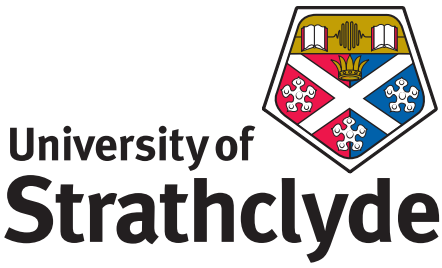

Numerical approximation and parametric statistical inference of stochastic differential equations, with applications to finance

Steven Craig
Department of Mathematics & Statistics
University of Strathclyde
Glasgow, UK
April 2014

This thesis is submitted to the University of Strathclyde for the degree of Doctor of Philosophy in the Faculty of Science.

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Acknowledgements

I am grateful to my supervisor, Xuerong Mao, for the guidance he has provided me in the course of my studies. I would also like to take this opportunity to thank my second supervisor, Vladislav Vyshemirsky, for introducing me to approximate Bayesian computation, providing me with lots of support and advice throughout the past few years, and for having the patience to answer my questions and decipher my C++ code on numerous occasions. Thirdly, I would like to thank Steven Morrisson and Graeme Lawson, both of whom acted as my CASE sponsor contacts at Barrie and Hibbert, and provided support when required. My thesis also benefitted from several discussions with Des Higham and Andrew Wade, to whom I am very grateful. I would also like to acknowledge the generous financial support provided by my CASE industrial sponsor, Barrie and Hibbert, and by the Engineering and Physical Sciences Research Council (EPSRC). Finally, I would like to extend a heartfelt thanks to my family and friends for putting up with me over the past three and a half years; in particular, I would like to thank my partner, Carolyn, for her continued support and encouragement.

Abstract

Stochastic differential equations (SDEs) have become an indispensable tool for modelling the dynamics of key state variables in mathematical finance such as instantaneous short rates of interest, share prices, and volatility processes. The appropriate application of SDEs requires reliable methods of generating sample paths from the equations, e.g. for use in Monte Carlo simulations, and robust parameter estimation methods to calibrate the SDEs to observed market data. Proposed stochastic models for financial variables are becoming increasingly complex in an effort to produce more realistic models, but only on rare occasions are the analytic expressions for the processes' transition densities available. Consequently, it is rare to be able to simulate sample data from the exact process, or conduct full likelihood-based inference. This difficulty motivates the need for approximation methods that are capable of simulating approximate sample paths with desirable convergence properties such that approximation errors can be controlled; and flexible parameter estimation methods that are not materially hampered by a paucity of analytic results associated with intractible SDEs. In this thesis we introduce a numerical approximation scheme for a class of SDEs that are widely applicable to finance. We prove the strong convergence of the numerical scheme and provide a lower bound on the convergence rate associated with the scheme. We also explore the subject of parameter estimation in the context of SDEs, and present three new parameter estimation techniques. By an application of approxiate

Bayesian computation (ABC) we develop two sampling algorithms that are capable of producing high quality approximations to the posterior distribution of model parameters, without any need to evaluate model likelihoods.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of subject area

Since the mid-eighties, stochastic analysis has been widely used within the financial services industry. In particular, stochastic differential equations (SDEs) have been used as parametric models to describe the dynamics of financial quantities such as equity prices, interest rates, and inflation rates. Although in practice financial quantities are quoted in markets at discrete times, and traded in discrete amounts, SDEs (which are continuous-time processs) have been successfully applied to the problem of modelling the behaviour of such quantities. Aside from technological advances in financial services, which have resulted in more high-frequency data being available for analysis (which puts the continuous-time approach associated with SDEs on more solid ground), there is a rich suite of analytic results and tools related to stochastic processes that are available to the analyst, which makes the task of managing and modelling the risk associated with financial variables more feasible.

Broadly speaking, there are two areas in finance in which SDEs are commonly used: the pricing of complex financial products; and evaluating portfolio risk. A famous

example of the former use is the celebrated option pricing formula developed by Fisher Black and Myron Scholes Black and Scholes (1973), which evaluates the fair price of a European call option (a type of financial derivative) assuming that the share price, whose value determines the value of the option, obeys geometric Brownian motion, a particular type of SDE which we will discuss in detail in the main body of this thesis. Derivative pricing formulae are functions of the parameters of the SDE being used to model the underlying financial process on which the derivative price depends and, as such, it is important that the model parameters are accurately estimated from the observed market data. Additionally, the pricing of exotic financial derivatives such as path dependent options often requires efficient numerical simulation schemes that converge strongly to the true solution of the underlying SDE with a favourable rate. Therefore, a failure to use appropriate numerical simulation schemes, or to correctly specify the model parameters can lead to serious derivative pricing errors, which in turn can result in damaging economic consequences for the financial services provider and their stakeholders. As well as being useful from the perspective of derivative pricing, SDEs are widely used to assess the risk associated with holding a portfolio of assets over a given time horizon. For example, a common method of evaluating the capital requirements of financial service providers is to use Monte Carlo methods to derive the empirical distribution of losses over a given time period (e.g. one year), and then use the nth percentile, e.g. the 99.5th percentile, of simulated losses as an indication of how much capital an organisation must hold to stand a good chance of avoiding insolvency during periods of excessive market volatility. Of course, the empirical distribution obtained from the Monte Carlo simulation exercise will be heavily dependent on the choices of model parameters used to simulate future scenarios; once again, a poorly calibrated parametric model could lead to inadequate capital

provision, which in turn could expose stakeholders (customers, employees, investors, etc.) to excessive levels of risk.

As the field of mathematical finance has evolved over the years, the complexity of financial models has grown due to the desire to find models that can more plausibly account for the rich variety of behaviour exhibited by financial variables like share prices. While this growth in model complexity is a positive development, in the sense that the most recent models can more realistically represent the dynamics oberserved in markets, it brings with it some serious challenges for the mathematical finance community, prime examples being model simulation and model estimation:

Simulation: Complex SDEs are rarely able to be solved analytically, and therefore exact simulation schemes for generating sample data from these SDEs are not usually available. While basic numerical simulation schemes can be used to generate approximate sample paths from the SDE that converge weakly to the true solution of the SDE, there are cases in which stronger convergence criteria are important, for example, in the pricing of path dependent options by Monte Carlo simulation. In addition, the fairly slow strong convergence rate associated with the most basic numerical simulation schemes associated with SDEs results in much more computational resources being utilised in order to achieve acceptable levels of accuracy of the approximate numerical solution, which can be costly in cases where computational speed is of great importance. The development of alternative numerical simulation schemes with faster rates of strong convergence is therefore a practically important, and mathematically challenging task.

Estimation: As models become more complex, they also become more difficult to estimate from the data. Parameter estimation is essentially a reverse engineering problem, and the more complicated the data generating process is assumed to be,

the more difficult it is to work backwards from the data to infer the best choice model parameter values.

1.2 Thesis overview

In this thesis we will explore aspects of both problems in the context of SDEs that are commonly used in mathematical finance. In Chapter 2 we analyse the strong convergence properties of a numerical approximation scheme associated with a particular class of SDEs that has been used to model equity prices, the stochastic volatility process associated with more complicated equity price models, and interest rates. We provide a proof of the strong convergence of the numerical scheme introduced, and provide a lower bound for the convergence rate. In Chapter 3 we investigate the parameter estimation of a highly nonlinear SDE that has been proposed as a viable model of short term interest rates. We introduce and test a completely new method of parameter estimation, based on the attractive qualities of implicit numerical discretisation schemes. The chapter ends with an in-depth discussion of the problems associated with conventional methods of parameter estimation. In Chapter 4 we introduce approximate Bayesian computation (ABC) methods, which combine a Bayesian approach to model estimation with a novel approximation of the model that allows parameter estimation to be conducted on models for which no information relating to the transition density (and hence the model likelihood) is required. Current approaches to parameter estimation of SDEs typically involve a significant amount of prior analysis before inference is conducted, for example the approach involving an approximation to the model likelihood via Hermite polynomial expansions, developed by Ait-Sähalia (2002), requires that the coefficients of the expansion be worked out in advance of the parameter estimation, which can be a time-consuming task given the complexity

of the expressions for the coefficients. For a detailed review of existing parameter estimation techniques for SDEs in finance, see Hurn et al. (2007). We survey the existing ABC sampling algorithms that have been developed for other applications in the environmental and biological sciences, and then introduce two brand new ABC sampling algorithms that we propose to apply to the task parameter estimation of SDEs used in finance. ABC methods represent a very promising method of parameter estimation, especially in financial applications, because it is a likelihood free method of inference—unlike most standard estimation methods, no analytic results pertaining to the model itself need to be known in order to carry out ABC based parameter estimation. Given the trend towards more complex financial models, reliable estimation techniques that can avoid having to deal explicitly with additional model complexity have great potential in the field of mathematical finance. Our work represents some novel first steps towards develping simulation based, likelihood free estimation techniques with financial applications in mind. From a statistical perspective, we contribute original knowledge via the development and detailed analysis of two new ABC based sampling algorithms. In Chapter 5 we apply the newly developed samplers introduced in Chapter 4 to two test models that are very often seen in the mathematical finance literature; geometric Brownian motion, and the mean-reverting square root process. We run a series of numerical experiments to obtain empirical samples from the posterior distribution of model parameters and compare the results against the analytic distributions to assess the quality of the newly introduced ABC samplers. In Chapter 6 we conclude with a recap of the main results and salient points raised in the body of the thesis, as well as a brief discussion of potential avenues for further research.

Chapter 2

A strong convergence rate for numerical simulation of a CEV-type diffusion process

2.1 Introduction

In computational finance, the dynamics of financial quantities such as equity prices are frequently modelled using stochastic differential equations (SDE). For example, in the well-known Black-Scholes-Merton model, asset prices are assumed to follow a Geometric Brownian Motion process

$$dS(t) = \mu S(t)dt + \sigma S(t)dB(t)$$

where $B(t)$ is a scalar Brownian motion and the rate of return, μ, and the volatility, σ, are assumed to be constant (Hull, 2009). For a more realistic representation of equity dynamics, the volatility of the equity process can also be considered random

(Hull, 2009). Various models have been proposed to describe the volatility process; in this chapter we consider one such model

$$dX(t) = \kappa(\lambda - X(t))dt + \nu X(t)^{\theta} dB(t) \tag{2.1}$$

where $\kappa, \lambda, \nu, \theta$ are assumed to be strictly positive constants. Additionally, equations of this class (linear drift with power law diffusion coefficient) are used to model the dynamics of equity prices. Thirdly, this class of SDE has been proposed as a model for the (instantaneous) short rate of interest, see Chan et al. (1992) and also Nowman (1997) for more details. The case $\theta = 0.5$ corresponds to the familiar 'square-root' process (this model was also used by Cox et al. (1985) as a model of the nominal short interest rate); if $\theta = 1$ equation (2.1) reduces to a linear mean-reverting process.

