In particular we would like to sample posterior distributions of model parameters α and β given data points $x = \{x_1, \cdots, x_{1200}\}$ assumed to be generated according to cusp SDE. We use Hamiltonian Monte Carlo to sample the posteriors. We take warm-up or burn-in period to be the first 200 runs, and we run HMC a total number of 22,000 runs, which makes the size of kept-samples being 2,000.

By using various MCMC convergence diagnostics, we would like to show this kept ones are indeed drawing from the desired equilibrium distribution hence a good representation of the true posterior distributions.

5.2 Trace plot

Trace plot is perhaps always a must in a MCMC diagnostic. In particular, trace plots are used to assess the quality of mixing of a chain by straightforward visualization. Beside, running multiple chains from different starting points, and then assess their trace plots to see whether they converge to the sample plot are even more helpful in practice.

In our example, the two chains have been run for each parameter. We do trace plot for the kept ones. Trace plots of both parameters clearly show the “caterpillar-like” behavior such as shown in Figure 5.1, it is in fact a good indication that the MCMC is efficiently sampling from its equilibrium distribution.

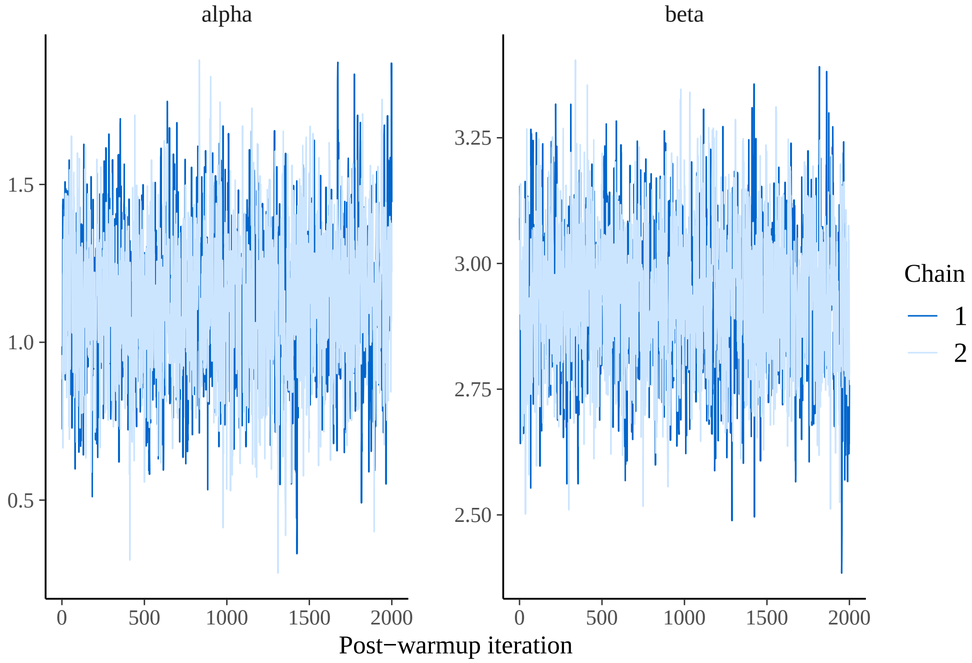


Figure 5.1: MCMC diagnostics: Trace plot

5.3 Autocorrelation plot

Another “must” diagnostic for MCMC convergence is by the autocorrelation plot. Here autocorrelation refers to the correlation between the samples drawn by MCMC. In particular, a lag-k autocorrelation is defined to be the correlation between every sample in the chain and the sample k steps before.

A converging chain is more likely to be seen accompanied by a decreasing autocorrelation when k is getting larger. Samples with smaller autocorrelation can be regarded as being more independent. On the other hand, a high degree of correlation is often accompanied by high autocorrelation values, especially with large k values, which further suggest slow or inefficient mixing. Some common explanation includes the chain being stuck in a local maximum therefore needs more MCMC runs to leave the local maximum in order to continue searching other parts of the parameter space.

Our autocorrelation plot Figure 5.2 shows an efficient hence desired mixing, which in turn supports the efficiency of Hamiltonian Monte Carlo.

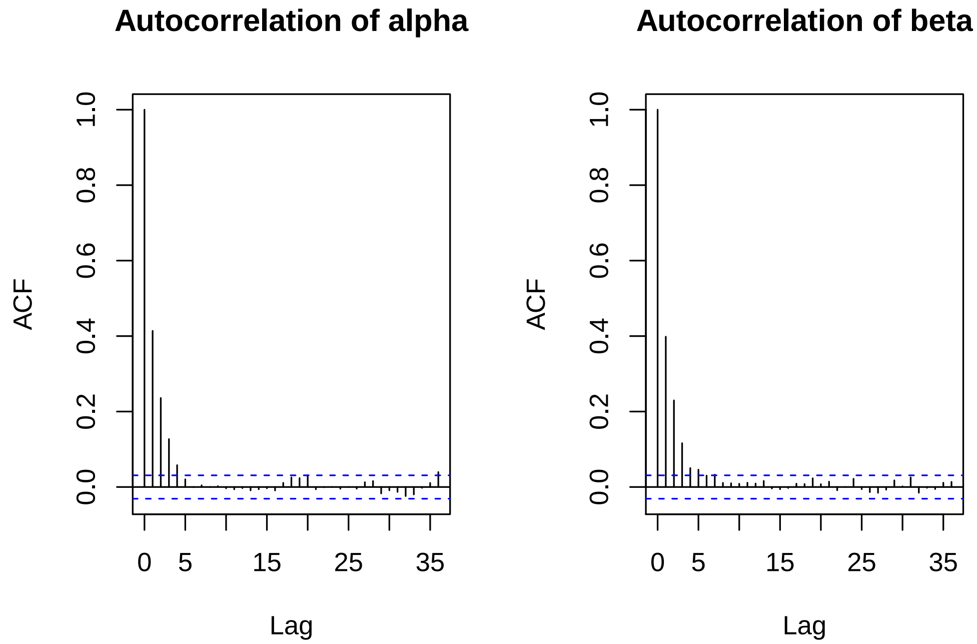


Figure 5.2: MCMC diagnostics: Autocorrelaiton plot

5.4 Effective sample size

Both trace plots and autocorrelation plots are visualizaiton techniques used to assess the convergence of MCMC chain. By purely look at plots, it's very straightforward and quick to spot if there's anything undesired happened.

Perhaps a better and more accurate estimate for identifying transient phase is by the effective sample size (ESS). As its name suggests, effective sample size measures the number of independent samples or information contained in an autocorrelated samples.

To see how it works, let's say we have a chain obtained by running an MCMC that includes all the draws starting from the very first starting point. Intuitively, samples in transient phase are often not very informative (due to the arbitrarily selected starting point that potentially far from the high probability region). Consequently,

1. if the burn-in period (up to the last sample of the transient phase) were under-estimated, it would reduce ESS because non-informative or noise samples are mixed with those desired ones, hence reduce the overall ESS size. By keeping removing those sample at transient phase, ESS size is expected to raise.

2. On the other hand, if the burn-in period is over-estimated instead, desired informative samples are in fact being thrown away. This would again reduce the ESS because informative samples are being discarded.

Therefore, the optimal estimate of the (end-of) transient phase would be the one that ESS is maximized.

From Figure 5.3, it clearly shows the ESS plots for both parameters are monotone decreasing, in this case, it indicates no transient phase in the chain; or in other words, the plots suggests the chain is drawn from its stationary phase.

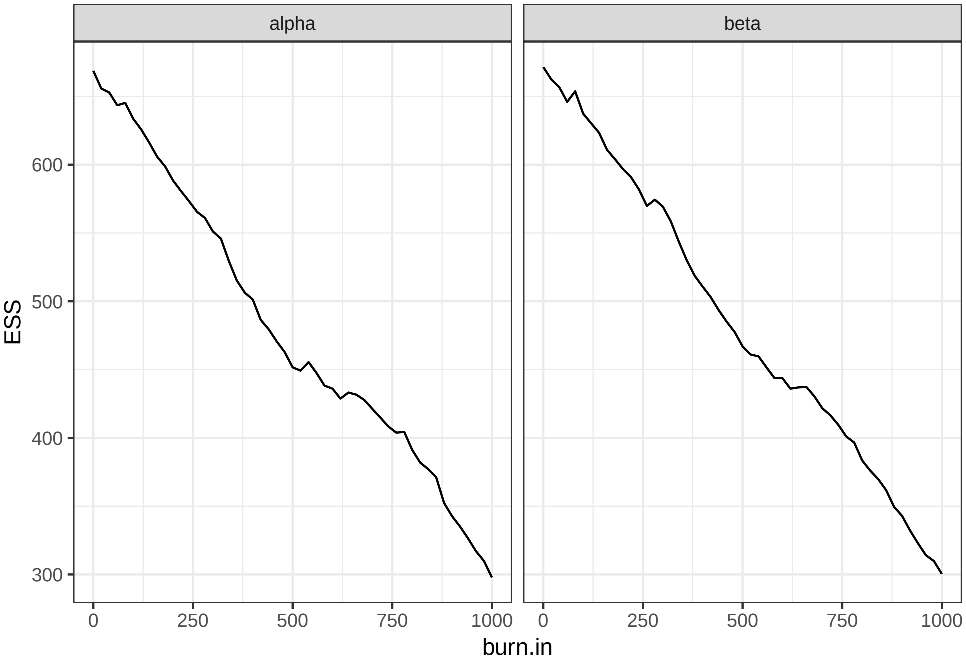


Figure 5.3: MCMC diagnostics: Effective sample size plot

5.5 Geweke

Geweke et al. [1991] contributed to the MCMC diagnostic community by proposing a single statistic test for identifying the transient phase.

The basic idea behind this test is that, if samples obtained by running MCMC were indeed drawn from desired equilibrium distribution, and if we split the chain into three parts, then mean of the first part are expected to be equal to the mean of the last part because the whole chain of samples are assumed to be drawn from the same equilibrium distribution.

Theory

To better illustrate how this works, let's assume we now have a long enough chain whose trace plot already suggests convergence to the target distribution. Since the convergence is supported by the trace plots, we simply assume the second half of the chain has converged to the equilibrium distribution. We are now interested in testing if the first 10% (can be 20%, or 30% etc., and it depends on the hypothesis) of the chain we constructed could be be identified as transient phase.

To do so, the diagnostic mimics the simple two□sample test of means. Two-sample X_1 and X_2 T test of equality of mean with unequal variance is computed as

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}. \tag{5.1}$$

We then perform hypothesis test with null hypothesis being the mean of the first 10% equals to the mean of the last 50%. If the result of hypothesis is statistically significant, we reject null hypothesis. It can then be perceived as the first 10% are in transient phase.

Result

We utilize the coda package in R, in particular the function geweke.diag for implementing Geweke diagnostics. This plot (Figure 5.4) describes Geweke's Z-scores when successively larger numbers of iterations are discarded from the beginning of the chain.

The first half of the Markov chain, in our cases the first $1,000$ runs, is divided into number of segments. Geweke's Z-score is then repeatedly computed with the first Z-score being calculated with all iterations in the chain, the second after discarding the first segment, the third after discarding the first two segments, and so on. We only use the samples in the second half chain to compute the very last Z-score.

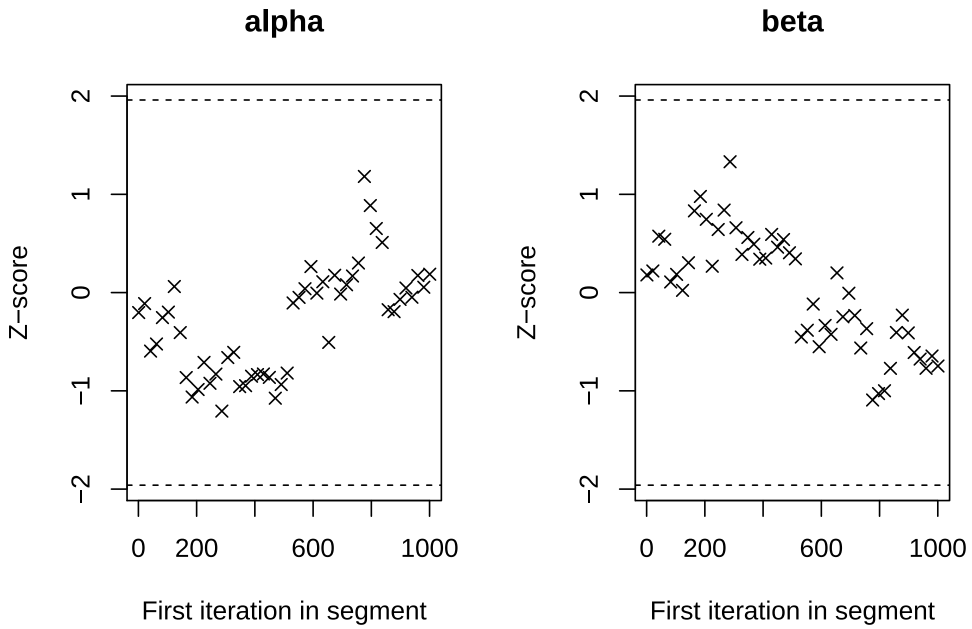


Figure 5.4: MCMC diagnostics: Geweke plot

5.6 Gelman-Rubin

Gelman and Rubin [Gelman et al., 1992] are two major contributors to the MCMC convergence diagnostic community. They proposed a profoundly useful and practical but also general approach to assess the convergence of MCMC where multiple parallel chains using different and arbitrary starting values are used.

The idea behind Gelman-Rubin diagnostics is intuitive and straightforward: convergence (to the stationary distribution) is often obtained when all different chains have passed their transient phase and the outputs from different chains are somehow “indistinguishable”.

Theory

The Gelman-Rubin diagnostic is based a comparison of within-chain and between-chain variances that in spirit similar to a analysis of variance (ANOVA) in statistical analysis.

To better illustrate the idea, let $\{x_{i,1}, \cdots, x_{i,N}\}$ be the ith Markov chain sampled, and suppose there are in total M independent chains sampled. Let $\overline{x}_{i.}$ be the mean from the ith chain, and $\overline{x}_{..}$ be the overall mean. Let W be the empirical mean of the variance of with-in chain with M independent chains in total, that is

$$W = \frac{1}{M}\sum_{m=1}^{M} s_m^2, \tag{5.2}$$

where

$$s_m^2 = \frac{1}{N-1}\sum_{t=1}^{N}\left(\overline{X}_{mt} - \overline{X}_{m\cdot}\right)^2, \tag{5.3}$$

where s_m^2 measures the with-in chain variance for the mth chain.

Let B be the variance between chain, that is

$$B = \frac{N}{M-1}\sum_{m=1}^{M}\left(\overline{X}_{m\cdot} - \overline{X}_{\cdot\cdot}\right)^2. \tag{5.4}$$

Now define a new statistics $\hat{V}$ that combine both the with-in chain and between-chain variances:

$$\hat{V} = \left(\frac{N-1}{N}\right)W + \left(\frac{M+1}{MN}\right)B, \tag{5.5}$$

the desired statistic proposed by Gelman-Rubin, $\sqrt{\hat{R}}$, is computed as

$$\hat{R} = \frac{\hat{V}}{W}\cdot\frac{df+3}{df+1}, \tag{5.6}$$

where $df = 2\hat{V}/\operatorname{Var}(\hat{V})$.

In particular, $\hat{R}$ near or below 1 suggest convergence.

5.7 Conclusion

Diagnostics cannot *guarantee* the chain has converged, and in fact the purpose of diagnostics are to spot anything undesired (i.e. not convergent) [Hoff, 2009]. However, results from various MCMC diagnostics run in this section are all *supporting* the convergence of the kept chain.

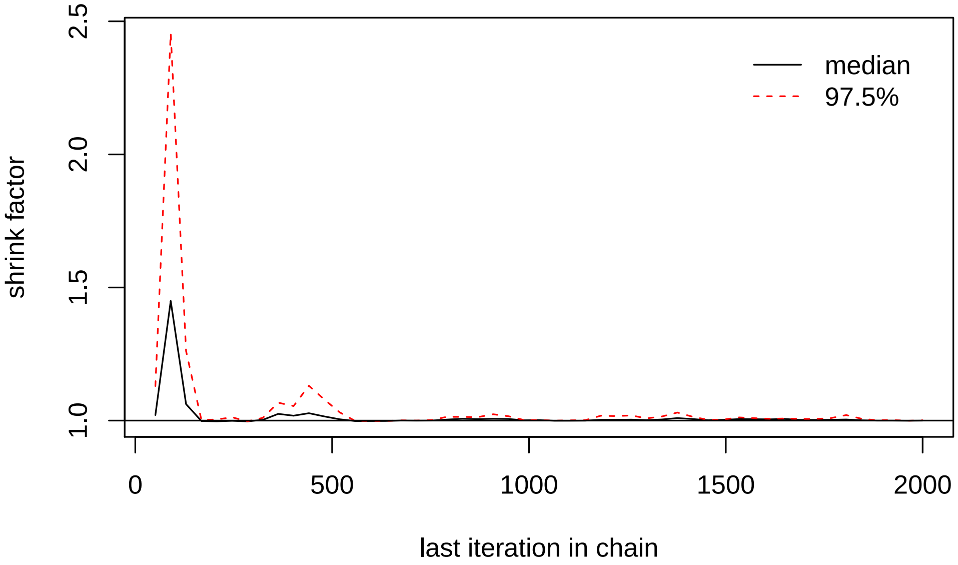


Figure 5.5: MCMC diagnostics: Gelman plot for α

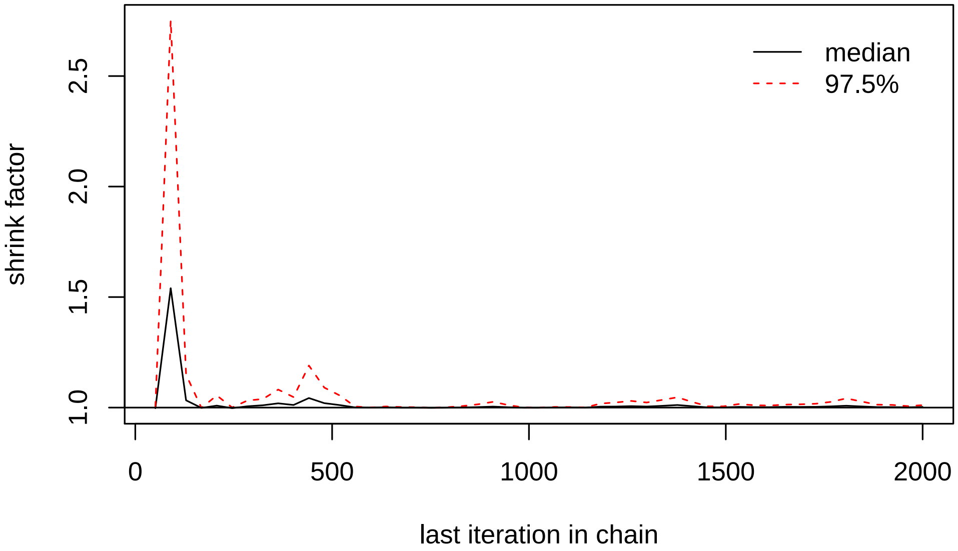


Figure 5.6: MCMC diagnostics: Gelman plot for β

CHAPTER 6
Inference from Complete Observations

The research problem of this thesis is to develop an accurate and computationally feasible parameter estimation algorithm based on Bayesian principle that can be implemented in absence of an exact transition distribution for cusp model using discretely sampled observations. Cloesd-form approximation using Hermite polynomials and Euler approximation are proposed to tackle the problem of intractable transition density, while Hamiltonian Monte Carlo is used to sample the posterior distribution. In this chapter, we carry out a series of simulation studies to verify the developed parameter estimation algorithm indeed works as desired under the complete observations scenario.

6.1 Model validation criterion

The most commonly used criterion for model validation and/or verification purpose is to generate simulated data with known model parameter values; the simulated data is then used as input of the developed parameter estimation algorithm to see whether or not the model can recover these parameters from the data. For example, one may use the maximum a posteriori (MAP) for point estimation, and Bayesian credible interval for interval estimation, etc. If the algorithm were not able to recover the pre-fixed parameters, it would be highly questionable to put the algorithm to work in analyzing real-world data [Jimenez et al., 2005], [Carpenter et al., 2017].

Procedures

The accuracy of the model is to be examined by a series of simulation studies designed as follow.

First and foremost, simulation experiment is designed to test the accuracy of the parameter estimation algorithm in *three* different case: they correspond to $\Delta^{\text{disc}} < 0$, $\Delta^{\text{disc}} = 0$, and $\Delta^{\text{disc}} > 0$. We knowingly choose parameters α, β to be $(1, 3)$, $(2, 3)$ and $(3, 3)$ to represent the three cases respectively. In all cases, we fix the diffusion coefficient to be constant $\sigma = 2$ and treat it as known instead of unknown model parameter. The primary goal of this section is to do parameter estimation on cusp model parameters α and β; however, the this can be extended easily to estimate more parameters when introduced. In the section of empirical study on foreign exchange rate, we do parameter estimation on generalized cusp SDE that consists 3 more parameters beside α and β.

For each of the three cases, after fixing parameter values, we use Euler's method with $\Delta = 0.1$ to perform trajectory simulation for T units of times, where $T = \{30, 60, 120\}$ so that each trajectory or *replication* contains $T \times 10$ time-series data points. We repeat this sample trajectory generating processes for $10,000$ times, so that eventually we will have $10,000$ replications in total, with each replication containing a length of $T \times 10$ time series data.

For each replication, we draw samples the posterior distribution $p(\theta | x_1, x_2 \cdots, x_{T\times 10})$ using Hamiltonian Monte Carlo. We let burn-in to be 200, and run the chain $2,200$ times in total, which makes the size of the kept draws $2,000$. The choice of burn-in and total number of MCMC runs were tuned and showed satisfying convergence result in chapter 5, hence being used here. In particular, in previous chapter, we showed the chain(s) have passed the transient phase of the constructed Markov chain, and therefore are good representation of the stationary hence the desired posterior distribution for both parameters.

Once the sample is obtained by performing Hamiltonian Monte Carlo, and its convergence is supported by different MCMC convergence diagnostic criteria, we are confident the kept samples in the chain are good empirical representation of the posterior distributions, we compute and record the following statistics for both α and β estimates:

- the mode of the posterior samples, denoted by α_{MAP} and β_{MAP};

- the standard error, denoted by SE α_{MAP} and SE β_{MAP};

- Bayesian 95-percent highest credible interval which is the tightest credible interval based on empirical posterior distributions;
- Whether or not the 95-percent highest credible interval successfully captured the true parameter values.

We repeat the above step for $10,000$ replications. Eventually we can obtain

- Two size $10,000$ vectors that hold all the α_{MAP} and β_{MAP} values, we use them as an empirical distribution of the estimator α_{MAP} whose size is $T \times 10$;
- a size $10,000$ vector that holds all the SEs;
- a scalar that shows total number of successfully captured replications (out of $10,000$). This is in fact the proportion of replications or sample trajectories where the (known) model parameter is captured in the confidence interval. This proportion will be used as an estimate for the *empirical coverage probability* for the constructed confidence interval.

The final output is summarized in the form of a table that contains 8 columns as follow

		α_{MAP}				β_{MAP}			
Par.	T	Mean	SE	Empirical SE	CP	Mean	SE	Empirical SE	CP

6.2 Simulation study and result

The simulation study results are shown in Table 6.1. Clearly the results supports the claim made on the parameter estimation algorithm. In particular, under all cases, mean values α_{MAP} and β_{MAP} are very close to the pre-selected parameter combinations, and this is a very important evidence that supports the accuracy of the algorithm. Meanwhile, the 95% CI is able to capture approximately 95% of the time in the total of $10,000$ replications.

Effect of T

Beside the accuracy of the algorithm has been verified, some interesting observation are worth pointing out. For example, naturally one would be interested in how the parameter estimation performance differs or perhaps improves by increasing the number of observations T. The simulation study can help to answer the question.

Intuitively, increasing the number of observations, hence more information would lead to a better estimation, in the sense of either reducing bias or variance, or both which is even better. This intuition is actually supported by the simulation study results. In particular, as first of all, more observations lead to reduced bias, where the SE is decreasing with an order of $\sqrt{n}$, where n is the number of observations.

The plots for all three testing cases, namely $\Delta < 0, \Delta > 0$, and $\Delta = 0$ clearly show that as more data points being collected, the posterior distributions become tighter and show higher peaks - which is consistent with our expectation. More data points can be viewed as more *information*, and a tighter posterior distribution implies more confident on the parameter estimations.