Typically these models are used in the pricing of financial options (via the usual expected present value calculation under the appropriate probability measure, see Shreve (2004)) and given that explicit formulae often do not exist for calculation of these prices, Monte Carlo techniques are required in order to compute approximate prices numerically. Ensuring the accuracy of such approximations motivates the investigation of the convergence properties of discretisation schemes such as the Euler-Maruyama approximation. A strong convergence result for the case $\theta = 0.5$, which corresponds to an SDE usually referred to as the *square root process* or the *Cox-Ingersoll-Ross (CIR)* model, was given in Dereich et al. (2012); in this chapter we consider the case $0.5 < \theta < 1$, and use a so called drift-implicit Euler approximation to prove the strong convergence of equation (2.1) above.

The difficulty in obtaining strong convergence rates for the approximation of the CIR process is described in Dereich et al. (2012); namely, the non-Lipschitz diffusion coefficient makes conventional error analysis redundant. The same issue arises

with equation (2.1). Dereich et al. (2012) use the so called Lamperti transform to transform the square-root process into a process with unit diffusion, thereby sidestepping the difficulties associated with non-Lipschitz diffusion coefficients mentioned above. We adopt this approach in our more general setting, and derive an equivalent result to that contained in Dereich et al. (2012), via differing arguments, that applies to a more general class of SDEs.

In what follows we will set out some preliminary steps that we take to make the problem more amenable to our analysis, and establish some basic results that will be relied on later in the proof of the main theorem; we then describe the steps in the proof of the main theorem; lastly, we present the results of a simulation study, the aim of which is to empirically demonstrate the strong convergence properties of the drift-implicit Euler approximation.

2.2 Preliminaries

Throughout this chapter we will make repeated use of several well known inequalities; these results are stated explicitly in Appendix 2.6. In order to deal with the non-Lipschitz diffusion coefficient, we transform equation (2.1) into a process with unit diffusion using the Lamperti transform (Iacus, 2008)

$$Y(t) \equiv \int_0^{X(t)} \frac{du}{\nu u^\theta} = \frac{1}{\nu(1-\theta)} X(t)^{1-\theta}, \tag{2.2}$$

for $X(t) > 0$, which gives

$$X(t) = [\nu(1-\theta)Y(t)]^{\frac{1}{1-\theta}}. \tag{2.3}$$

The positivity of $X(t)$, i.e. $\mathbb{P}(0 < X(t) < \infty \text{ for all } t \geq 0) = 1$, is proved in Mao et al. (2006). Note that, as a result of the restriction $0.5 < \theta < 1$, the positivity is preserved under the transformation from $X(t)$ to $Y(t)$. Itô's Lemma shows

$$dY(t) = \Big(\frac{\kappa}{\nu X(t)^{\theta}}(\lambda - X(t)) + \frac{\theta\nu}{2}X(t)^{\theta-1}\Big)dt + dB(t).$$

Substituting (2.3) into the equality above gives the following SDE for $Y(t)$

$$dY(t) = \left(\hat{a}Y(t)^{-p} - \hat{b}Y(t) - \hat{c}Y(t)^{-1}\right)dt + dB(t), \tag{2.4}$$

where

$$\hat{a} = \frac{\kappa\lambda}{(\nu(1-\theta)^{\theta})^{\frac{1}{1-\theta}}}, \qquad \hat{b} = \kappa(1-\theta), \qquad \hat{c} = \frac{\theta}{2(1-\theta)}, \qquad p = \frac{\theta}{1-\theta}.$$

Note that due to the restriction on θ, p lies in the interval $1 < p < \infty$. We now introduce a drift-implicit discretisation of equation (2.4) on which our error analysis will be based

$$y_{k+1} - ay_{k+1}^{-p} + by_{k+1} + cy_{k+1}^{-1} = y_k + \Delta_k B, \quad k = 0, 1, 2, \ldots \tag{2.5}$$

with

$$y_0 \equiv Y(0) = \frac{X(0)^{1-\theta}}{\nu(1-\theta)},$$

where the step size associated with the discretisation is denoted by Δ, and $a = \Delta\hat{a}$, $b = \Delta\hat{b}$, $c = \Delta\hat{c}$. $\Delta_k B = B((k+1)\Delta) - B(k\Delta)$ is a Gaussian increment with zero mean and variance Δ. The idea here is to simulate the process y_k as an approximation to $Y(k\Delta)$ and then transform back from y_k to x_k via (2.3) to obtain an approximation of the process $X(k\Delta)$.

In order to use (2.5) reliably, we need to show that the approximate process yields unique, positive solutions, i.e. we need to show that $y_{k+1} > 0$ for all $y_k > 0$. The following lemma provides this result.

Lemma 1. *Define*

$$F(y) = (1+b)y + cy^{-1} - ay^{-p} \quad \textit{for } y > 0, \tag{2.6}$$

where a, b, c and p are defined as before. If

$$\hat{b}\left(\frac{\hat{a}p(p+1)}{2\hat{c}}\right)^{\frac{p+1}{p-1}} - \frac{1}{2}\hat{a}p(p-1) > 0$$

(where $\hat{a}$, $\hat{b}$, and $\hat{c}$ are, again, defined as before), i.e.

$$\frac{\kappa\lambda^{2(1-\theta)}}{\nu^2} > \left(\frac{\theta(2\theta-1)}{2}\right)^{2\theta-1}(1-\theta)^{4(1-\theta)} \tag{2.7}$$

then for any $z \in (-\infty, \infty)$, the equation $F(y) = z$ has a unique, positive solution for any step size, Δ. Otherwise,

$$\frac{\kappa\lambda^{2(1-\theta)}}{\nu^2} \leq \left(\frac{\theta(2\theta-1)}{2}\right)^{2\theta-1}(1-\theta)^{4(1-\theta)} \tag{2.8}$$

and $F(y) = z$, $z \in (-\infty, \infty)$ yields unique, positive solutions provided $\Delta < \Delta^*$, where

$$\Delta^* = \frac{1}{\kappa(1-\theta)\left(\frac{\theta(2\theta-1)}{2}(1-\theta)^{\frac{4(1-\theta)}{2\theta-1}}\left(\frac{\nu^2}{\kappa\lambda^{2(1-\theta)}}\right)^{\frac{1}{2\theta-1}} - 1\right)} \tag{2.9}$$

Proof The lemma follows if we can show that the function F is continuous, coercive[1], and strictly monotone (Zeidler, 1989). Clearly F is continuously differentiable and coercive on $\mathbb{R}_+$. Taking the derivative of F w.r.t. y we see that

$$\lim_{y\to 0} F'(y) = +\infty \qquad \lim_{y\to\infty} F'(y) = 1 + b > 0.$$

If we can show that $\min F'(y) > 0$ then the function F is strictly increasing and the result follows. Differentiaing $F'(y)$ and setting the result equal to zero, we obtain

$$2cy^{-3} - ap(p+1)y^{-(p+2)} = 0.$$

After rearranging for y, we see that the minimum of $F'(y)$ occurs at the point

$$\tilde{y} = \left(\frac{ap(p+1)}{2c}\right)^{\frac{1}{p-1}} > 0. \tag{2.10}$$

Substituting (2.10) into $F'(y)$ and imposing the condition that $F'(\tilde{y}) > 0$ yields the following inequality

$$\Delta\left(\hat{b}\Big(\frac{\hat{a}p(p+1)}{2\hat{c}}\Big)^{\frac{p+1}{p-1}} - \frac{1}{2}\hat{a}p(p-1)\right) > -\Big(\frac{\hat{a}p(p+1)}{2\hat{c}}\Big)^{\frac{p+1}{p-1}}. \tag{2.11}$$

The RHS of (2.11) is strictly negative. Keeping in mind that $\Delta > 0$ we see that two situations arise:

1. $\hat{b}\left(\frac{\hat{a}p(p+1)}{2\hat{c}}\right)^{\frac{p+1}{p-1}} - \frac{1}{2}\hat{a}p(p-1) > 0$

   In this case, Δ must be strictly greater than a negative number, which is obviously always true. Therefore (2.6) always yields unique positive solutions

[1]In this instance, a function is coercive if $\lim_{y\to\infty} F(y) = \infty$ and $\lim_{y\to 0} F(y) = -\infty$, or $\lim_{y\to\infty} F(y) = -\infty$ and $\lim_{y\to 0} F(y) = \infty$.

for all step sizes.

2. $\hat{b}\left(\frac{\hat{a}p(p+1)}{2\hat{c}}\right)^{\frac{p+1}{p-1}} - \frac{1}{2}\hat{a}p(p-1) \leq 0$

   In this case, rearranging (2.11) for Δ yields, after some algebra, the condition: $\Delta < \Delta^*$, where Δ^* is given by (2.9). This completes the proof. □

We now prove that the function $f : (0, \infty) \mapsto \mathbb{R}$, defined by

$$f(y) = \hat{a}y^{-p} - \hat{b}y - \hat{c}y^{-1} \tag{2.12}$$

is one-sided Lipschitz-continuous. Note that the function f is of the same form as the drift-coefficient in (2.4). Higham et al. (2002) used this feature to control error propagation using a drift-implicit approximation in the context of non-linear SDEs with polynomial drift coefficients.

Lemma 2. *If* (2.7) *is satisfied, then the following holds*

$$(x - y)\left(f(x) - f(y)\right) < 0 \tag{2.13}$$

for $x, y > 0$*, where* f *is defined by* (2.12) *above.*

Proof By inspection, f is clearly continuously differentiable and therefore, by the mean value theorem, if $f'(y) < 0$, $\forall y > 0$ then f is one-sided Lipschitz-continuous and the result follows. After differentiating f w.r.t. y and rearranging, we see that $f'(y)$ has a maximum at

$$\tilde{y} = \left(\frac{\hat{a}p(p+1)}{2\hat{c}}\right)^{\frac{1}{p-1}} > 0. \tag{2.14}$$

Substituting (2.14) into the expression for $f'(y)$ and rearranging we arrive at

$$f'(\tilde{y}) = \tilde{y}^{-(p+1)}\Big\{\frac{1}{2}\hat{a}p(p-1) - \hat{b}\bigg(\frac{\hat{a}p(p+1)}{2\hat{c}}\bigg)^{\frac{p+1}{p-1}}\Big\}.$$

We see that f is one-sided Lipschitz iff the term in parenthesis on the RHS of the above equality is strictly less than zero. As outlined in Lemma 1, this corresponds to condition (2.7). The proof is complete. □

Note that although the numerical approximation (2.5) may yield unique positive solutions under condition (2.7) or (2.8), the one-sided Lipschitz condition on which our proof relies is only satisfied under condition (2.7). Henceforth we assume condition (2.7) is satisfied.

Condition (2.7) imposes a certain restriction on the parameters of the original diffusion process, (2.1). We note that reasonable parameter values typically satisfy this condition. When $\theta \to 0.5$ condition (2.7) reduces to the following:

$$4\kappa\lambda > \nu^2$$

which closely resembles the so-called Feller condition[2] that guarantess the positivity of solutions to the Cox-Ingersoll-Ross process (Brigo and Mercurio, 2006). Note that our condition on the parameter values guarantees unique positive solutions for the numerical approximation to our original process, (2.1), whereas the Feller condition guarantees the positivity of the analytic solution to the Cox-Ingersoll-Ross process; therefore there is no reason to expect our condition to reduce to the Feller condition in the limit $\theta \to 0.5$

[2]The Feller condition is as follows: $2\kappa\lambda > \nu^2$.

2.3 Main Result

In the interests of clarity, we now state the main result that we will prove in the course of this section:

Theorem 1. *Let $x_0 > 0$ and $T > 0$. Then for all $q \geq 1$, there exists a constant $K_q > 0$ such that*

$$\left(\mathbb{E} \sup_{0 \leq t \leq T} |X_t - \bar{x}_t|^q\right)^{\frac{1}{q}} \leq K_q \cdot \sqrt{|log(\Delta)|} \cdot \sqrt{\Delta}$$

for all $\Delta \in (0, 1/2]$, where $\bar{x}_t = x_k = [\nu(1-\theta)y_k]^{\frac{1}{1-\theta}}$, $t \in [k\Delta, (k+1)\Delta))$, $k \geq 0$ and y_k is defined by (2.5).

Note that $\bar{x}_t$ is the piecewise constant interpolation for $t \in [k\Delta, (k+1)\Delta)$ and hence $\bar{x}_t$ is an $\mathcal{F}_t$-adapted process. This is different from that in Dereich et al. (2012) where they use the linear interpolation between x_k and x_{k+1}, meaning $\bar{x}_t$ is not $\mathcal{F}_t$-adapted.

We now work through the steps involved in proving this result.

2.3.1 Moment Bounds of $X(t)$

To begin, we prove that the moments and inverse moments of the original process, $X(t)$, are bounded in finite time.

Lemma 3. *For all $p \in (-\infty, \infty)$*

$$\sup_{0 \leq t \leq T} \mathbb{E}|X(t)|^p < \infty$$

Proof First, let us consider the case where $p \geq 2$. Let $V(X(t)) = X(t)^p$. A simple application of Itô's Lemma shows that the infinitesimal generator (see Øksendal (2007)) denoted by $\mathcal{L}V(X(t))$ is given by [3]

$$\mathcal{L}V(X) = \kappa\lambda p X^{p-1} - \kappa p X^p + \frac{1}{2}\sigma^2 p(p-1)X^{p-2+2\theta}$$

which is clearly bounded from above by some constant, K. This observation allows upper bounds for the higher moments of $X(t)$ to be established by virtue of the fact that

$$\mathbb{E}V\left(X(t)\right) = V(X_0) + \mathbb{E}\int_0^t \mathcal{L}V(s)\mathrm{d}s.$$

This result is extended to $0 \leq p < 2$ by noting that $X^p \leq 1 + X^2$ for any $X > 0$. The case where $p < 0$ is dealt with by taking $V(X(t)) = X(t)^{-p}$, $p > 0$, and applying the infinitesimal generator, as before, to obtain

$$\mathcal{L}V(X) = -\kappa\lambda p X^{-(p+1)} + \kappa p X^{-p} + \frac{1}{2}\sigma^2 p(p-1)X^{-p-2+2\theta},$$

which is also bounded from above by a constant. Upper bounds for the inverse moments follow by similar reasoning used for the case $p \geq 0$. □

2.3.2 Smoothness of the Transformed Process, $Y(t)$

In this subsection, we will prove a smoothness result for the process obtained by the transformation (2.2). For the remainder of this chapter, any unimportant constant shall be denoted by C for brevity.

[3] Note that we have omitted the time dependence of X for clarity.

Lemma 4. *Let $T > 0$. Then, for all $q \geq 1$ we have*

$$\mathbb{E}\left(\sup_{\substack{0 \leq s < t \leq T \\ |t-s| \leq \Delta}} |Y(t) - Y(s)|^q \right) \leq C \cdot (|log(\Delta)|\Delta)^{\frac{q}{2}}$$

and

$$\mathbb{E} \sup_{0 \leq t \leq T} |Y(t)|^q < \infty$$

for $\Delta \in (0, 0.5]$.

Proof We have[4]

$$Y_t - Y_s = \hat{a} \int_s^t Y_u^{-p} \mathrm{d}u - \hat{b} \int_s^t Y_u \mathrm{d}u - \hat{c} \int_s^t Y_u^{-1} \mathrm{d}u + B_t - B_s.$$

By an application of the Hölder inequality (see Mao (2008)) followed by the discrete Hölder inequality (see Mao and Yuan (2006)) one obtains

$$|Y_t - Y_s|^q \leq C \cdot \left(\Big(\int_s^t Y_u^{-2p} \mathrm{d}u \Big)^{\frac{q}{2}} + \Big(\int_s^t Y_u^2 \mathrm{d}u \Big)^{\frac{q}{2}} + \Big(\int_s^t Y_u^{-2} \mathrm{d}u \Big)^{\frac{q}{2}} + |B_t - B_s|^q \right). \tag{2.15}$$

After taking expectations of both sides above, what remains is to bound the first three terms in parenthesis; an application of Itô's Lemma, and Theorem 2.12 in Mao and Yuan (2006) is sufficient for this purpose (see Appendix 2.7 at the end of this chapter for an example of this calculation). The first result follows by using the inequality (Müller-Gronbach, 2002)

$$\mathbb{E}\left(\sup_{\substack{0 \leq s < t \leq T \\ |t-s| \leq \Delta}} |B_t - B_s|^q \right) \leq C \cdot (|log\Delta|\Delta)^{\frac{q}{2}}$$

[4]We abbreviate the time argument in what follows, i.e. $Y(t) \equiv Y_t$.

for $\Delta \in (0, 0.5]$. The second assertion follows by a similar argument, and by noting that $\mathbb{E}\sup_{0\leq t\leq T}|B_t|^q \leq K_{T,q}$ which can be obtained using the Burkholder-Davis-Gundy inequality (Mao, 2008). □

2.3.3 Error Bound for Implicit Euler Scheme for Y(t)

We now derive an upper bound for the strong error of the drift-implicit approximate solutions of equation (2.4).

Lemma 5. *For $T > 0$, there exists a constant, C, such that*

$$\mathbb{E}\Big(\sup_{k=0,\ldots,\lceil T/\Delta\rceil}|Y_{k\Delta} - y_k|^q\Big) \leq C \cdot (|log(\Delta)|\Delta)^{q/2}$$

for $\Delta \in (0, 0.5]$, where $\lceil T/\Delta\rceil$ denotes the smallest integer which is no less than T/Δ.

Proof Let

$$e_k = Y_{k\Delta} - y_k$$

be the local error introduced by the approximation scheme. We have the following recursive relationship for the error at time $k\Delta$

$$e_0 = 0$$
$$e_{k+1} = e_k + (f(Y_{(k+1)\Delta}) - f(y_{k+1}))\Delta + r_k$$

with

$$r_k = -\int_{k\Delta}^{(k+1)\Delta}(f(Y_{(k+1)\Delta}) - f(Y_t))\mathrm{d}t.$$

Utilising the one-sided Lipschitz result (2.13), we have that

$$\begin{aligned} e_{k+1}^2 &= e_{k+1}e_k + e_{k+1}(f(Y_{(k+1)\Delta}) - f(y_{k+1}))\Delta + e_{k+1}r_k \\ e_{k+1}^2 &\leq e_{k+1}(e_k + r_k) \\ |e_{k+1}| &\leq |e_k| + |r_k|. \end{aligned}$$

Making use of the recursive nature of this inequality, we arrive at

$$\sup_{k=0,\ldots,\lceil T/\Delta\rceil} |Y_{k\Delta} - y_k| \leq \sum_{k=0}^{\lceil T/\Delta\rceil - 1} |r_k|.$$

As in Dereich et al. (2012), we need to bound $|r_k|$

$$|r_k| = \left| \int_{k\Delta}^{(k+1)\Delta} (f(Y_{(k+1)\Delta}) - f(Y_t))\mathrm{d}t \right|.$$

Note that

$$|f(u) - f(v)| \leq C \cdot \left(u^{-(p+1)} + v^{-(p+1)} + 1 + u^{-2} + v^{-2}\right) \cdot |u - v|$$

for $u, v > 0$, where f is, again, defined by (2.12). Thus

$$\begin{aligned} \sum_{k=0}^{\lceil T/\Delta\rceil - 1} |r_k| \leq C \cdot \sup_{\substack{0\leq s<t\leq T \\ |t-s|\leq\Delta}} |Y_t - Y_s| \Bigg(1 + \Delta \sum_{k=0}^{\lceil T/\Delta\rceil - 1} Y_{(k+1)\Delta}^{-(p+1)} \\ + \Delta \sum_{k=0}^{\lceil T/\Delta\rceil - 1} Y_{(k+1)\Delta}^{-2} + \int_0^T Y_s^{-(p+1)}\mathrm{d}s + \int_0^T Y_s^{-2}\mathrm{d}s \Bigg). \end{aligned}$$

Raising both sides to the power q and applying the discrete Hölder inequality, we obtain

$$\left(\sum_{k=0}^{\lceil T/\Delta\rceil-1}|r_k|\right)^q \leq C\cdot\sup_{\substack{0\leq s<t\leq T\\ |t-s|\leq\Delta}}|Y_t-Y_s|^q\left(1+\Delta\sum_{k=0}^{\lceil T/\Delta\rceil-1}Y_{(k+1)\Delta}^{-q(p+1)}\right.$$
$$\left.+\Delta\sum_{k=0}^{\lceil T/\Delta\rceil-1}Y_{(k+1)\Delta}^{-2q}+\int_0^T Y_s^{-q(p+1)}\mathrm{d}s+\int_0^T Y_s^{-2q}\mathrm{d}s\right).$$

We now take expectations of both sides and apply the Hölder inequality once more. We have already shown each of the four terms in parenthesis on the RHS to be bounded from above from earlier (see section 3.2). The final result follows from the smoothness result derived in section 3.2. □

2.3.4 Bounded Moments of the Approximation, y_k

Lemma 6. *Let $\Delta > 0$ and $T > 0$. Then for all $r \geq 1$ we have*

$$\mathbb{E}\sup_{k=0,..,\lceil T/\Delta\rceil}|y_k|^r < \infty$$

Proof Note that y_k are random variables with respect to some probability space and corresponding sigma-algebra $\{\Omega, \mathcal{F}\}$. To prove the boundedness of the moments of $y_k, k = 0, .., \lceil T/\Delta\rceil$, we define the following partition of Ω:

$$A_{k+1} = \{\omega : y_{k+1}(\omega) \leq \Delta\} \in \Omega,$$
$$A_{k+1}^c = \{\omega : y_{k+1}(\omega) > \Delta\} \in \Omega.$$

1. If $\omega \in A_{k+1}$:

$$y_{k+1} \leq \Delta.$$

2. If $\omega \in A_{k+1}^c$:

$$y_{k+1} = y_k + f(y_{k+1})\Delta + \Delta B_k < y_k + f(\Delta)\Delta + \Delta B_k.$$

Therefore we have that, for all $\omega \in \Omega$

$$y_{k+1} \leq \Delta + y_k + f(\Delta)\Delta + \Delta B_k$$

which implies that

$$y_{k+1} \leq T(1 + f(\Delta)) + y_0 + B_{k+1}.$$

Utilising once more the discrete Hölder inequality we have that

$$|y_{k+1}|^r \leq C \cdot (y_0^r + T^r(1 + f(\Delta))^r + |B_{k+1}|^r).$$

After taking the expectation of the supremum over k, we see by inspection that the first two terms on the RHS are obviously bounded. The Burkholder-Davis-Gundy inequality provides the bound for the third term, and the proof is complete. $\square$

2.3.5 Error Bound for the Drift-Implicit Approximation of $X(t)$

We are now in a position to prove the main result contained within this chapter, which is reproduced below for the reader's convenience.