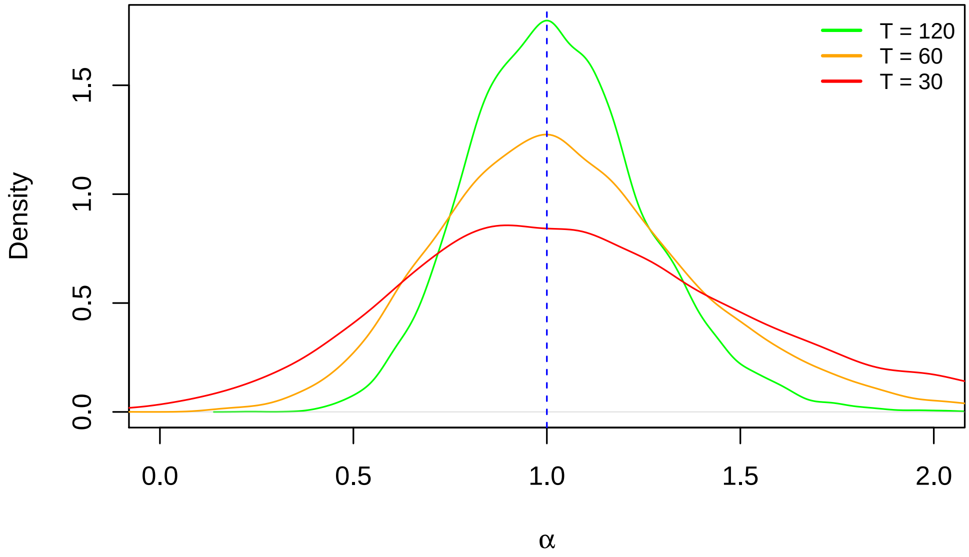


Figure 6.1: Complete observations: Empirical α_{MAP} with $\alpha = 1, \beta = 3$

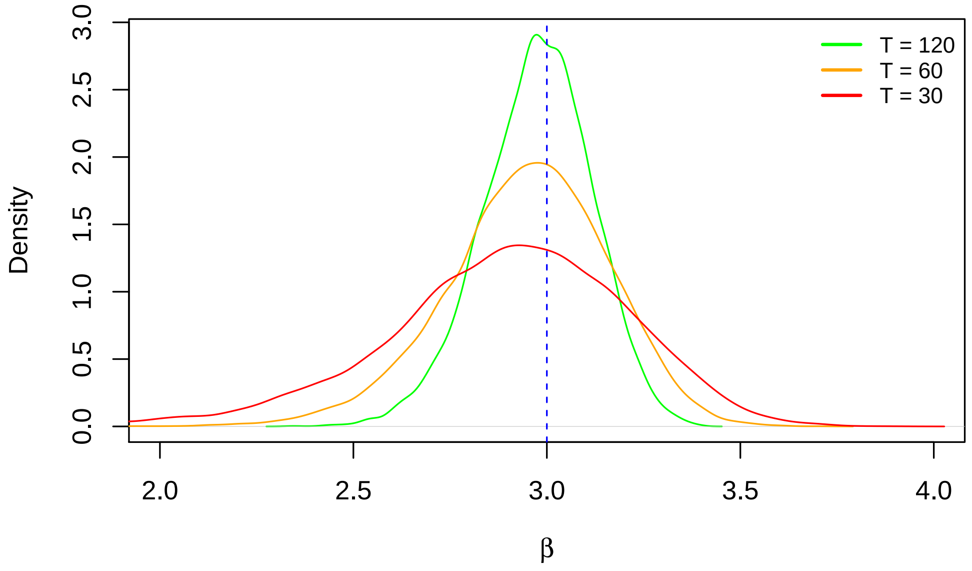


Figure 6.2: Complete observations: Empirical β_{MAP} with $\alpha = 1, \beta = 3$

6.3 Empirical example: USD/EUR Exchange Rate

In this section, we use an empirical example to show that cusp SDE performs “better” than the Vasicek model when both applied to the USD/EUR exchange rate example.

6.3.1 Model identification via AIC

The comparison criterion used here is the Akaike information criterion (AIC). AIC is constructed to find the best model embedded in a wider class of models [Iacus, 2009].

The idea behind is that AIC rewards models for high likelihood value while penalizing complexity. In particular, too many parameters makes the model over-specified hence less valuable and less favorable. So when picking a class of competing models, AIC criterion chooses the optimal one with minimum AIC criterion. Moreover, the true model is assumed to be among the competing ones.

In particular, the AIC statistic is defined as

$$\text{AIC} = -2\ell_n\left(\hat{\theta}_n^{(ML)}\right) + 2\dim(\Theta), \tag{6.1}$$

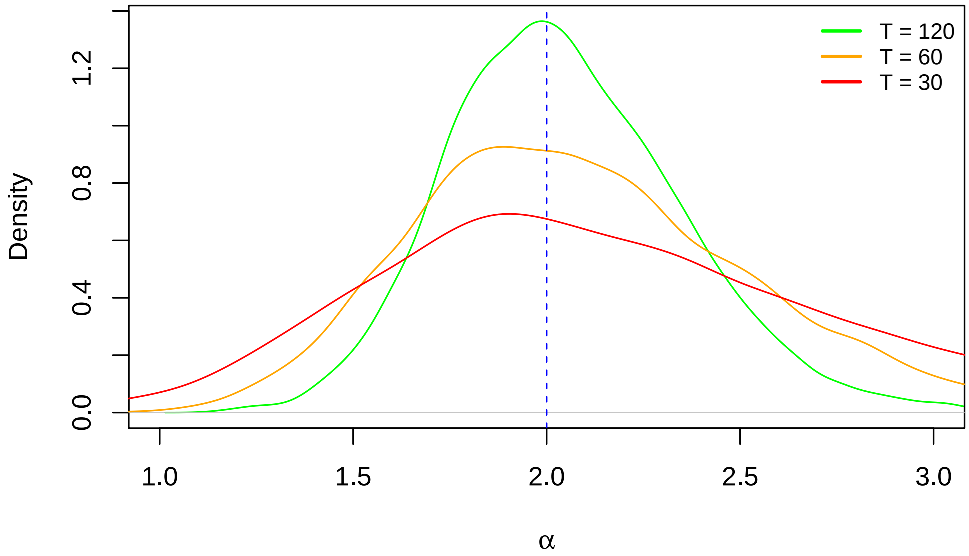


Figure 6.3: Complete observations: Empirical α_{MAP} with $\alpha = 2, \beta = 3$

where ideally $\hat{\theta}_n^{(ML)}$ will be the true maximum likelihood estimator. However, in most cases including our cusp SDE case, there is not analytic likelihood function, hence we use approximated likelihood instead.

6.3.2 Generalized cusp model

Cusp SDE (Equation 1.8) can be generalized by incorporating two additional parameters, namely the location parameter λ and scaling parameter r,

$$dX_t = r\left(\alpha + \beta\left(X_t - \lambda\right) - \left(X_t - \lambda\right)^3\right)dt + \sqrt{\epsilon}dW_t, \tag{6.2}$$

which is equivalent to the following expression

$$dX_t = \left(a + bX_t + cX_t^2 + dX_t^3\right)dt + fdW_t. \tag{6.3}$$

We now consider the famous Vasicek model:

$$dX_t = \theta\left(\mu - X_t\right)dt + \sigma dW_t \tag{6.4}$$

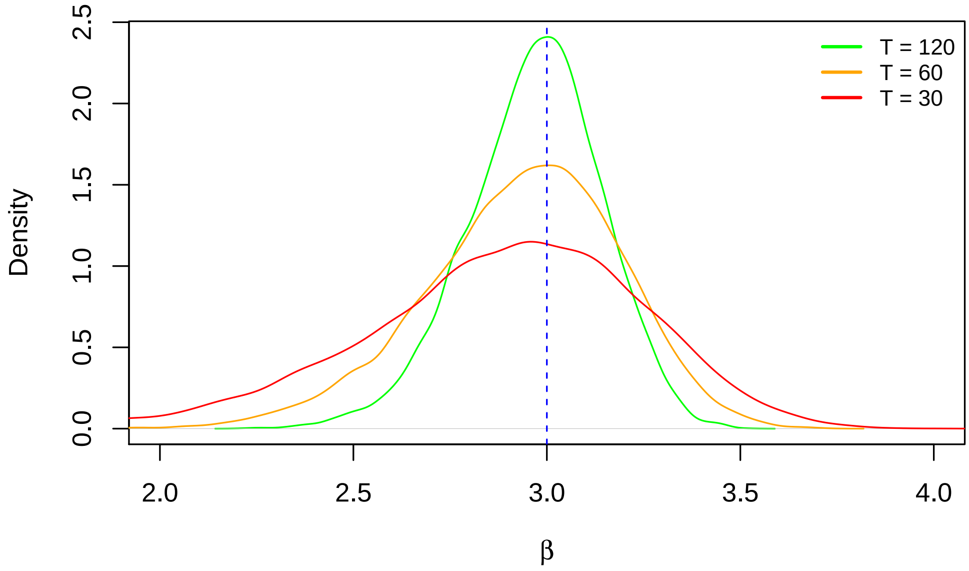


Figure 6.4: Complete observations: Empirical β_{MAP} with $\alpha = 2, \beta = 3$

as a competing model for the given empirical data.

6.3.3 USD/EUR exchange rate

We first define a re-scaling function that maps the actual empirical data to $[-2, 2]$, this can be achieved by following piece of R code:

```
ReScaling <- function(x){
4 * ((x-min(x))/(max(x)-min(x)) - 0.5)
}
```

Listing 6.1: User-defined function

We now apply the re-scaling function to the empirical data, and perform parameter estimation respectively by letting $\Delta^{\text{obs}} = 0.1$. Maximum likelihood parameter estimations hence AIC under Vasicek (Equation 6.4) can be computed in R:

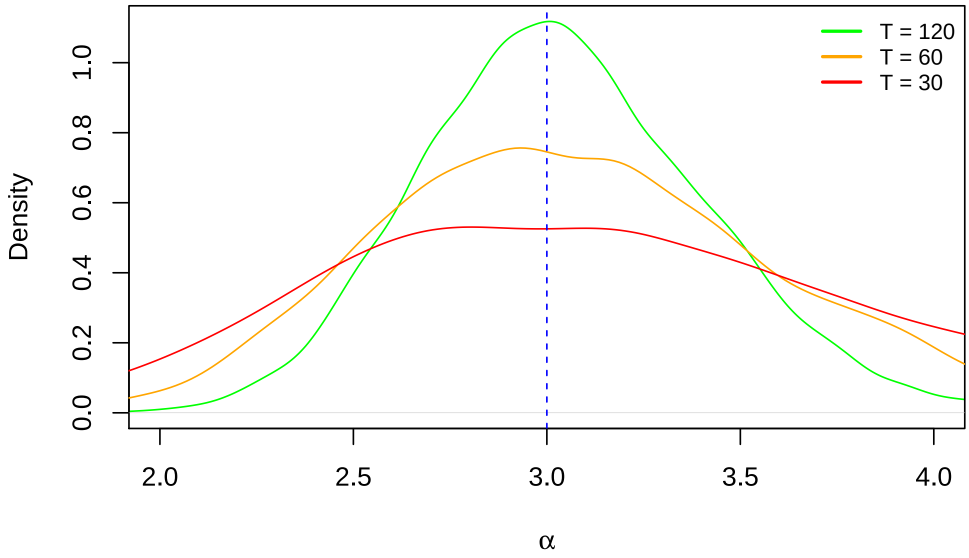


Figure 6.5: Complete observations: Empirical α_{MAP} with $\alpha = 3, \beta = 3$

```
## Maximum likelihood estimation
##
## Call:
## mle(minuslogl = OU.lik, start = list(theta1 = 1, theta2 = 0.5,
##     theta3 = 1), method = "BFGS")
##
## Coefficients:
##              Estimate   Std. Error
##  theta1  0.01344113   0.09090498
##  theta2  0.35880719   0.10449990
##  theta3 -0.71284861   0.02055915
##
## -2 log L: -110.1343
##
## AIC = -104.1343
##
```

Listing 6.2: AIC from Vasicek model

The corresponding estimations and AIC under generalized cusp model (Equation 6.3) can be computed:

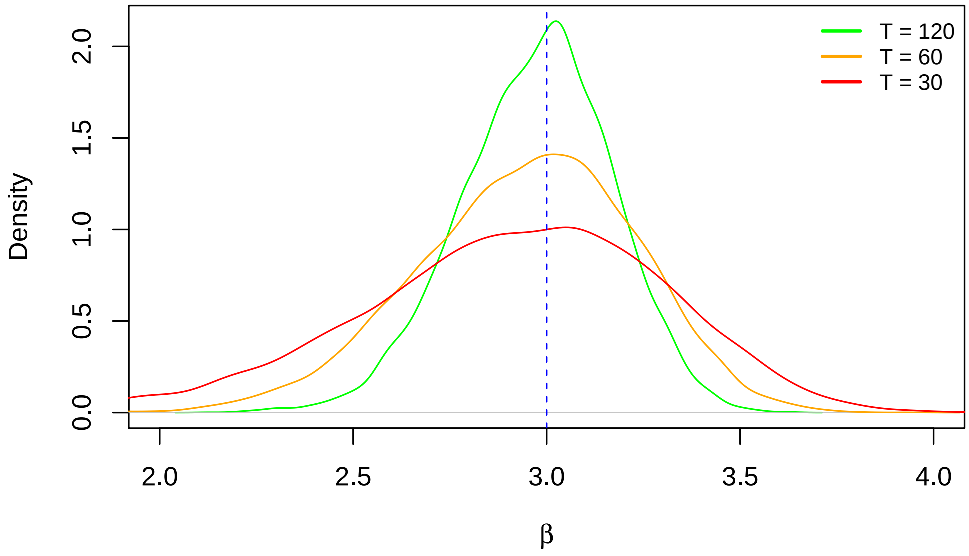
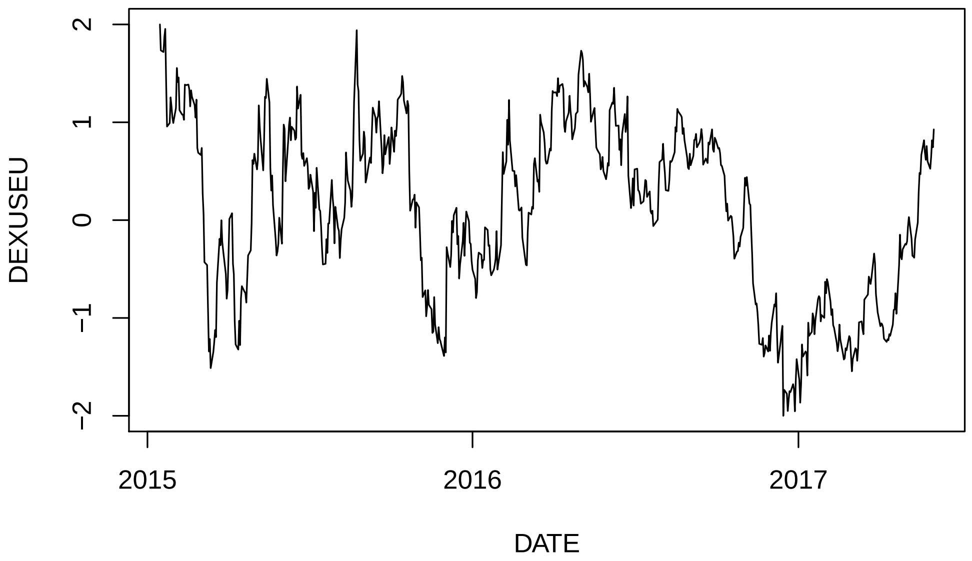


Figure 6.6: Complete observations: Empirical β_{MAP} with $\alpha = 3, \beta = 3$

```
## Maximum likelihood estimation
##
##
## Coefficients:
##        Estimate
##   a   0.07596694
##   b   0.1999819
##   c   -0.1546965
##   d   -0.3916001
##   f   0.7143831
##
## -2 log L: -121.6358
##
## AIC = -111.6358
##
```

Listing 6.3: AIC from Cusp model

Therefore, by AIC criterion, the generalized cusp SDE (Equation 6.3) performs better than the Vasicek model Equation 6.4 for the given data set.



Empirical study on FX: USD/EUR rescaled

		α_{MAP}				β_{MAP}			
Par. values	Num. days	Mean	SE	Empirical SE	CP	Mean	SE	Empirical SE	CP
$\alpha = 1,\ \ \beta = 3$	T = 30	1.131	0.536	0.457	0.930	2.891	0.320	0.288	0.932
	T = 60	1.061	0.343	0.313	0.940	2.948	0.212	0.197	0.933
	T = 120	1.031	0.230	0.218	0.943	2.975	0.141	0.137	0.944
$\alpha = 2,\ \ \beta = 3$	T = 30	2.197	0.650	0.612	0.945	2.881	0.367	0.351	0.942
	T = 60	2.102	0.445	0.612	0.945	2.937	0.253	0.242	0.942
	T = 120	2.050	0.304	0.295	0.945	2.968	0.173	0.169	0.947
$\alpha = 3,\ \ \beta = 3$	T = 30	3.175	0.753	0.751	0.948	2.91	0.398	0.399	0.950
	T = 60	3.094	0.533	0.526	0.947	2.948	0.282	0.279	0.949
	T = 120	3.047	0.372	0.370	0.948	2.973	0.195	0.196	0.951

Table 6.1: Simulation study: Inference from complete observations

For each setting (row), posterior analysis is performed based on 10,000 replications, each contains $T \times 10$ data points

CHAPTER 7

Inference from Partial Observations

In reality, complete observations may not be available all the time, and we call this the *partial observation* scenario. As the names suggest, partial observation scenario differs from complete observations scenario by admitting unobserved observations. One apparent difference is that two consecutive observed data points are often considerable sparse with partial observations. In cases where time interval Δ^{obs} between to consecutive observation points are large, the discrepancy between continuous model assumption and discretely observed data points should not be neglected, because otherwise it could lead to inconsistent estimators [Melino, 1996], [Jones, 1998], [Ait-Sahalia et al., 2008]. In this section, we run simulation studies to compare different methods and aim for improvement under the partial observations scenario.

7.1 Bayesian data augmentation

Under the partial observations scenario, data points are assumed to be observed more sparsely in time than the complete observations scenario studied in previous chapter. In particular, if we let Δ_{obs} denote the time difference between two consecutive observed data points, this Δ_{obs} would be considerably bigger than Δ_{obs} that corresponds to the complete observation scenario. We further let $n = n_{obs}$ be the total number of actual observed data points.

One approach to tackle the problem is to formulate this partial observations scenario as missing value problem and attempt to improve the estimation accuracy with data augmentation. Simply saying, data augmentation in Bayes treats those unobserved or missing data points between two consecutive observed as unknown parameters in addition to the unknown model parameters.

That being said, Bayesian inference treats unknowns as random variables and naturally associates them to probability distributions. In notation,

$$p(\theta, \tilde{x}|x) \propto p(\theta, \tilde{x}, x) \propto p(\theta)p(\tilde{x}, x|\theta), \tag{7.1}$$

where $x = \{x_1, \cdots, x_n\}$ is the actual observed data points, $\tilde{x}$ is latent or unobserved data points.

By doing so, the unknown model parameters θ (α and β in our cusp model) are to be estimated along with the incorporation $\tilde{x}$, potentially a very high dimensional data. On one hand, the complexity hence the computation cost of the problem has increased substantially due to the introduction of the high dimensional missing values, on the other hand however, the approximated transition density now has a much smaller discretization Δ^{obs} hence leads to less biased approximation - we are hoping more "accurate" estimation result could be attained.

A simple example

To better illustrate how data augmentation works, particularly how the latent variable $\tilde{x}$ are incorporated into the parameter estimation algorithm, we will now look at a concrete example:

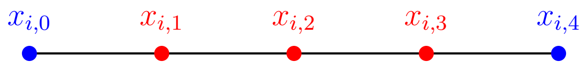


Figure 7.1: Illustration of Bayesian data augmentation

In the above plot, two end points in blue, namely $x_{i,0}$ and $x_{i,4}$ denote two consecutive *observed* data points. Beside, there are three points in red, namely $x_{i,1}, x_{i,2}$, and $x_{i,3}$. Not surprisingly, the three in red represent unobserved data points. They are "synthetic" data in the sense that we do not have the actual observation, either because we were not able to do so, or we just simply did not do so, but we assume data points have been *generated* at those time points. By incorporating those unobserved data points into our model, here's some immediate consequences:

1. the dimension of parameters increased substantially. To see this, let's suppose we have n observed data points. Now for each time interval between two consecutive observed data points, we introduced 3 unobserved data points. Consequently, we introduced $(n-1) \times 3$ new unknowns to be estimated in addition to the 2 unknown model parameters α and β.

2. On the other hand, by introducing the unobserved data points, the time difference two the combined data points (i.e. both the actual observed and the synthetic data points) is getting smaller. To see this, without introducing synthetic data, let's the time time difference between the measurement of two consecutive data points is Δ^{obs}, then by introducing additional 3 synthetic data points and inserting them into one observation interval, the difference becomes $\delta = \Delta^{obs}/4$. Recall that the approximated transition density $p_\theta^{\text{approx.}}(x,t)$ will have a better approximation to the true transition density since Δ^{obs} has been reduced to $\delta = \Delta^{obs}/4$. From this perspective, we would expect an improvement in the accuracy of the approximation due to a better approximated transition density hence likelihood function.

General case

Above example showed how data augmentation works by giving a simple illustration with 3 synthetic observation inserted between two consecutive observed data points. Of course the number of synthetic observation can be generalized to m.

Let $\{x_{i,j}\}_{j=1}^m$ be m (m is a non-negative integer) unobserved hence synthetic observations between two consecutive observed data points $x_i = x_{i,0}$ and $x_{i+1} = x_{i+1,0}$. Furthermore, let's assume $i = \{1, \cdots .n\}$ and let Δ^{obs} be the time difference between two observed data points, and let $\delta = \frac{\Delta^{obs}}{m+1}$.

Without assuming any synthetic data points between observed data points, the log-likelihood function given only the observed ones is:

$$\ell_n(\theta) \equiv \sum_{i=1}^{n} \ln \{p_x(\Delta, x_i | x_{i-1}; \theta)\}, \tag{7.2}$$

where the log-likelihood function given both observed and synthetic data points is

$$\ell_n(\theta;\tilde{x}) \equiv \sum_{i=1}^{n}\sum_{j=1}^{m} \ln \left\{p_x\left(\delta, x_{i,j}|x_{i,j-1};\theta\right)\right\}, \tag{7.3}$$

where $\delta = \frac{\Delta}{m+1}$.

7.2 Closed-form approximation using Hermite polynomials

In chapter 4, we showed in plot that the closed-form approximation using Hermite polynomial by Ait-Sahali is able to retain and capture the bimodality feature of the transition density for large Δ (Figure 4.3) when $\Delta^{\text{disc}} < 0$ - the case when cusp *stationary* distribution is actually bimodal. Thus, this theoretically promising approach can be used under this sparse-sampling scenario. Along with the ordinary Euler approximation, we compare the three different approaches, with the other two being and the Hermite polynomial approximation, and Euler approximation with data augmentation.

7.3 Simulation study and result

To compare the three proposed methods, we run simulation study in two cases, one is assuming $\Delta^{obs} = 0.2$, and the other assumes $\Delta^{obs} = 0.4$. Furthermore under both cases, we assume data are generated at 10 data points per unit time interval (i.e. $\Delta^{gen} = 0.1$) and is simulated using Euler's method.

In both demonstrated cases (Table 7.1), Euler approximation with data augmentation outperforms Euler approximation and approximation using Hermite polynomial, in terms of coverage probability and bias.

Moreover, if there were a considerable number of many missing values, which often resulted in more sparse sampling than "continuous" which it should be (for example the case with 4 observed data points and 6 unobserved data points in a unit time interval), Neither Hermite polynomial nor regular Euler would be able to obtain satisfying simulation study results. The simulation-study result is not surprising, since as pointed by Melino [1996], Jones [1998], and Ait-Sahalia et al. [2008], in cases where time interval Δ^{obs} between to consecutive observation points are large, the discrepancy between continuous model assumption and discretely observed data points should not be neglected, because otherwise it could lead to inconsistent estimators. However, by applying Bayesian data augmentation with Euler, the augmented Euler's method was able to improve the performance supported by increasing coverage probabilities with cost of increasing variance (i.e. bias-variance trade-off).

7.4 Effect of number of augmented data points

In this section, we investigate how number of augmented data points would affect the parameter estimation result by running simulation studies.

W generate 500 sample trajectories with parameter values $\alpha = 1$ and $\beta = 3$. Each sample trajectory is simulated using Euler's method with discretization step size $\Delta = 0.1$ for 60 unites of times; therefore 600 data points were generated and considered a nearly "continuous" sample trajectory. We pretend that we were only able to observe 1 observation per unit time, which leads us to a total number of 60 observe data points. The goal is to use this 60 data points as observation to perform parameter estimations. The result is summarized at Table 7.2.

From the Table 7.2, we've observed that the coverage probability for both α and β increase as number of augmented data points being inserted increase; meanwhile the cost is the variance of the estimator (i.e. α_{MAP} and β_{MAP}) is getting bigger.

According to the simulation study result, by increasing the number of augmented data points per unit time interval, what we have gained is the increasing coverage probabilities for both α and β - this might be more practically desirable in reality; at the same time, the trade-off is the estimator itself become *skewed*. From the plots, we clearly see that the empirical estimators become more skewed as more augmented data points being inserted.