Theorem 1. *Let $x_0 > 0$ and $T > 0$. Then for all $q \geq 1$, there exists a constant $K_q > 0$ such that*

$$\left(\mathbb{E} \sup_{0 \leq t \leq T} |X_t - \bar{x}_t|^q\right)^{\frac{1}{q}} \leq K_q \cdot \sqrt{|log(\Delta)|} \cdot \sqrt{\Delta}$$

for all $\Delta \in (0, 1/2]$.

Proof Denote by $\bar{X}_t$ the piecewise constant interpolation of the mean reverting θ-process with stepsize $\Delta > 0$[5], i.e.

$$\bar{X}_t = X_{k\Delta}, \qquad t \in [k\Delta, (k+1)\Delta).$$

Now recall the transformation defined in equation (2.2)

$$Y(t) = \frac{1}{\nu(1-\theta)} X(t)^{1-\theta}.$$

Using this and a well-known inequality (see Appendix 2.6), we obtain the following relationship

$$|X_t - X_s|^q = C \cdot |Y_t^r - Y_s^r|^q \leq C \cdot (Y_t^{r-1} + Y_s^{r-1})^q |Y_t - Y_s|^q$$

where $r = \frac{1}{1-\theta}$ and $q \geq 1$. This, along with Hölder's inequality, implies that

$$\mathbb{E}\left(\sup_{\substack{0\leq s<t\leq T \\ |t-s|\leq\Delta}} |X_t - X_s|^q \right) \leq C \cdot \left(\mathbb{E} \sup_{\substack{0\leq s<t\leq T \\ |t-s|\leq\Delta}} (Y_t^{r-1} + Y_s^{r-1})^{2q} \right)^{\frac{1}{2}} \left(\mathbb{E} \sup_{\substack{0\leq s<t\leq T \\ |t-s|\leq\Delta}} |Y_t - Y_s|^{2q} \right)^{\frac{1}{2}}.$$

The RHS is bounded following Lemma 4.[6] Thus

$$\mathbb{E}\left(\sup_{\substack{0\leq s<t\leq T \\ |t-s|\leq\Delta}} |X_t - X_s|^q \right) \leq C \cdot \big(|log(\Delta)|\Delta \big)^q \tag{2.16}$$

for $q \geq 1$.

Next, note that

$$X_t - \bar{X}_t = X_t - X_{k\Delta},$$

[5]Note that $\bar{x}_t$ is taken to represent the equivalent piecewise constant interpolation of the numerical solution, x_k.

[6]Provided that the stepsize $\Delta \in (0, 0.5]$.

where $t \in [k\Delta, (k+1)\Delta)$.

Given that

$$\sup_{t\in[k\Delta,(k+1)\Delta)} |X_t - X_{k\Delta}| \leq \sup_{\substack{0\leq s<t\leq T \\ |t-s|\leq\Delta}} |X_t - X_s|$$

we have that

$$\sup_{0\leq t\leq T} |X_t - \bar{X}_t| \leq \sup_{\substack{0\leq s<t\leq T \\ |t-s|\leq\Delta}} |X_t - X_s|. \tag{2.17}$$

Finally, we use the fact that

$$X_t - \bar{x}_t = X_t - \bar{X}_t + \bar{X}_t - \bar{x}_t$$

to arrive at

$$\sup_{0\leq t\leq T} |X_t - \bar{x}_t|^q \leq C \left(\sup_{0\leq t\leq T} |X_t - \bar{X}_t|^q + \sup_{k=0,..,\lceil \frac{T}{\Delta} \rceil} |X_{k\Delta} - x_k|^q \right).$$

After taking expectations of both sides, we use (2.16) and (2.17) to bound the first term on the RHS.

To bound the second term, first note that

$$|X_{k\Delta} - x_k| = C \cdot |Y^r_{k\Delta} - y^r_k| \leq C \cdot (Y^{r-1}_{k\Delta} + y^{r-1}_k)|Y_{k\Delta} - y_k|, \quad r = \frac{1}{1-\theta}$$

which follows from (2.3) and an elementary inequality (which is reproduced in Appendix 2.6 for the reader's convenience). We therefore have that

$$\mathbb{E} \sup_{k=0,..,\lceil \frac{T}{\Delta} \rceil} |X_{k\Delta} - x_k|^q \leq C \cdot \mathbb{E} \left(\sup_{k=0,..,\lceil \frac{T}{\Delta} \rceil} (Y^{r-1}_{k\Delta} + y^{r-1}_k)^q \cdot \sup_{k=0,..,\lceil \frac{T}{\Delta} \rceil} |Y_{k\Delta} - y_k|^q \right).$$

After an application of Hölder's inequality, we arrive at our result by virtue of Lemmas 4, 5 and 6. The proof of Theorem 1 is now complete. □

2.4 Simulation Study

In this section we produce some results to empirically demonstrate the strong convergence of the so-called drift-implicit Euler approximation to equation (2.1). All simulations of the process (2.1) were carried out using reasonable values for the drift and volatility parameters[7], and with $\theta = 0.75$. The steps taken in this analysis were as follows:

1. A sample path of the process (2.4) over one time interval was produced using the explicit Euler-Maruyama algorithm, with a time-step of 10^{-6}; this sample path was used to approximate the exact process given that (2.4) cannot be sampled exactly due to the transition density of the process being unknown.

2. An approximate path was then generated from the same Brownian motion used in step 1, but using a larger time-step. The approximate path was generated using the drift-implicit Euler approximation. Note that this approximation method involves solving (2.5) for y_{t+1} at each time step; a simple Newton-Raphson algorithm was implemented to carry out this task[8].

3. The 'exact' and approximate sample paths were then transformed into the original process (2.1) using the Lamperti transform (2.2). A sample path of the 'exact' process and the corresponding approximation is illustrated in 2.1.

4. The square of the absolute difference between the end-points of the exact and approximate sample paths was recorded. This error criterion was used as a proxy for the supremum error over the whole path in order to reduce memory usage and therefore reduce CPU run-time.

[7]Parameter values were as follows: $\kappa = 0.2$, $\lambda = 0.09$ and $\nu = 0.08$.
[8](2.5) was solved for Y_{t+1} to within an accuracy of 10^{-7}.

5. Steps 1, 2, 3 and 4 were repeated 1000 times and the square root of the sample mean taken to represent the strong error of the drift-implicit approximation.

6. Steps 1, 2, 3,4 and 5 were repeated for 4 different step-sizes: $\Delta = 10^{-5}, \Delta = 10^{-4}, \Delta = 10^{-3}$and $\Delta = 10^{-2}$.

The results of the above experiment are illustrated in 2.2.

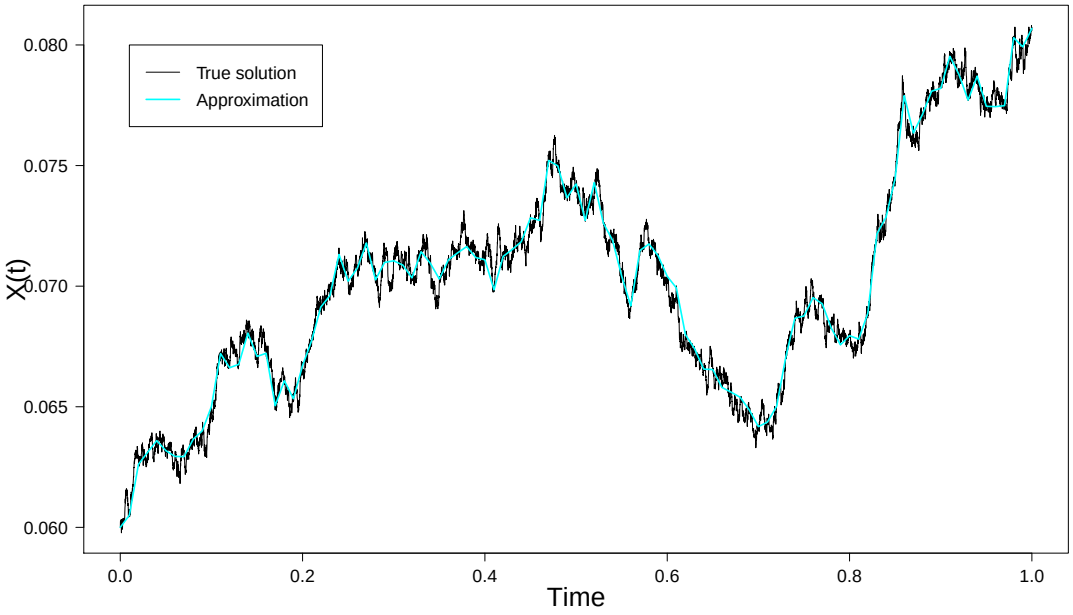

Figure 2.1: A sample path of the true solution, $X(t)$, and the drift-implicit approximation, $x(t)$, based on the same Brownian sample path.

Note that the drift-implicit approximation appears to converge strongly to the true solution with a rate of 1, rather than 0.5 as indicated in our earlier analysis. Although this finding does not invalidate the analysis presented in this chapter, this strong convergence rate is a much stronger result than that indicated by our analysis; further work may be required to fully understand this feature.

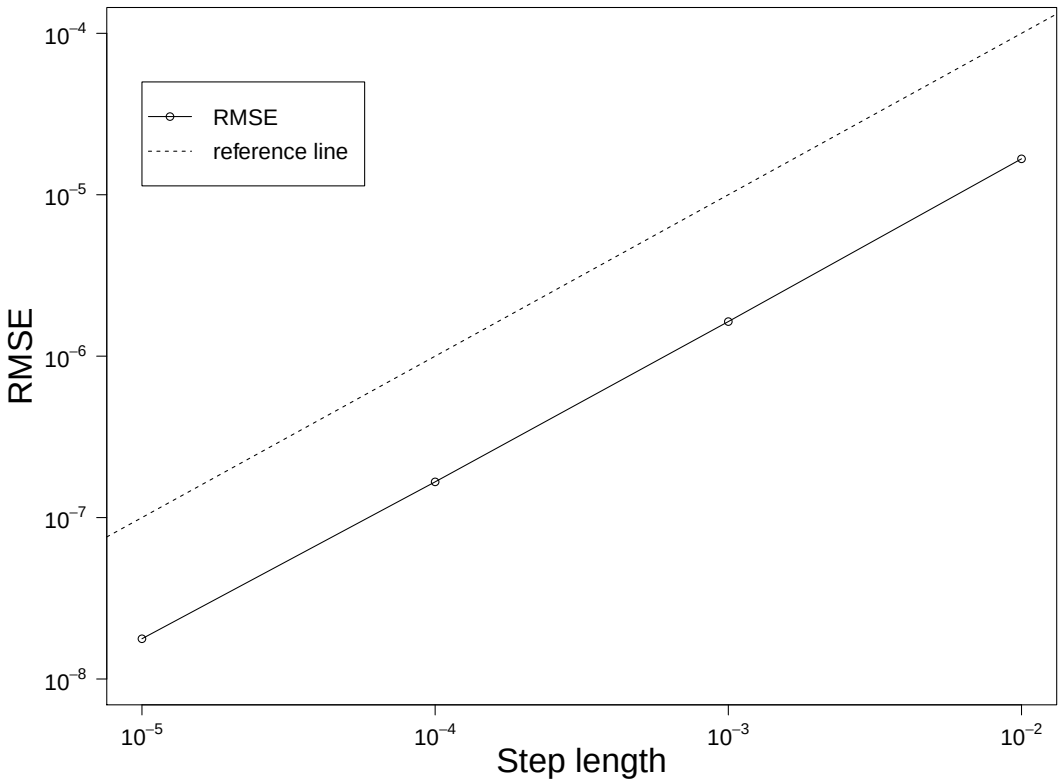

Figure 2.2: The strong error plot for the drift-implicit Euler approximation to the mean reverting CEV process. The dashed line of slope 1 is the reference line.

2.5 Discussion & summary

In this chapter we provided a lower bound for the strong convergence rate associated with a numerical approximation of a class of SDEs commonly known as CEV processes. We began by introducing some preliminary steps that made the problem simpler to analyse, which was then followed by proving some basic results concerning the numerical approximation to the true process, which were then used to prove the main result in this chapter. The chapter ends with a numerical experiment to illustrate the main result. As mentioned earlier, the empirical results seem to indicate a faster convergence rate for the numerical approximation used here than

the theoretical analysis suggests. Note that since the results contained within this chapter were produced (during 2011), other parties have independently derived a stronger convergence rate result for this class of processes (Neuenkirch and Szpruch, 2014) and subsequently published these results. The result proved by Neuenkirch and Szpruch (2014) confirms the empirical work carried out in this chapter—that the strong convergence rate of the numerical approximation to (2.1) is equal to one. Having analysed a typical problem associated with generating data from stochastic models commonly used in finance, demonstrating that a judicial choice of numerical approximation scheme can yield accurate approximations to the strong solutions of such processes, in the forthcoming chapters we turn our attention to another important area in the field of mathematical finance: the parameter inference of stochastic models.

2.6 Appendix A

The following are a list of some common results that have been used in this chapter:

- Hölder's Inequality

$$|\mathbb{E}XY| \leq (\mathbb{E}|X|^p)^{\frac{1}{p}}(\mathbb{E}|Y|^q)^{\frac{1}{q}}$$

  if $p > 1$, $1/p + 1/q = 1$, provided $\mathbb{E}|X|^p < \infty$ and $\mathbb{E}|Y|^q < \infty$.

- Discrete Hölder's Inequality

$$\Big|\sum_{i=1}^{k} a_i b_i\Big| \leq \Big(\sum_{i=1}^{k} |a_i|^p\Big)^{1/p}\Big(\sum_{i=1}^{k} |b_i|^q\Big)^{1/q}$$

  if $p, q > 1$, $1/p + 1/q = 1$, $k \geq 2$, and $a_i, b_i \in \mathbb{R}$.

- Power inequality

$$|u^p - v^p| \leq p|u - v|(u^{p-1} + v^{p-1}), \quad \forall u, v > 0 \text{ and } p \in (-\infty, \infty).$$

- Moment inequality for stochastic integrals

  Let $p \geq 2$. Let g be a process adapted to the filtration generated by a one-dimensional Brownian motion, $B(t)$, such that

$$\mathbb{E}\int_u^v |g(s)|^p \mathrm{d}s < \infty.$$

  Then

$$\mathbb{E}\Big|\int_u^v g(s)\mathrm{d}B(s)\Big|^p \leq \Big(\frac{p(p-1)}{2}\Big)^{\frac{p}{2}}(v-u)^{\frac{p-2}{2}}\mathbb{E}\int_u^v |g(s)|^p \mathrm{d}s.$$

2.7 Appendix B

In the proof of Lemma 4 we had to bound the expectation of an integral of the stochastic process, $Y(t)$; in this section we work through the steps required to do this. We bound one of the terms to illustrate the method - the other two terms are handled in a similar manner. Consider the first term inside parenthesis on the RHS of (2.15), the integral is raised to the power $q/2$; we can show that the expectation of this term is bounded by showing that higher moments are bounded. Therefore, we will bound the following term

$$\mathbb{E}\Big(\int_s^t Y_u^{-2p}\mathrm{d}u\Big)^{\delta}, \quad \delta \geq 1.$$

Firstly, we transform back from $Y(t)$ to $X(t)$ using (2.3) so that we may use some of the results for the process $X(t)$ derived in earlier sections of this chapter. After transforming to $X(t)$, we apply Itô's lemma to obtain the following

$$Y_u^{-2p} \equiv X_u^{-2\theta} = X_0^{-2\theta} + \int_0^u \mathcal{L}(X_v^{-2\theta})\mathrm{d}v - 2\theta\sigma\int_0^u X_v^{-(1+\theta)}\mathrm{d}B_v.$$

From Lemma 3 we know that the infinitesimal generator under the first integral on the RHS is bounded from above by some constant, K. Taking the integral over u, we have that

$$\int_s^t X_u^{-2\theta}\mathrm{d}u \leq X_0^{-2\theta}(t-s) + \frac{1}{2}K(t^2-s^2) - 2\theta\sigma\int_s^t\int_0^u X_v^{-(1+\theta)}\mathrm{d}B_v\mathrm{d}u.$$

After reversing the order of the double integral, raising both sides to the power δ, using the discrete Hölder inequality (see Appendix 2.6) and taking the expectation we have that

$$\mathbb{E}\Big(\int_s^t X_u^{-2\theta}\mathrm{d}u\Big)^{\delta} \leq CX_0^{-2\theta\delta}(t-s)^{\delta} + C(t^2-s^2)^{\delta} + C\mathbb{E}\Big(\int_s^t X_v^{-(1+\theta)}(t-v)\mathrm{d}B_v\Big)^{\delta}$$

where, as before, C represents any unimportant constant. Only the last term on the RHS requires further analysis (the first two are clearly finite for $t, s \leq T$). Using a moment inequality for stochastic integrals (which is stated in Appendix 2.6 for convenience), we see that the final term on the RHS is also finite.

Chapter 3

Drift-implicit pseudo-maximum-likelihood estimation of the parameters of the Ait-Sahalia short rate model

3.1 Introduction

In the previous chapter we focussed on the problem of simulating sample paths from a particular class of SDEs, and demonstrated the convergence properties of a particular numerical approximation to the underlying class of SDEs considered. In this chapter we switch focus to the problem of parameter estimation, still in the context of SDEs used in financial applications. Parameter estimation is a subject of great practical interest for industry practitioners, as well as being a very challenging problem from a researcher's standpoint. For quantitative analysts, choosing a model that is able to reproduce realistic dynamics of economic variables observed in markets is a challenging task, but even once such a model has been chosen there still remains the challenge of calibrating the model to real data. In recent years various new models have been proposed for different financial quantities, e.g.

interest rates, stock prices, exchange rates, that are increasingly able to capture a wider variety of qualitative behaviours that one might observe in market variables; however, increasing the complexity of models, while perhaps facilitating more realistic representations of key variables, also leads to a correspondingly more difficult parameter inference problem. Ensuring that 'good' model parameter values are chosen is extremely important in practice; poorly chosen parameter values can lead to materially different behaviour being exhibited by the model generated data, relative to the observed market data, which in turn can lead to errors in the pricing of financial derivatives and other contingent claims, as well as potentially leading to insufficient capital being set aside to cope with extreme risk events. For these reasons it is important to develop robust parameter estimation techniques for models used in finance.

In this chapter, we focus our attention on a six-parameter Itô diffusion process that conveniently nests a family of SDEs frequently used to model the dynamics of the instantaneous short rate of interest, a key state variable for pricing securities, such as bonds, whose price depends on the term structure of interest rates. There currently exists a relatively large number of parameter estimation techniques in the literature that could be applied to this particular type of model (a univariate diffusion process); however, most of these techniques either involve some non-trivial work to be carried out prior to the parameter estimation (for example, in order to implement the Hermite polynomial likelihood approximation presented by Ait-Sähalia (2002), one must first determine the coefficients of the expansion which can be a challenging task due to the complexity of the terms), or are computationally intensive and time-consuming. In this chapter we investigate a new technique for estimating the six-parameter SDE which we will label the *Ait-Sähalia* short rate model, or AS model for brevity. The technique we develop is closely related to

the Pseudo-Maximum-Likelihood (PML) approach in which the SDE is discretised using an explicit Euler-Maruyama approximation, effectively imposing on the model the assumption that the transition density of the process is Gaussian. It is known that the PML method is too crude for many SDEs, as parameter estimates produced via this method can possess significant biases, especially in the parameters appearing in the drift coefficient of the diffusion process. However, the method is very straightforward to implement and can be useful for finding crude parameter estimates that are then used as a starting point for more complex estimation techniques. Our new technique involves making a drift-implicit Euler-Maruyama discretisation of the AS SDE and deriving a transition density for the resulting approximation. Given the often attractive properties of drift-implicit discretisations of SDEs in the context of numerical simulation, which will be discussed below, our goal is to investigate whether the corresponding parameter estimation utilising the drift-implicit discretisation (which we will label DI-PML in what follows), yields better parameter estimates, both in terms of closeness to the 'true' parameter values, and in terms of the size of the confidence intervals of the estimates, than the traditional PML estimates.

In the following sections we introduce the model being studied, along with some interesting features of the model; this is followed by a brief survey of the traditional PML approach to parameter estimation; we then introduce the new technique we developed to estimate the AS model parameters, and derive the transition density associated with the drift-implicit approximation of the SDE; we then carry out a numerical experiment to compare the parameter estimates obtained via our method against the estimates obtained via the traditional PML approach; the chapter ends with a detailed discussion of the results along with what implications they have

for parameter inference in mathematical finance in general, and a recap of the key points covered in the chapter.

3.2 The model

The SDE that we will focus our analysis on in this chapter is the Ait-Sähalia short rate SDE, which is given by

$$\mathrm{d}r_t = (\alpha_{-1}r_t^{-1} - \alpha_0 + \alpha_1 r_t - \alpha_2 r_t^2)\mathrm{d}t + \sigma r_t^{\gamma}\mathrm{d}B_t, \quad r_0 = R_0 \in (0, \infty) \tag{3.1}$$

where B_t is a scalar Brownian motion. $\alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \sigma$ are positive constants[1], and $1 < \gamma < 1.5$, which we collectively represent as a vector of parameters θ. We assume that the initial value, R_0, is known. The upper bound on γ is required in order to bound the moments of the process; for a derivation of this bound see Szpruch et al. (2011). This form of SDE was first considered in Ait-Sähalia (1996)[2]. This particular SDE was the focus of the analysis conducted in Szpruch et al. (2011), in the context of the design of numerical approximation schemes for the simulation of sample traces from the SDE. In Szpruch et al. (2011), the authors demonstrate that an appropriate implicit numerical approximation can preserve the positivity of the process (a key qualitative feature of the strong solution to the SDE), as well as demonstrating the strong convergence of the numerical approximation to (3.1). The qualitative advantage of using an implicit numerical scheme, namely that positivity of the approximate process is preserved, might indicate that there are potentially additional benefits to be had by considering alternative parameter

[1]The drift parameters are bounded from below in order to ensure the existence and uniqueness of solutions to (3.1).