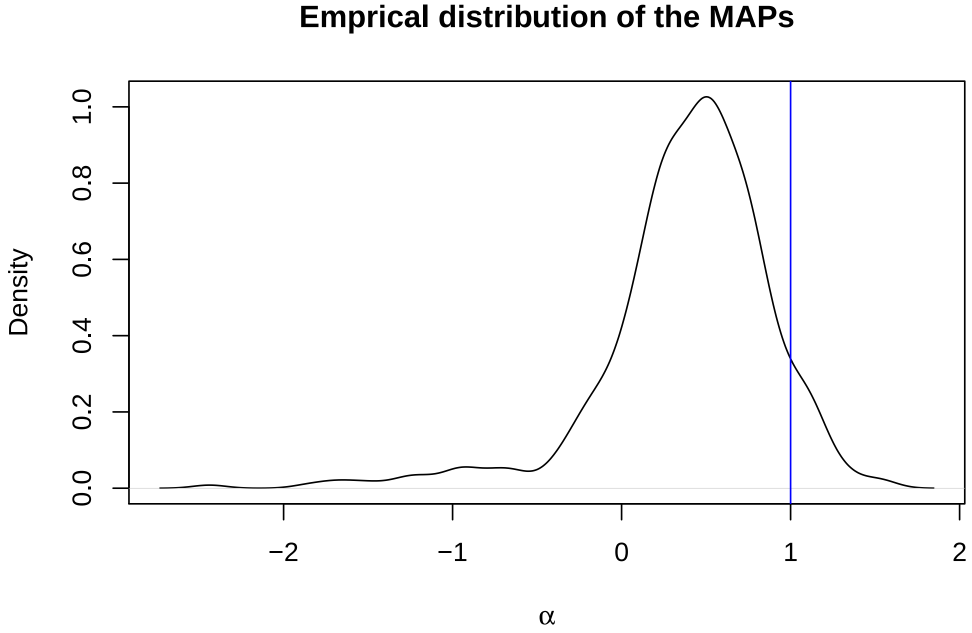


Figure 7.2: Partial observations: Empirical $\alpha_{\text{MAP}}$with 1 augmented data point

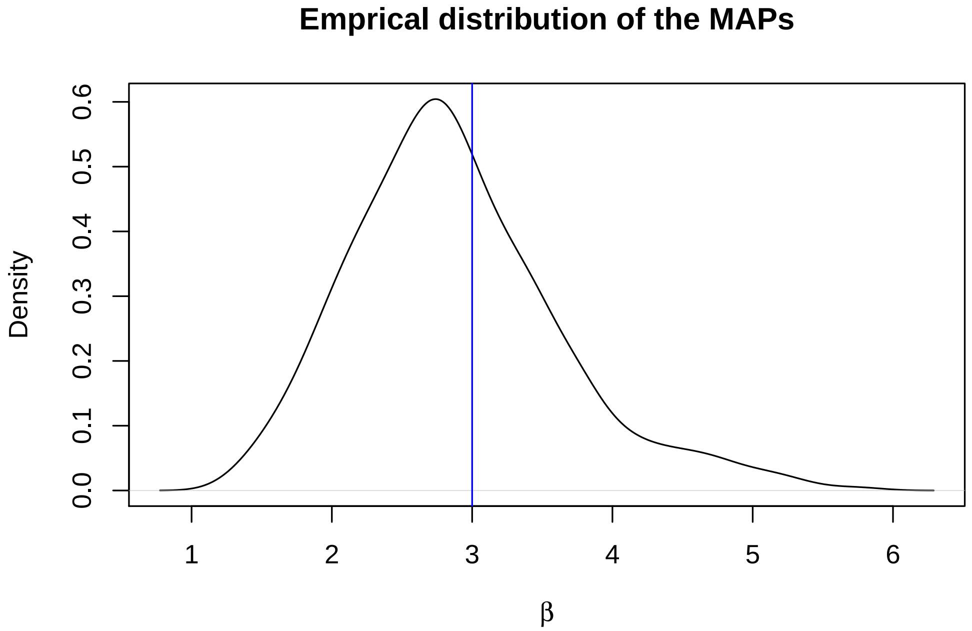


Figure 7.3: Partial observations: Empirical β_{MAP} with 1 augmented data point

		α_{MAP}				β_{MAP}			
Sampling freq.	Method	Mean	SE	Empirical SE	CP	Mean	SE	Empirical SE	CP
5 obs. or $\Delta^{\text{obs}} = 0.2$	Hermite	0.555	0.228	0.313	0.751	3.445	0.269	0.198	0.407
	Euler	0.638	0.272	0.312	0.812	3.431	0.274	0.196	0.425
	Euler (augmented)	4.368	6.968	0.409	0.775	0.244	5.607	0.263	0.761
2.5 obs. or $\Delta^{\text{obs}} = 0.4$	Hermite	0.414	0.309	0.311	0.548	3.910	0.376	0.194	0.071
	Euler	0.322	0.429	0.312	0.462	3.785	0.416	0.197	0.157
	Euler (augmented)	3.378	05.793	0.451	0.713	1.190	4.510	0.333	0.776

Table 7.1: Simulation study: Inference from partial observations

For each setting, posterior analysis is performed based on 1,000 replications, each contains $60/\Delta^{\text{obs}}$ data points.

	α_{MAP}					β_{MAP}				
Num. Augmented Pts	Mean	Mode	SE	Empirical SE	CP	Mean	Mode	SE	Empirical SE	CP
1	0.389	0.503	0.506	0.432	0.732	2.874	2.739	0.753	0.346	0.624
2	1.003	0.841	0.517	0.474	0.918	2.144	2.349	0.734	0.453	0.490
4	1.653	1.172	0.970	0.613	0.780	2.300	2.654	0.895	0.622	0.782
9	4.065	0.946	6.681	0.644	0.772	0.728	2.858	5.000	0.514	0.786
15	3.048	0.816	8.880	0.587	0.892	1.334	2.748	5.732	0.416	0.846

Table 7.2: Simulation study: Effect of number of augmented data points

For each setting, posterior analysis is performed based on 500 replications, each contains 60 data points;
Parameter values were set to be $\alpha = 1$ and $\beta = 3$.