[2]In fact, the form of diffusion coefficient considered here is slightly less general than the diffusion coefficient introduced in Ait-Sähalia (1996); nonetheless, all commonly used univariate models for the short rate of interest are encapsulated in the SDE considered in this chapter.

estimation techniques, the basis for which is some form of implicit discretisation of (3.1).

3.3 Pseudo-maximum likelihood estimation

One of the simplest methods of parameter estimation is to simply discretise the SDE by using the explicit Euler-Maruyama (EM) approximation which, in effect, imposes on the model the assumption that the transition density of the process is Gaussian. Substituting the Gaussian approximate transition density into the expression for the model likelihood of a diffusion process and optimising over the range of the parameter space in which we are interested yields the PML parameter estimates. The EM approximation of (3.1) is as follows

$$r_{k+1} - r_k = (\alpha_{-1} r_k^{-1} - \alpha_0 + \alpha_1 r_k - \alpha_2 r_k^2)\Delta t + \sigma r_k^{\gamma} \Delta B_k, \tag{3.2}$$

where $r_k = r_{k\Delta t} = r(k\Delta t)$ and $\Delta B_k = B((k+1)\Delta t) - B(k\Delta t)$. This approximation implies that the transition density of the approximate process, i.e. the probability density function of r_{k+1}, conditioned on r_k and some paramter value θ, is given by

$$f^e(r_{k+1}|r_k, \theta) \sim \mathcal{N}(\tilde{\mu}(r_k, \theta), \tilde{\sigma}^2(r_k, \theta)), \tag{3.3}$$

where $\tilde{\mu}(r_k, \theta) = (\alpha_{-1} r_k^{-1} - \alpha_0 + \alpha_1 r_k - \alpha_2 r_k^2)\Delta t + r_k$ and $\tilde{\sigma}^2(r_k, \theta) = \sigma^2 r_k^{2\gamma} \Delta t$. Itô diffusions are, by construction, Markov processes, which allows one to write down the model likelihood, $L(\underline{r}|\theta)$, of such a process as a product of the transition densities evaluated at the observed values of the time-series that we are modelling (here labelled $\underline{r}$), i.e.

$$L(\underline{r}|\theta) = \prod_{k=0}^{N-1} f^e(r_{k+1}|r_k, \theta),$$

where $f^e(r_{k+1}|r_k, \theta)$ is defined in (3.3). The PML parameter estimates are then obtained by maximising this quantity with respect to the parameters θ. We have already noted that the resulting PML parameter estimates can exhibit significant biases away from the true parameter values when the time increment between observations is finite. See Iacus (2008), page 122, for a discussion and illustration of the discretisation bias associated with PML estimates. For this reason, this particular method of parameter estimation is not terribly reliable, especially for SDEs that possess nonlinear drift and/or diffusion coefficients; however, the method is still useful for situations in which the time step between observations is very small (a regime for which the Gaussian transition density is a reasonable assumption), and also in situations where a crude parameter estimate is needed to initialise a more complicated method of estimation, e.g. MCMC based estimation—a numerical technique that will be discussed in detail later in this thesis.

3.4 The drift-implicit approximation

In the previous section we introduced the Pseudo-Maximum-Likelihood method of parameter estimation, in which the assumption of a Gaussian transition density for the process under consideration is used to derive a tractable approximation to the model likelihood which, in turn, allows parameter estimates to be obtained from this pseudo likelihood. Underlying the PML method is the explicit EM discretisation of the SDE. In this section, we derive an approximate transition density, which is not Gaussian, that is associated with a drift-implicit EM discretisation of the underlying SDE (3.1). The rationale for this approach is that the drift-implicit discretisation yields qualitatively superior numerical simulation schemes for the SDE of interest (in particular, the positivity of the strong solution is preserved in the numerical solution obtained under the drift-implicit discretisation of (3.1)),

therefore the task is to determine whether a drift-implicit discretisation can lead correspondingly to superior parameter estimates. We now describe the steps taken to determine the DI-PML estimates, starting with the basic discretisation of (3.1) that leads to the approximate transition density, developed for the first time in this thesis.

First, we discretise (3.1) using a drift-implicit EM approximation,

$$r_{k+1} - r_k = (\alpha_{-1} r_{k+1}^{-1} - \alpha_0 + \alpha_1 r_{k+1} - \alpha_2 r_{k+1}^2)\Delta t + \sigma r_k^{\gamma} \Delta B_k, \tag{3.4}$$

where r_k and ΔB_k are defined as before. Note that this is very similar to the discretisation in (3.2), the only difference being the appearance of r_{k+1} in the drift coefficient, rather than r_k. As mentioned earlier in the chapter, Szpruch et al. (2011) proved that this discretisation scheme converges strongly to the solution of (3.1). Rearranging the drift-implicit discretisation, (3.4), we obtain

$$r_{k+1}(1 - \alpha_1 \Delta t) - \alpha_{-1} \Delta t r_{k+1}^{-1} + \alpha_0 \Delta t + \alpha_2 \Delta t r_{k+1}^2 = r_k + \sigma r_k^{\gamma} \Delta B_k. \tag{3.5}$$

Note that the conditional density of the RHS of (3.5) (i.e. the distribution of the RHS conditioned on the value of r_k) is normal, as it consists of a non-random term, r_k, plus another non-random term multiplied by an increment of Brownian motion. Let us represent the normally distributed RHS of (3.5) by u. Let the LHS, which is a function of r_{k+1}, be represented by the function $F(r_{k+1})$. The goal is now to derive an expression for $f^i(r_{k+1}|r_k, \theta)$, the drift-implicit analogue of $f^e(r_{k+1}|r_k, \theta)$

defined above. In what follows we will denote the transition density, $f^i(r_{k+1}|r_k,\theta)$, by $f_{r_{k+1}}(x)$ in order to keep the notation as clear as possible.

$$\begin{aligned} f_{r_{k+1}}(x) &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}(x \leq r_{k+1} \leq x + \Delta x) \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}(x \leq F^{-1}(u) \leq x + \Delta x) \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}(F(x) \leq u \leq F(x + \Delta x)) \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \int_{F(x)}^{F(x+\Delta x)} f_u(v) \mathrm{d}v \end{aligned}$$

Changing integration variables from v to s via the transformation $v = F(s)$ yields

$$\lim_{\Delta x \to 0} \frac{1}{\Delta x} \int_{x}^{x+\Delta x} f_u(F(s)) F'(s) \mathrm{d}s,$$

which is simply equal to

$$\begin{aligned} f_u(F(x))F'(x) = & \frac{1}{\sqrt{2\pi\Delta t}\sigma r_k^{\gamma}} \exp\left(\frac{-1}{2\sigma^2 \Delta t r_k^{2\gamma}} (F(x) - r_k)^2\right) \cdot \\ & (1 - \alpha_1 \Delta t + \alpha_{-1} \Delta t x^{-2} + 2\alpha_2 \Delta t x), \end{aligned}$$

recalling that u, the RHS of (3.5), is normally distributed. Note that in order to carry out the steps above (in particular, the first steps involving the inverse of the LHS of (3.5)) one must first demonstrate the existence of an inverse function $F^{-1}(u)$ used in the above derivation. For a proof of the existence of an inverse function see Szpruch et al. (2011). We have now derived the drift-implicit approximation to the transition density of the AS short rate model. One can then use this approximate transition density to form pseudo likelihood functions by taking advantage of the Markovian nature of diffusions in the manner described above.

3.5 Numerical experiment

In order to test the newly developed approximation scheme outlined above, we conducted a numerical experiment to empirically test whether the new approximation yielded better results relative to the standard, explicit PML estimates obtained via the Gaussian approximation to the transition density of the process, which was briefly outlined in section 3.4 above. We now provide some details regarding the experiment set up, which will be followed by a presentation of the results.

3.5.1 Experiment design

Our methodology involved simulating a single trace from the process of interest (3.1)[3] using a pre-determined parameter vector, which we will label θ_0 for the remainder of this chapter[4] The sample trace from the model contained $N = 10,001$ observations; 1000 observations per time period, and 10 time periods, plus one additional observation at $t = 0$, which we will assume is known, and not random. We label the N data points from the model D in the following analysis, i.e. $D = \{D_i\}_{i=0,\ldots,N}$. The approximate loglikelihood functions associated with the model 3.1 were then derived by taking advantage of the Markovian nature of Itô diffusions, i.e. because diffusion processes are Markovian by construction, the likelihood function can be written as the product of the transition densities associated with successive observations of the process

$$L(D|\theta) = \prod_{i=1}^{N} f(D_i|D_{i-1}, \theta). \tag{3.6}$$

[3]In practice, an approximate numerical simulation scheme with a very small step size ($1,000,000$ per time period) was used to simulate the sample path of the process due to the analytical form of the true transition density of the process not being known.

[4]Thoughout this chapter we will refer to θ_0 as either the 'true' parameter values, or the 'data generating' parameters.

By substituting either the drift-explicit or drift-implicit EM approximation to the transition density for the true density in (3.6), one obtains an approximate likelihood function that leads to PML or DI-PML parameter estimates respectively. The true parameter values, θ_0, that were used to generate the M sample paths were chosen as follows

$$\theta_0 \equiv (\alpha_{-1}, \alpha_0, \alpha_1, \alpha_2, \sigma, \gamma) = (1.0, 3.0, 2.0, 4.0, 0.8, 1.3).$$

The optimisation was performed over a large region of the parameter space; the lower and upper bounds of the search space were as follows:

$$B_l = (0.0, 0.0, 0.0, 0.0, 0.0, 1.0), \quad B_u = (50.0, 50.0, 50.0, 50.0, 50.0, 1.5).$$

The lower bounds (and the upper bound for the CEV parameter, γ) are consistent with the parameter bounds introduced in section 3.2, while the upper bounds were chosen in order to ensure that the search space was large enough such that the optimisation could explore a wide range of possible solutions.

The approximate loglikelihood functions were optimised in C++, using a Nelder-Mead simplex optimisation routine (see Nelder and Mead (1965) for further details of this technique) from the Gnu Scientific Library (GSL). The approximate likelihood functions associated with the simulated data D were each optimised 1500 times, with each optimisation being initialised at a different, randomly generated, point in the constrained search space. The $M = 500$ best estimates[5] were then used to construct point estimates and confidence intervals for the model parameters. The rationale behind using only a subset of the 1500 optimisations was that a small number (around 50 in each case) of optmisations failed to converge to within the

[5]By 'best' we mean the estimates that yielded the largest pseudo-likelihood value.

specified tolerance, while a similarly small number of estimates that did converge to within the specified tolerance were located in regions of the parameter space that were very far from the bulk of the estimates, with these points evaluating to pseudo-likelihood values that were significantly lower than the vast majority of the other converged parameter estimates. Removing these points from the results ensured that valid pseudo-maximum-likelihood estimates were used to derive our conclusions. Point estimates were derived by evaluating the sample mean for each collection of estimates; confidence intervals are empirical, i.e. they are generated using the 2.5th and 97.5th percentiles of the sample estimates. Point estimates and confidence intervals derived for each method of approximation (PML and DI-PML) were then compared to one another and compared to the data generating parameters, θ_0, to evaluate the relative efficiency of each method.

3.5.2 Results

Table 3.1 contains the results of the numerical experiments.