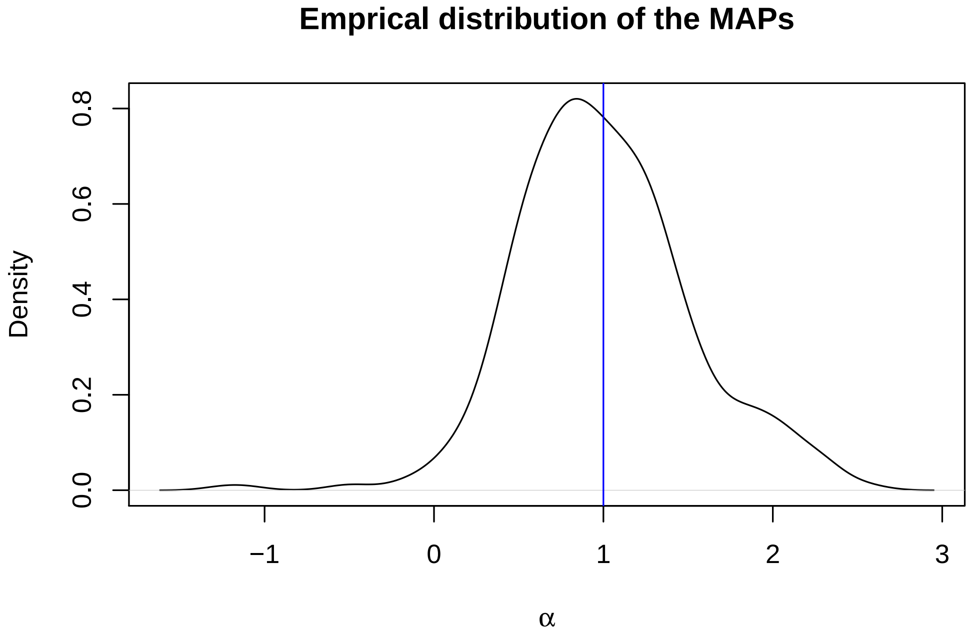


Figure 7.4: Partial observations: Empirical α_{MAP} with 2 augmented data points

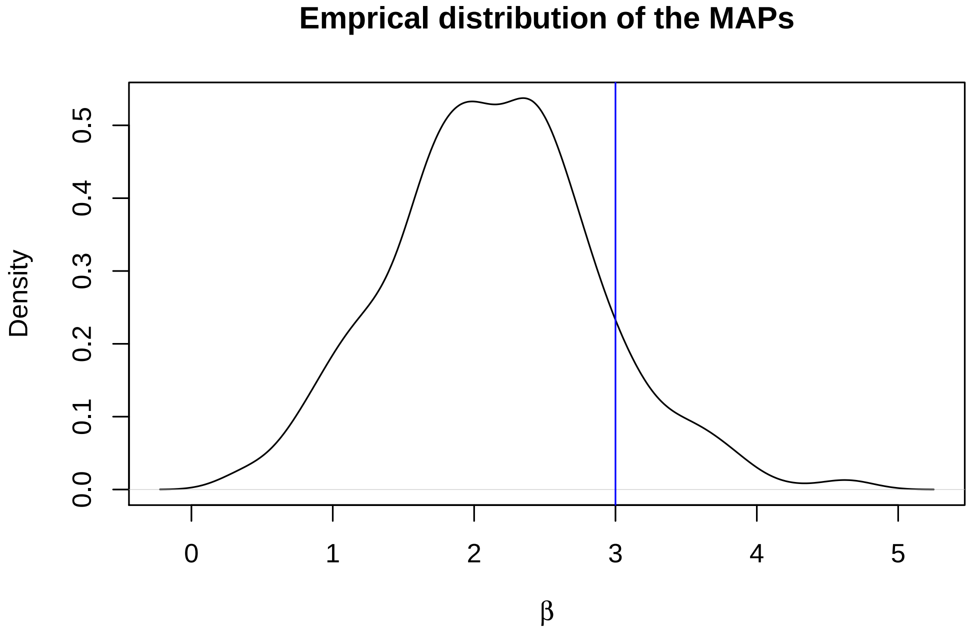


Figure 7.5: Partial observations: Empirical β_{MAP} with 2 augmented data points

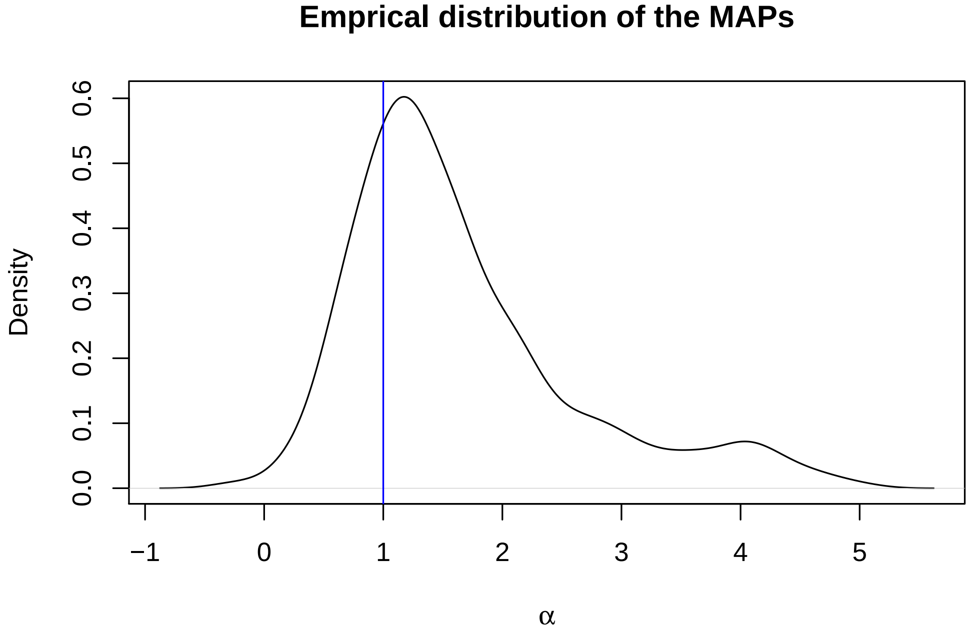


Figure 7.6: Partial observations: Empirical α_{MAP} with 4 augmented data points

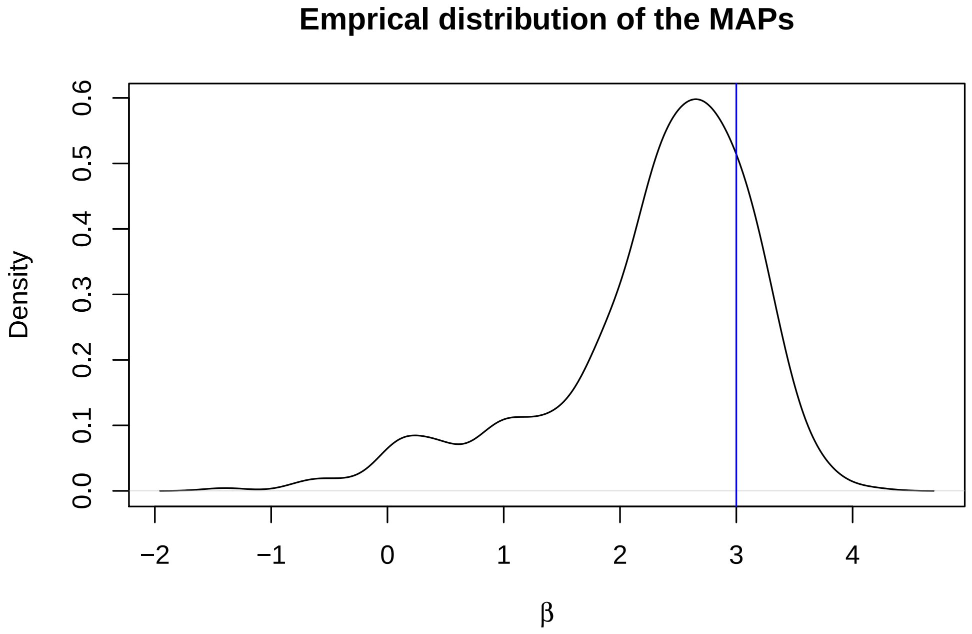


Figure 7.7: Partial observations: Empirical β_{MAP} with 4 augmented data points

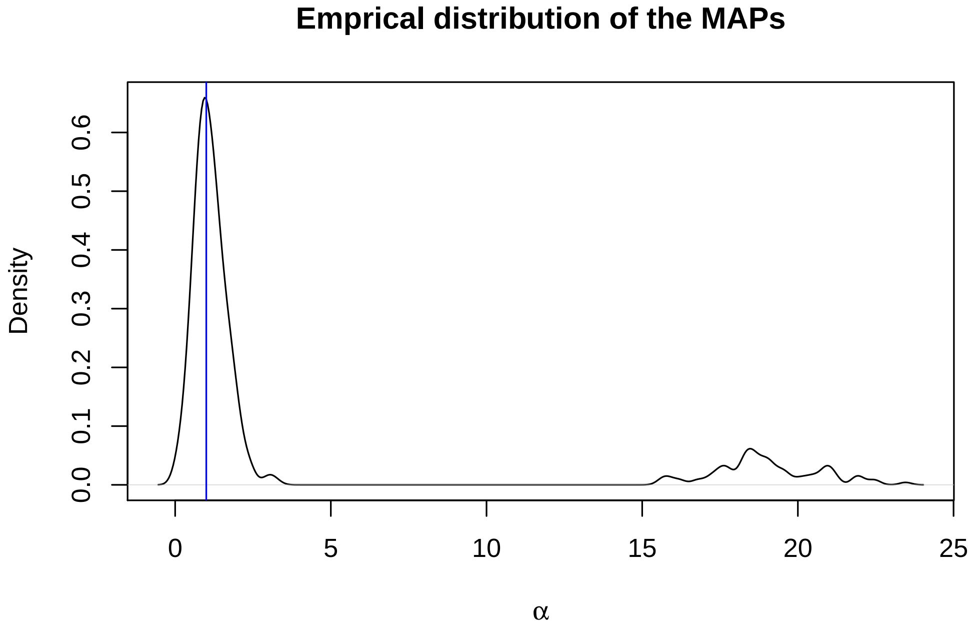


Figure 7.8: Partial observations: Empirical α_{MAP} with 9 augmented data points

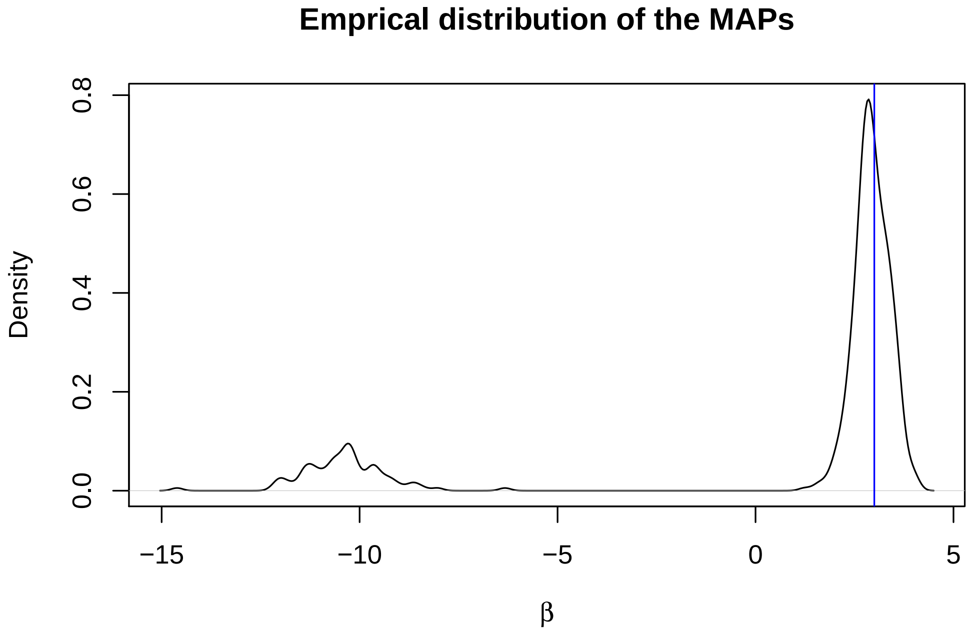


Figure 7.9: Partial observations: Empirical β_{MAP} with 9 augmented data points

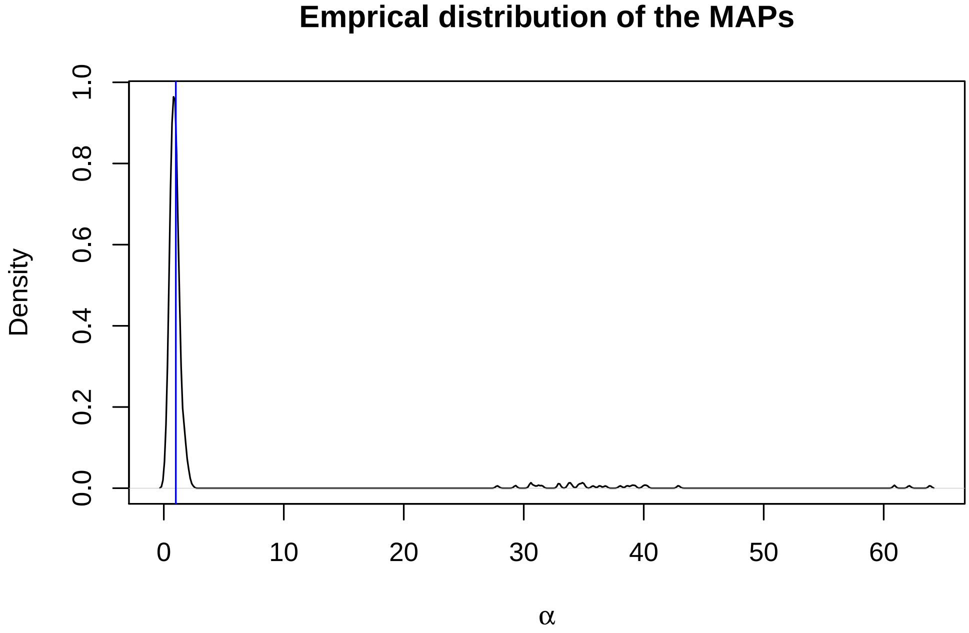


Figure 7.10: Partial observations: Empirical α_{MAP} with 15 augmented data points

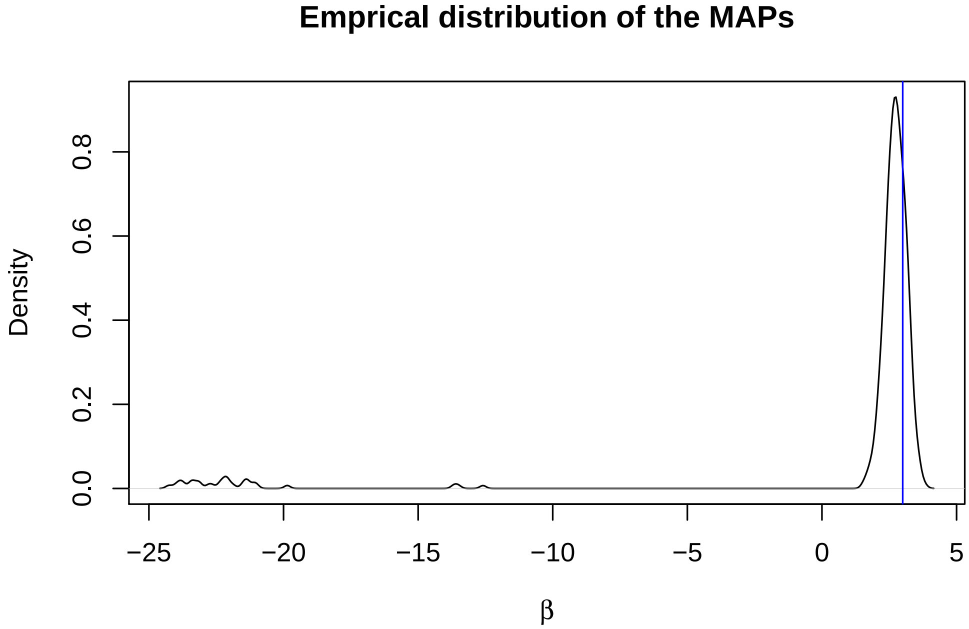


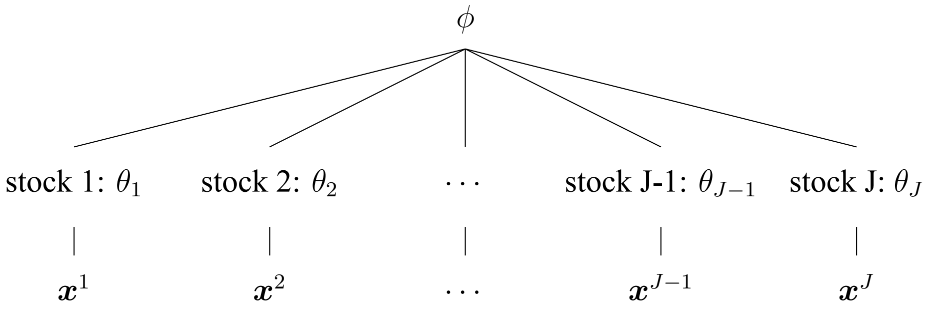
Figure 7.11: Partial observations: Empirical β_{MAP} with 15 augmented data points

CHAPTER 8
Cusp Model with More Complex Structure

Bayesian hierarchical modeling, as its name suggests, is a statistical modeling approach that constructs a model with multiple levels or in hierarchical form and estimates the parameters using Bayesian inference. Bayesian hierarchical modeling is commonly used when the problem can be formulated with several levels or hierarchies of observational units, and this hierarchical organization often helps understand the relationship of parameters on and between different levels.

8.1 Bayesian hierarchical modeling

A simple Bayesian hierarchical model often consists of three layers, namely the data layer, process layer, and prior layer. This three-layer organization can be further used to describe two levels of units, namely the individual level and the population level.



8.1.1 Population and individual

To better illustrate how this hierarchical organization works, we now use a portfolio of stocks as an example.

Individual level describes the behavior of individuals over time. For example, if individual is thought as a particular stock, and we are interested in the stock price movement over some time interval, denoted by X_t^j. We can further make the underlying model assumption that X_t^j is a diffusion process governed by cusp SDE with model parameters θ_j. We track the trajectory X_t^j over some time interval by taking measurements on discrete time points with interval Δ^{obs}, and the time-series data is denoted by $\boldsymbol{x}^j = \{x_1^j, \cdots, x_n^j\}$.

Now suppose there's a portfolio consists of several individual stocks, this portfolio is the population level. In particular, Bayesian hierarchical modeling assumes individual parameters $\theta_1, \theta_2, \cdots, \theta_J$ are generated from a common population p that is further governed by a hyper parameter ϕ. At population level, across-unit analysis can be used to capture the heterogeneity or the diversity across individuals.

Structurally, a Bayesian hierarchical model often consists of three layers, namely the data layer, a process layer, and a prior layer. More specifically, in the data layer, observations associated to jth individual is assumed to be generated by cusp process (Equation 1.8), i.e. $\boldsymbol{x}^j \mid \theta_j$; In process layer, individual model parameter is assumed to be generated from population, i.e. $\theta_j \sim p(\theta_j|\phi)$; The prior level simply assigns a prior on hyper-parameters, i.e. $\phi \sim p(\phi)$.

In our case, Bayesian hierarchical model is mainly used to describe (1) the behavior of individuals in a study, which models the with-in unit behavior over a time interval, and (2) the distribution of responses across individuals, which reflects cross-sectional variation in model parameters, or the heterogeneity.

8.1.2 Simulation study and result

In simulation study, we first fix the population level parameter to be

$$\mu_\alpha = 1, \;\; \sigma_\alpha = 0.1; \quad \mu_\beta = 4, \;\; \sigma_\beta = 0.1.$$

Then, we generate 20 pairs of parameters, each pair are generated according to the population-level parameters:

$$\alpha_i \sim \mathcal{N}(\mu_\alpha, \mu_\alpha), \quad \beta_i \sim \mathcal{N}(\mu_\beta, \sigma_\beta).$$

We then simulate 20 trajectories for each pair of parameters

$$dX_t^j = \left(\alpha + \beta X_t^j - \frac{1}{4} X_t^{j^3}\right) dt + \sqrt{\varepsilon} dW_t$$

using Euler's method with discretization step size $\Delta = 0.1$ for $T = 180$ units of time. The parameters of interest are population-level parameters $\mu_\alpha, \mu_\beta,$ and , while we treat σ_α and σ_β as known quantities. Again, here we fix $\sqrt{\varepsilon} = 2$ and treat it as a constant when estimating other model parameters. Posteriors are given by Figure 8.1 and Figure 8.2

The posterior plots support the accuracy of the algorithm.

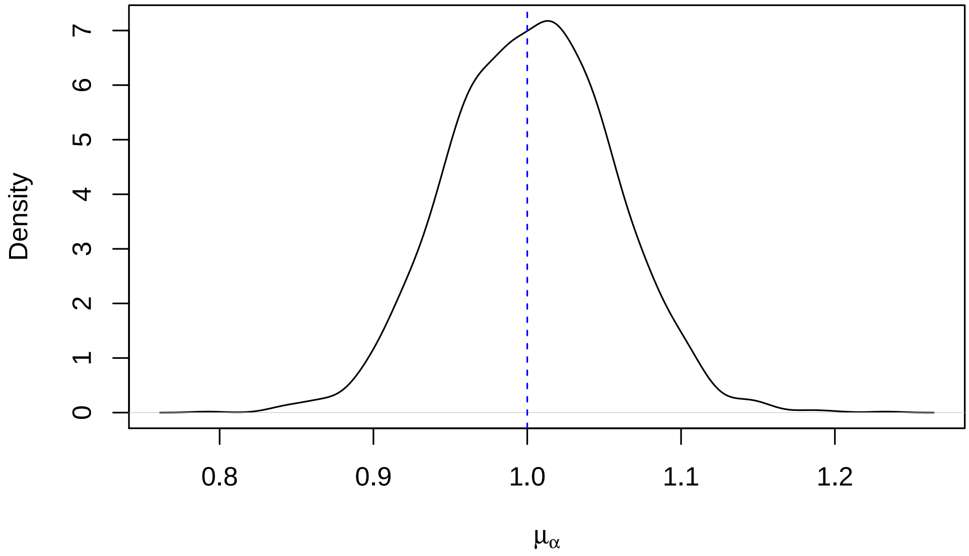


Figure 8.1: Bayesian hierarchical modeling: Posterior αs

8.2 Cusp model with time-varying parameters

In reality, there are abundant reasons to believe that the underlying data generating process for many real-world situation might change over time. Needless to say, economic data that reflect

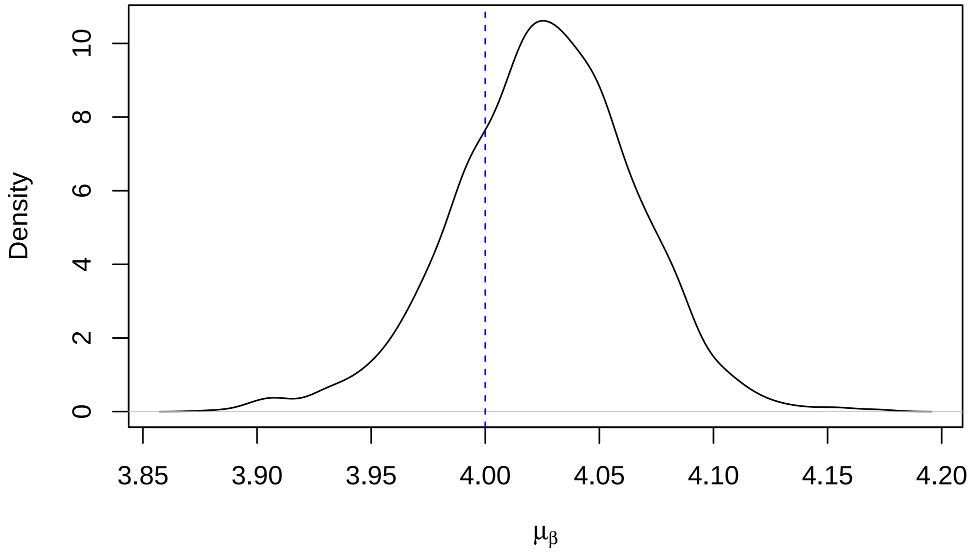


Figure 8.2: Bayesian hierarchical modeling: Posterior βs

an actual economy in a economy unit may depend on many other economic indicators, economic policy, etc. This motivates us to extend our cusp model into time-varying process with exogenous processes.

8.2.1 Time-varying parameters

Following the work of Creedy et al. [1996] and Fernandes [2006], one could naturally extent parameters α and β governing the number of stable equilibria of the cusp model to be time-varying, specifically they can be model in such a way that both parameters are depended on some strictly *exogenous* process or processes ζ_t. Here *strictly* is with respect to process X_t.

The stochastic differential equation of the cusp model with time-varying parameters $\alpha(\zeta_t)$ and $\beta(\zeta_t)$ takes the form

$$dX_t = \left(\alpha\left(\zeta_t\right) + \beta\left(\zeta_t\right) X_t - \left(X_t\right)^3\right) dt + \sigma dW_t. \tag{8.1}$$

Let's assume we have two strictly exogenous process ζ_t^1 and ζ_t^2 which will be used as covariate processes, and consider the linear dependence of both α_t and β_t on ζ^1, ζ^2 in the way

$$\alpha_t \equiv \alpha\left(\zeta_t^1,\ \zeta_t^2\right) = \alpha_0 + \alpha_1\zeta_t^1 + \alpha_2\zeta_t^2$$
$$\beta_t \equiv \beta\left(\zeta_t^1,\ \zeta_t^2\right) = \beta_0 + \beta_1\zeta_t^1 + \beta_2\zeta_t^2,$$

where $\theta = \{\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2\}^\top$ are the model parameters hence to be estimated.

8.2.2 Simulation study and result

In our simulation study, the pre-fixed model parameters are

$$\alpha_0 = 0,\ \ \alpha_1 = 2,\ \ \alpha_2 = -2; \quad \beta_0 = 5,\ \ \beta_1 = -2,\ \ \beta_2 = 1.$$

The choice of ζ_1 and ζ_2 being trigonometric (periodic) functions is motivated by two major reasons, one is because covariate (or independnent) processes are often more "predictable" than the target process; the other is because some economic factors indeed exhibit periodic behaviors or *cycles*.

We are hoping to use observations denoted by y plus two exogenous processes, denoted by ζ_1 and ζ_2 to estimate model parameters $\{\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2\}$. Sample trajectories of y, ζ_1 and ζ_2 with pre-fixed model parameters are simulated as shown in Figure 8.3 with discretization step-size $\Delta^{\text{gen}} = 0.01$ for $T = 180$ units of time. Posterior distributions are given below. The simulation study result supports the accuracy of the parameter estimation method.

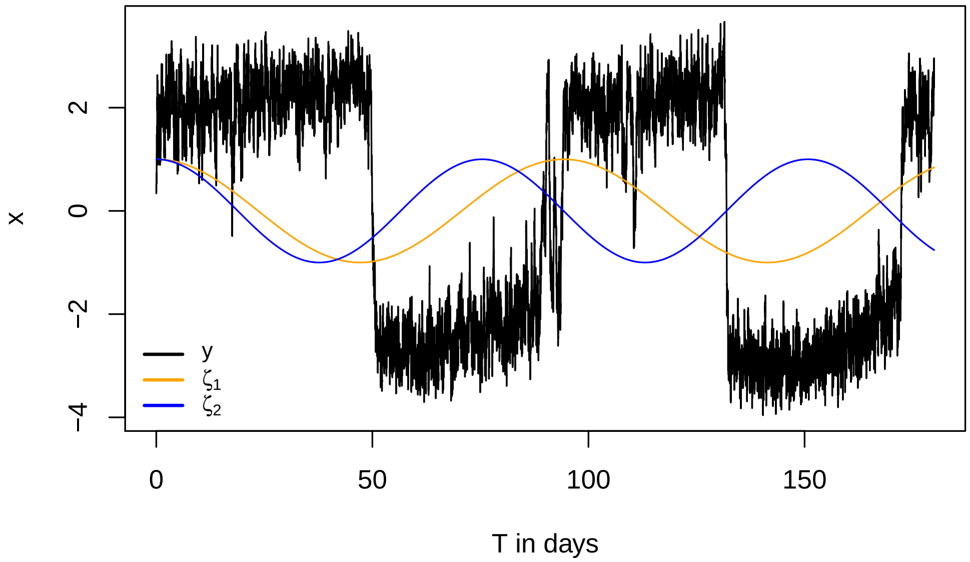


Figure 8.3: Time-varying cusp: Sample trajectories

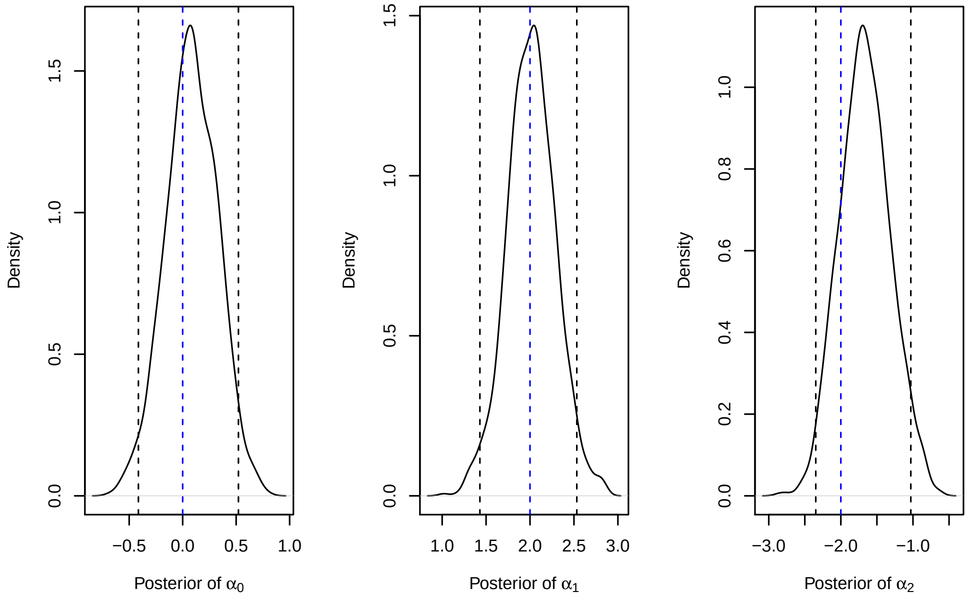


Figure 8.4: Time-varying cusp: Posterior αs

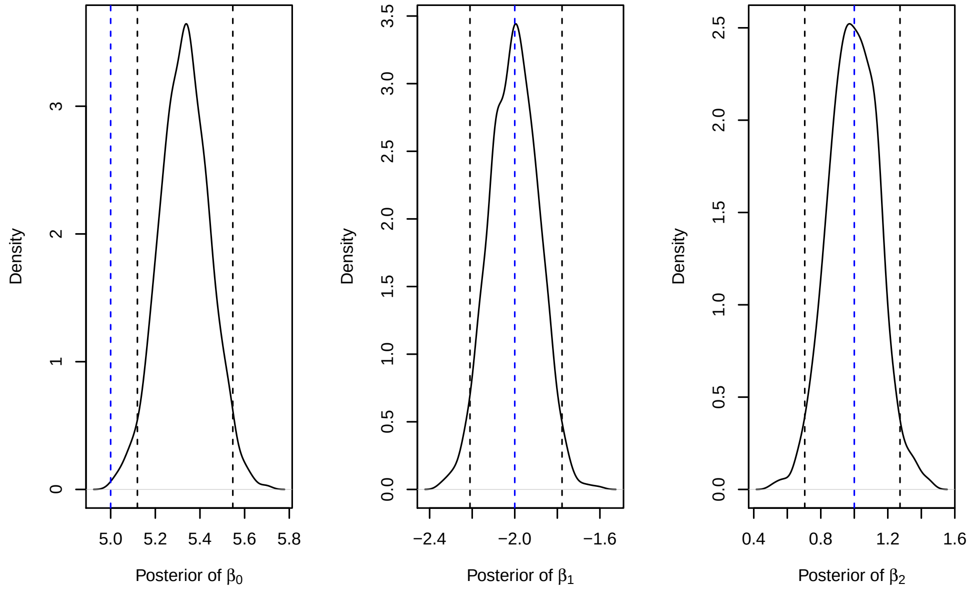


Figure 8.5: Time-varying cusp: Posterior βs

CHAPTER 9

Conclusion

Bayesian inference on non-linear diffusion models such as cusp model, is an interesting research topic. Exploring how one can apply cusp model in financial econometrics or other subjects is also very interesting and exciting.