Table 3.1: AS SDE parameter estimation results

	Parameter					
	α_{-1}	α_0	α_1	α_2	σ	ρ
True value	1.0	3.0	2.0	4.0	0.8	1.3
Explicit estimate	3.80	22.06	44.23	33.37	0.81	1.32
Explicit CIs	(2.88,4.13)	(15.41,24.52)	(28.62,50.00)	(21.67,37.70)	(0.81,0.81)	(1.32,1.32)
Implicit estimate	3.35	18.68	36.00	27.00	0.80	1.30
Implicit CIs	(1.29,4.17)	(3.60,24.65)	(0.52,50.0)	(0.34,37.48)	(0.80,0.81)	(1.30,1.30)

Evidently both methods fail to accurately estimate the data generating parameters that appear in the drift coefficient of (3.1); not only are the point estimates associated with both approximations significantly biased away from the true values, but the confidence intervals associated with the estimates are extremely wide. It appears that the DI-PML point estimates (which correspond to the means of the $M = 500$ sample parameter estimates) are slightly less biased than the point estimates obtained via the PML approximation, but the degree of variablity in the drift parameter estimates associated with the DI-PML method is greater than the variability in the PML estimates. As was pointed out earlier in the chapter, the crude PML approach to parameter inference is known for producing extremely biased estimates, especially in the case of models with nonlinear terms, such as the model considered here. Unfortunetely it seems that the DI-PML approach does not offer any material improvement over the PML approach in terms of accuracy of estimates.

The estimates of parameters appearing in the diffusion coefficient (σ and γ) of (3.1) are considerably better relative to the drift parameter estimates, for both methods of approximation. Both approximations yield point estimates that are close to the true parameter values, but the DI-PML point estimates are marginally superior to the PML point estimates, which exhibit a slight upward bias for both the σ and the γ parameters. Additionally, despite the point estimates associated with the PML estimates being very close to the true parameter values, the empirical 95% confidence intervals around the estimates do not actually include the true parameter values, whereas the confidence intervals associated with the DI-PML approximation do include the true parameter values. This result suggests that, despite these discretisation schemes being ineffective at generating accurate drift parameter estimates, this method of approximation might be used to derive relatively accurate

parameter estimates for those parameters appearing in the diffusion coefficient of the model, with the DI-PML method being preferrable to the PML method. The following boxplots (figures 3.1 and 3.2) provide a more detailed description of the distribution of parameter estimates obtained from the experiment.

Figure 3.1 reinforces the conclusions that were drawn above; namely, that both methods of estimation are inadequate for the purposes of accurate parameter inference for (3.1). The interquartile ranges (IQRs) of both the PML and DI-PML estimates both span large regions of the parameter space, indicating that there may be parameter identifiability issues associated with this particular model; a consideration that we will return to later in this chapter.

Figure 3.2 clearly illustrates that both methods of approximation, PML and DI-PML, provide good estimates of the diffusion coefficient parameters, σ and γ, with the DI-PML method producing marginally superior results. Both methods of parameter inference result in tight box and whisker plots, which indicates that there was very little variability in the estimation of the diffusion parameters.

In the course of conducting the numerical experiment described above, we ran into various difficulties associated with our method of inference that made obtaining parameter estimates from the model observations particularly problematic. In what follows we will discuss some of the reasons underlying these difficulties.

During preliminary testing of the optimisation routine used in the experiment, we noticed that each optimisation of the approximate likelihood function[6] produced converged parameter estimates that differed greatly in their position in the parameter space, but evaluated to the same loglikelihood value. Discretisation bias associated with the PML and DI-PML approximations do not explain this

[6]The phenomenon we describe was observed using both approximations to the loglikelihood function considered in this chapter, but for illustrative purposes the analysis that follows in this section is focussed on the PML approximation only.

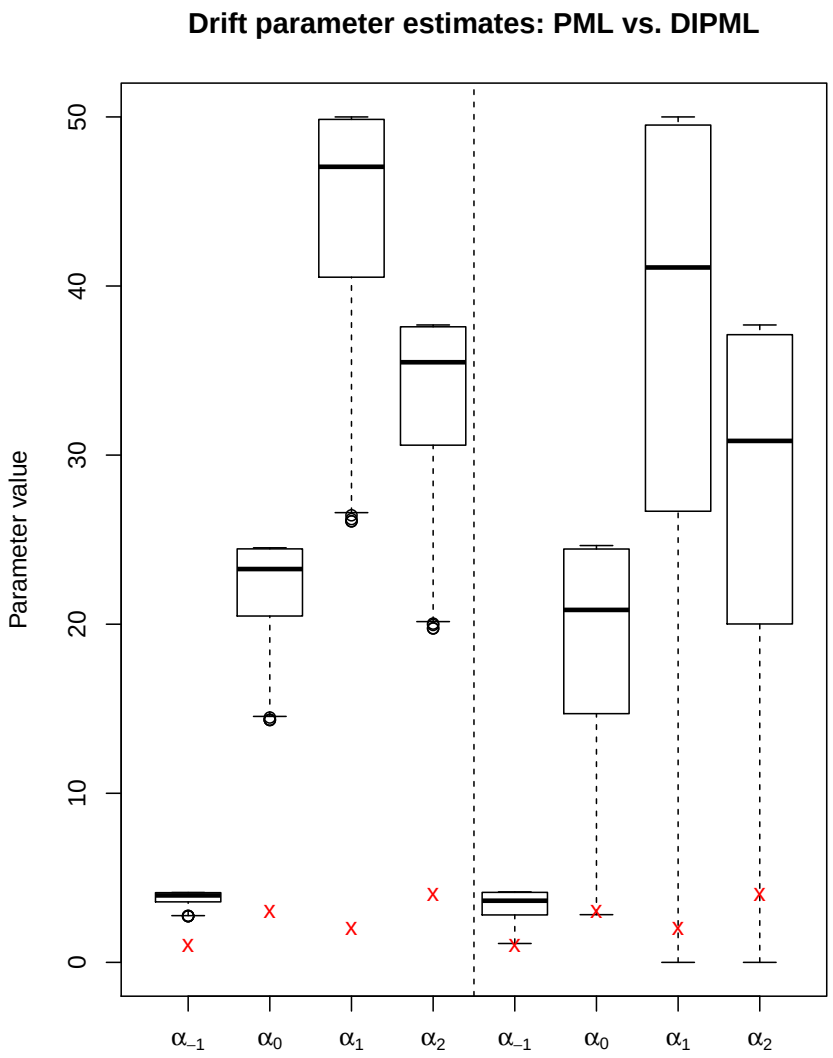

Figure 3.1: The first four boxplots relate to the ensemble of drift parameter estimates associated with the PML approximation; the last four relate to estimates of the drift parameters obtained via the DI-PML approximation. The red crosses represent the true parameter values—the parameter values used to generate the observations used in the experiment.

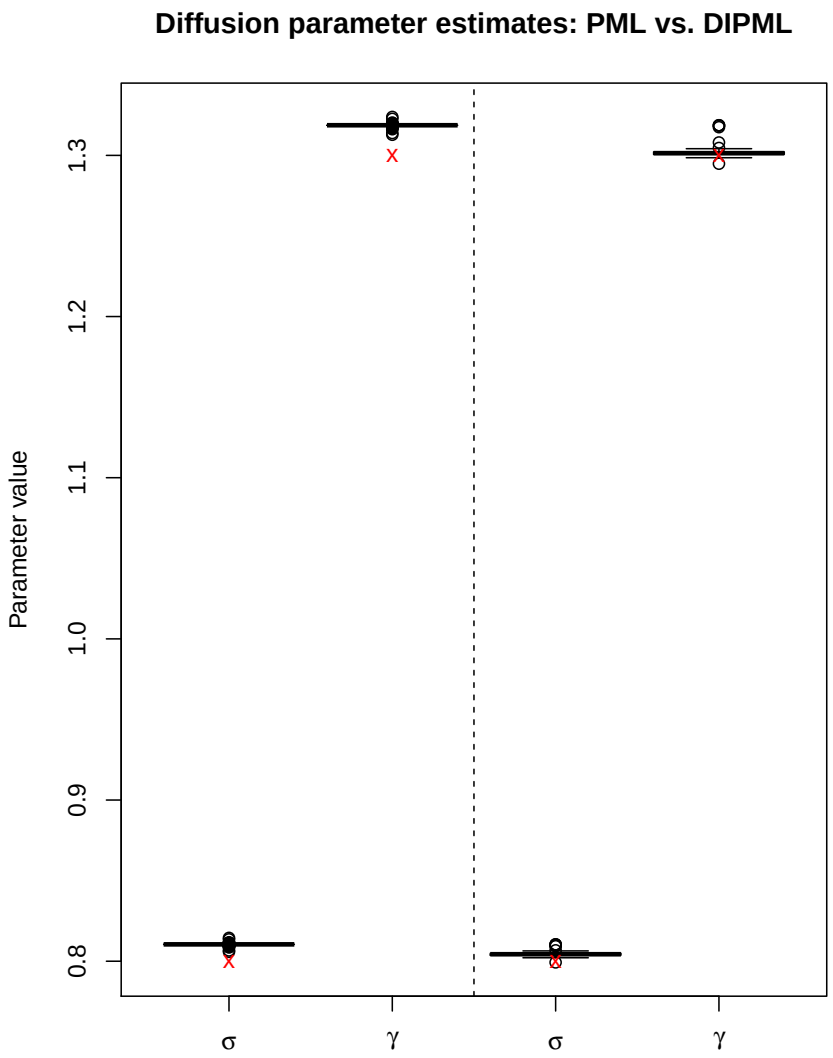

Figure 3.2: The first two boxplots relate to the ensemble of diffusion parameter estimates associated with the PML approximation; the last two relate to estimates of the diffusion parameters obtained via the DI-PML approximation. The red crosses represent the true parameter values—the parameter values used to generate the observations used in the experiment.

observation; the fact that the locations of the parameter estimates differed significantly from the locations of the data generating parameters can be explained by the well-documented discretisation bias problems associated with the pseudo-likelihood methods investigated here, but the wide spread of estimates associated with the drift parameters suggests that, in addition to the problem of biased estimates, there was an additional parameter identifiability problem associated with this model—the model observations could be explained equally well by a wide range of different parameter values. The distribution of sampled parameter estimates is illustrated in Figure 3.3. It is clear from Figure 3.3 that no unique combination of drift parameters best explains the data, and this explains why point estimates and associated confidence intervals of the drift parameters were so poor for both types of approximation. The optimisation stage was repeated using another GSL optimisation routine, the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method[7]—a quasi-Newton method that utilises the gradient of the loglikelihood function to move towards local optima—in order to test whether the loglikelihood optimisation was sensitive to the choice of solver used, but the same behaviour was observed using this solver. If the ridge-like features that appear in Figure 3.3 were indeed present in the approximate loglikelihood functions, they would almost certainly manifest themselves in the form of parameter identifiability problems at the optimisation stage. In order to confirm our hypothesis that a ridge is present in the parameter space which makes identifying a unique PMLE impossible, we generated an empirical sample from the approximate loglikelihood function using a Markov chain Monte Carlo (MCMC) sampling algorithm. As a brief aside: MCMC samplers generate samples from a target function (in this case, the approximate loglikelihood function associated with the explicit EM discretisation of (3.1)) by constructing a

[7]See Nocedal and Wright (2006) for details of this method.