In this thesis, we considered cusp model, one of the elementary catastrophe models studied in catastrophe theory. We demonstrated how Bayesian inference can be used as a solution to cusp SDE inference problem via Hamiltonian Monte Carlo with different likelihood approximation methods. The proposed method has been tested by a series of intensive simulation studies under different scenarios, including inference from complete observations, from partial observations, as well as inference for more complex models such as Bayesian hierarchical modeling and time-varying parameters setting. Particularly, in the partial observations scenario, we (1) showed how Bayesian data augmentation could be used to help remedy the model-observation discrepancy when observations are sparsely taken, and (2) investigated how number of augmented data points would affect the result of parameter estimation by running simulation studies.

Advantages and limitations of the methods from the simulation studies have been presented hoping to support practitioners to select suitable methods for their real-world problems, as well as to encourage further theoretical and empirical studies related to cusp model.

APPENDIX A
STATIONARY DISTRIBUTION OF CATASTROPHE MODEL

Fokker-Planck equation reveals that the *transition* probability density $p(t, x|x_0, \theta)$ of an stochastic differential equation obeys a deterministic *partial differential equation*. Instead of focusing on the individual trajectory x, this section explores the time evolution of the transition probability density $p(t, x|x_0, \theta)$.

In one spatial dimension x, for an Itô process driven by standard Wiener process W_t and described by the stochastic differential equation

$$dX_t = \mu\left(X_t, t\right)dt + \sqrt{\nu\left(X_t, t\right)}dW_t \tag{A.1}$$

with drift $\mu(X_t, t)$ and diffusion coefficient ν, the Fokker-Planck equation, also known as the Kolmogorov forward equation for the probability density

$$p(x,t) \equiv \frac{d}{du}\,\text{Prob}\{x(t) < u|x(0) = x_0\}$$

of the random variable X_t is

$$\frac{\partial}{\partial t}p(x,t) = -\frac{\partial}{\partial x}[\nu(x,t)p(x,t)] + \frac{1}{2}\frac{\partial^2}{\partial x^2}[\nu(x,t)p(x,t)]. \tag{A.2}$$

The Fokker-Planck equation can be used to obtain the transition probability density of the solution process $p(t, x|x_0, \theta)$ and perhaps a less ambitious goal - the stationary solution of the transition probability density.

The *stationary distribution* denoted by p_s can be obtained by solving $\partial_t p(x,t) = 0$ (i.e. constant in t), that is

$$\frac{d}{dx}\left[\mu(x)p_\text{s}(x)\right] - \frac{1}{2}\frac{d^2}{dx^2}\left[\nu(x)p_\text{s}(x)\right] = 0. \tag{A.3}$$

Upon integrating once we have

$$\mu(x)p_{\mathrm{s}}(x) - \frac{1}{2}\frac{d}{dx}\left[\nu(x)p_{\mathrm{s}}(x)\right] = c_1 = \text{const.} \tag{A.4}$$

If we write

$$\nu(x)p_{\mathrm{s}}(x) = h(x) \tag{A.5}$$

then equation becomes

$$\frac{d}{dx}h(x) - 2\frac{\mu(x)}{\nu(x)}h(x) = -2c_1. \tag{A.6}$$

Hence the general solution for p_{s} is, upon substituting into

$$p_{\mathrm{s}}(x) = \frac{c_2}{\nu(x)}\exp\left[2\int_0^x \frac{\mu(s)}{\nu(s)}ds\right] - \frac{2c_1}{\nu(x)}\int_0^x \exp\left[2\int_r^x \frac{\mu(s)}{\nu(s)}ds\right]dr \tag{A.7}$$

The constants of integration, c, and c_2, are determined from normalization and boundary conditions. If we assume that

$$p_{\mathrm{s}}(x)(\pm\infty) = 0 \quad \text{and} \quad \left.\frac{d}{dx}p_{\mathrm{s}}(x)\right|_{x=\pm\infty} = 0, \tag{A.8}$$

It's seen from Eqs that $c_1 = 0$ and we have

$$\begin{aligned} p_{\mathrm{s}}(x) &= \frac{c_2}{\nu(x)}\exp\left[2\int^x \frac{\mu(s)}{\nu(s)}ds\right] \\ &= c_2 \cdot \exp\left[2\int^x \frac{\mu(s)}{\nu(s)} - \frac{1}{2}\frac{\nu'(s)}{\nu(s)}ds\right] \end{aligned} \tag{A.9}$$

If the underlying stochastic differential equation happens to have a constant diffusion term ε, (A.9) can be further simplified to

$$p_s(x) = c_2 e^{-\frac{2V(x)}{\varepsilon}} \tag{A.10}$$

APPENDIX B

CUSP TRANSITIONAL PDF BY HERMITE POLYNOMIALS

For cusp stochastic differential equation expressed in the form

$$dX_t = (a + bX_t + cX_t^2 + dX_t^3)dt + f dW_t, \tag{B.1}$$

by using (4.18), we can obtain a closed-form approximation up to Δ^2 to the transition probability density $p(\Delta, x|x_0, \theta)$, and its log-likelihood function is given by

$$p^{(2)}(\Delta, x|x_0, \theta) = -\log(2\pi\Delta)/2 - \log(f) + cm_1/\Delta + c_0 + c_1\Delta + c_2\Delta^2/2. \tag{B.2}$$

Closed-form approximation using Hermite polynomial up to the second order is given in form of R user-defined function:

```
cusp2 <- function(x, x0, delt, a, b, c, d, f ){

  sx = f

  cm1 = -(x - x0) ^ 2 / (2 * f ^ 2)

  c0 = (4 * c * x ^ 3 + 3 * d * x ^ 4 + 12 * a * (x - x0) - 4 * c * x0 ^ 3 - 3 *
          d * x0 ^ 4 + 6 * b * (x ^ 2 - x0 ^ 2)) / (12 * f ^ 2)

  c1 = -1 / (420 * f ^ 2) * (
    210 * a ^ 2 + 70 * b ^ 2 * (x ^ 2 + x * x0 + x0 ^ 2) +
      35 * a * (
        6 * b * (x + x0) + 4 * c * (x ^ 2 + x * x0 + x0 ^ 2) +
          3 * d * (x ^ 3 + x ^ 2 * x0 + x * x0 ^
                     2 + x0 ^ 3)
      ) +
      21 * b * (
        10 * f ^ 2 + 5 * c * (x ^ 3 + x ^ 2 * x0 + x * x0 ^ 2 + x0 ^ 3) +
          4 * d * (x ^ 4 + x ^ 3 * x0 + x ^ 2 *
                     x0 ^ 2 + x * x0 ^ 3 + x0 ^ 4)
      ) +
      2 * (
        21 * c ^ 2 * (x ^ 4 + x ^ 3 * x0 + x ^ 2 * x0 ^ 2 + x * x0 ^ 3 + x0 ^ 4) +
          35 * c * (x + x0) * (3 * f ^ 2 + d * (x ^
                                                  4 + x ^ 2 * x0 ^ 2 + x0 ^ 4)) +
          15 * d * (
            7 * f ^ 2 * (x ^ 2 + x * x0 + x0 ^ 2) +
              d * (x ^ 6 + x ^ 5 * x0 + x ^ 4 *
                     x0 ^ 2 + x ^ 3 * x0 ^ 3 + x ^ 2 * x0 ^ 4 + x * x0 ^ 5 +
                     x0 ^ 6)
          )
      )
  )

  c2 = 1.0 / 210 * (
        -35 * b ^ 2 - 105 * d * f ^ 2 - 63 * c ^ 2 * x ^ 2 - 140 * c * d * x ^ 3 - 75 *
        d ^ 2 * x ^ 4 -
        84 * c ^ 2 * x * x0 - 210 * c * d * x ^ 2 * x0 - 120 * d ^2 *
        x ^ 3 * x0 - 63 * c ^ 2 * x0 ^ 2 -
        210 * c * d * x * x0 ^ 2 - 135 * d ^ 2 * x ^ 2 * x0 ^2 -
        140 * c * d * x0 ^ 3 - 120 * d ^ 2 * x * x0 ^ 3 -
        75 * d ^ 2 * x0 ^ 4 - 35 * a * (2 * c + 3 * d * (x + x0)) -
        21 * b * (5 * c * (x + x0) + 2 * d * (3 * x ^ 2 + 4 *
        x * x0 + 3 * x0 ^ 2)
        )
  )
  return(exp(-log(2 * pi * delt) / 2 - log(sx) + cm1 / delt + c0 + c1 * delt + c2 *
               delt ^ 2 / 2 ))
}
```

Listing B.1: HPE: J = 2

BIBLIOGRAPHY

Ait-Sahalia, Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies,9,385-426*.

Ait-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. *Econometrica*, 70(1):223–262.

Ait-Sahalia, Y. et al. (2008). Closed-form likelihood expansions for multivariate diffusions. *The Annals of Statistics*, 36(2):906–937.

Baruník, J. and Vosvrda, M. (2009). Can a stochastic cusp catastrophe model explain stock market crashes? *Journal of Economic Dynamics and Control*, 33(10):1824–1836.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Chen, D.-G. D., Chen, X. J., and Zhang, K. (2016). An exploratory statistical cusp catastrophe model. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 100–109. IEEE.

Cobb, L. (1978). Stochastic catastrophe models and multimodal distributions. *Behavioral Science*, 23(4):360–374.

Cobb, L. (1981). Parameter estimation for the cusp catastrophe model. *Behavioral Science*, 26(1):75–78.

Cobb, L., Koppstein, P., and Chen, N. H. (1983). Estimation and moment recursion relations for multimodal distributions of the exponential family. *Journal of the American Statistical Association*, 78(381):124–130.

Costantino, R. F., Desharnais, R. A., Cushing, J. M., Dennis, B., Henson, S. M., and King, A. A. (2005). Nonlinear stochastic population dynamics: the flour beetle tribolium as an effective tool of discovery. *Advances in Ecological Research*, 37:101–141.

Creedy, J., Lye, J., and Martin, V. L. (1996). A non-linear model of the real us/uk exchange rate. *Journal of Applied Econometrics*, 11(6):669–686.

Creedy, J. and Martin, V. (1993). Multiple equilibria and hysteresis in simple exchange models. *Economic Modelling*, 10(4):339–347.

Durham, G. B. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business & Economic Statistics*, 20(3):297–338.

Fernandes, M. (2006). Financial crashes as endogenous jumps: estimation, testing and forecasting. *Journal of Economic Dynamics and Control*, 30(1):111–141.

Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics: A Journal of Theoretical and Applied Statistics*, 20(4):547–557.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Geweke, J. et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

Iacus, S. M. (2009). *Simulation and inference for stochastic differential equations: with R examples*. Springer Science & Business Media.

Ivancevic, V. G. and Ivancevic, T. T. (2007). *Computational mind: a complex dynamics perspective*, volume 60. Springer.

Jimenez, J., Biscay, R., and Ozaki, T. (2005). Inference methods for discretely observed continuous-time stochastic volatility models: a commented overview. *Asia-Pacific Financial Markets*, 12(2):109–141.

Jones, C. S. (1998). Bayesian estimation of continuous-time finance models. *manuscript University of Rochester*.

Melino, A. (1996). Estimation of continuous-time models in finance. In *Advances in Econometrics Sixth World Congress, ed. Sims, CA, Cambridge: Cambridge University Press*, volume 2, pages 313–351.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2.

Platen, E. (1999). An introduction to numerical methods for stochastic differential equations. *Acta numerica*, 8:197–246.

Zeeman, E. (1973). On the unstable behaviour of stock exchanges. *Journal of Mathematical Economics 1 (1974) 39-49*.