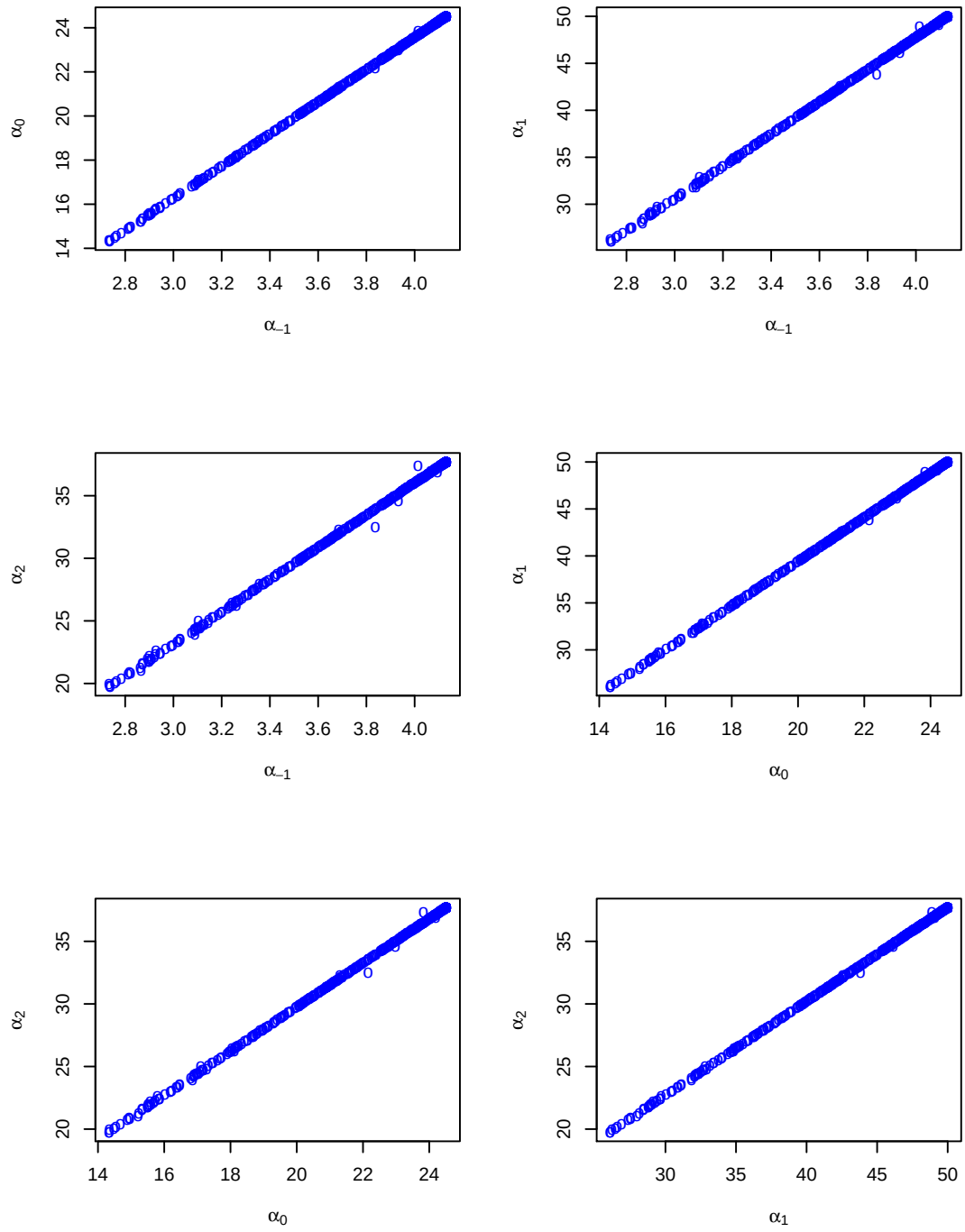

Figure 3.3: These plots demonstrate that the $M = 500$ parameter estimates lie on a ridge, running across large portions of the parameter space. The pseudo-loglikelihood function evaluates to the same value all along this ridge.

Markov chain with limiting distribution equal to the target function Gilks et al. (1996). By running the constructed Markov chain for a sufficiently long period of time (long enough for the chain to converge to its limiting distribution), draws from the Markov chain should represent correlated samples from the target distribution, which can be used to estimate model parameters or, as in this case, to visualise the features of a multidimensional function. This sampling method is extremely common in Bayesian statistics due to the sampler's ability to generate samples from complex distributions whose nomalising constant is not computable, but it is equally valid as a means of sampling from complex functions, outside the Bayesian paradigm. This sampling technique will be discussed at length in the following chapter, where we will make use of MCMC techniques to develop new samplers that can be used for parameter inference in cases where the model likelihood is not even computable pointwise. The rationale for employing MCMC here is that it allowed us to visualise the shape of the loglikelihood function in the vicinity of the ridge, rather than simply assuming that the ridge was present on the basis of the location of the converged parameter estimates. Figure 3.4 plots the samples generated by the MCMC sampler. The MCMC samples in Figure 3.4 provide further evidence of parameter identifiabiliy problems associated with (3.1). One can clearly observe the same ridge-like pattern in the MCMC sample that was observed in the scatter plot of the parameter estimates (see Figure 3.3). To further illustrate this problem, consider Figure 3.5. The black line represents the magnitude of the drift coefficient of (3.1) as a function of the process value, given the parameter values are equal to θ_0. The red line represents the drift coefficient resulting from a sample parameter

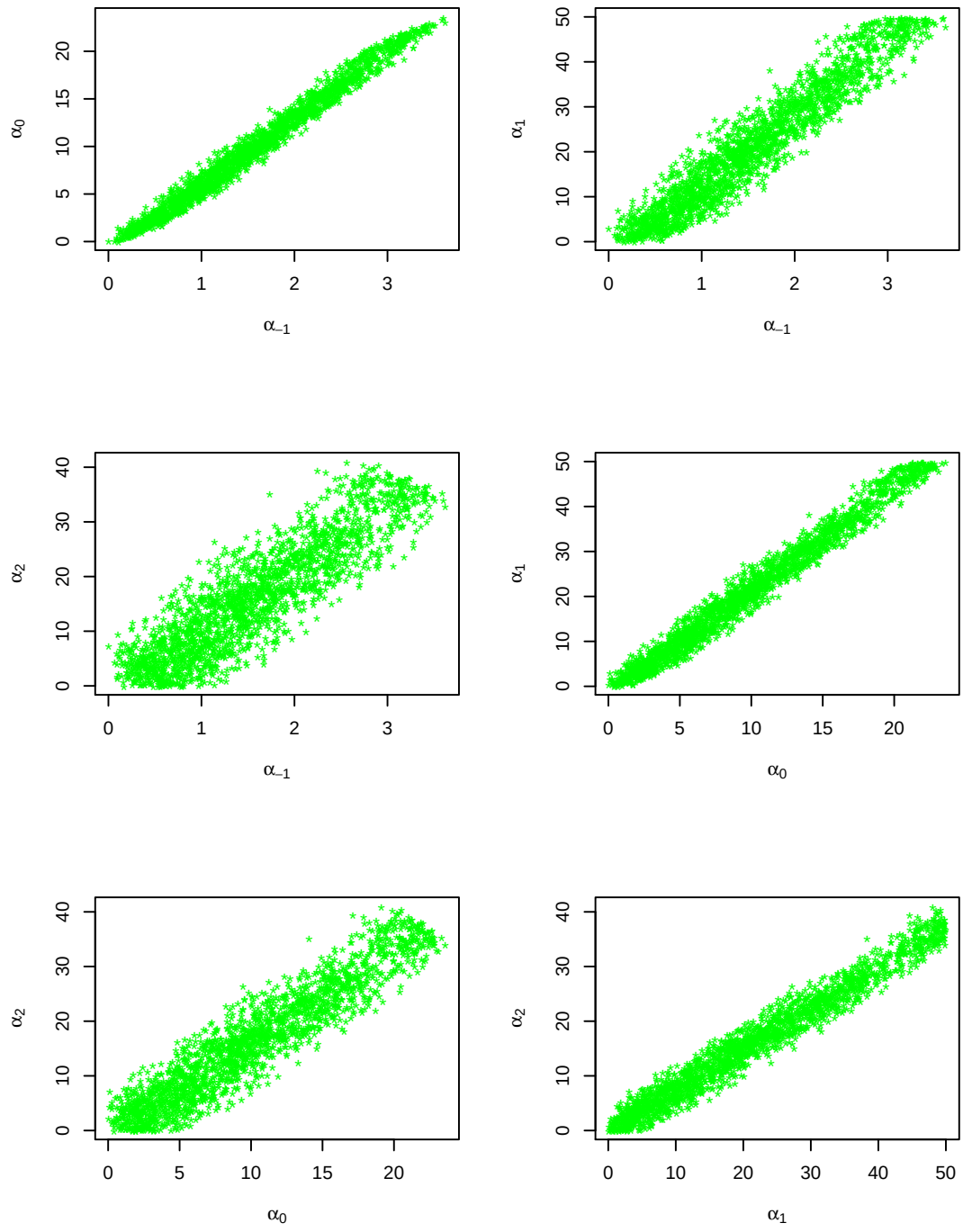

Figure 3.4: Marginal loglikelihood samples associated with the explicit Euler discretisation of (3.1). Diffusion parameters were held constant, at their true values $(\sigma, \gamma) = (0.80, 1.30)$, when running the MCMC sampler in order to simplify the analysis; this is appropriate given that it is the structure of the approximate loglikelihood function in the drift parameter space that is of interest here.

estimate obtained from the numerical experiment. The estimated parameter values used to draw the red line were as follows:

$$\hat{\theta} = (2.07, 9.45, 14.47, 10.92, 0.81, 1.32). \tag{3.7}$$

Note that only the first four parameter values are associated with the drift coefficient of the model. Clearly, these parameter estimates differ, to a significant extent, from the true parameter values, θ_0, and yet both the mean reversion levels and the magnitudes of the drift coefficients are very similar. With such similarities between the drift coefficients associated with parameter values that are significantly different from one another, one would indeed expect the data generated using $\hat{\theta}$ in (3.7) to be virtually indistinguishable from data generated using θ_0.

Figures 3.6, 3.7 and 3.8 further illustrate the problems associated with the estimation of (3.1). These 'funnel' charts attempt to convey graphically the evolution of the distributional properties of the data generated from (3.1) using a particular choice of parameter values. Chart one was generated using the true parameter values, funnel charts two and three were generated using samples from the ensemble of parameter estimates obtained during the numerical experiment. The two sets of parameter estimates used to generate figures 3.7 and 3.8 were chosen because the drift parameter estimates are significantly different in each case. Despite each set of parameters differing to a significant extent, the distributions of the data produced in each case are very similar. This provides further evidence to support the claim that identifying unique parameter estimates using the maximum likelihood estimation approach is very problematic in this context. One can rationalise these observations by noting that the parametric form of (3.1) was first introduced in Ait-Sähalia (1996), in which the author tests a variety of stochastic diffusion processes for their

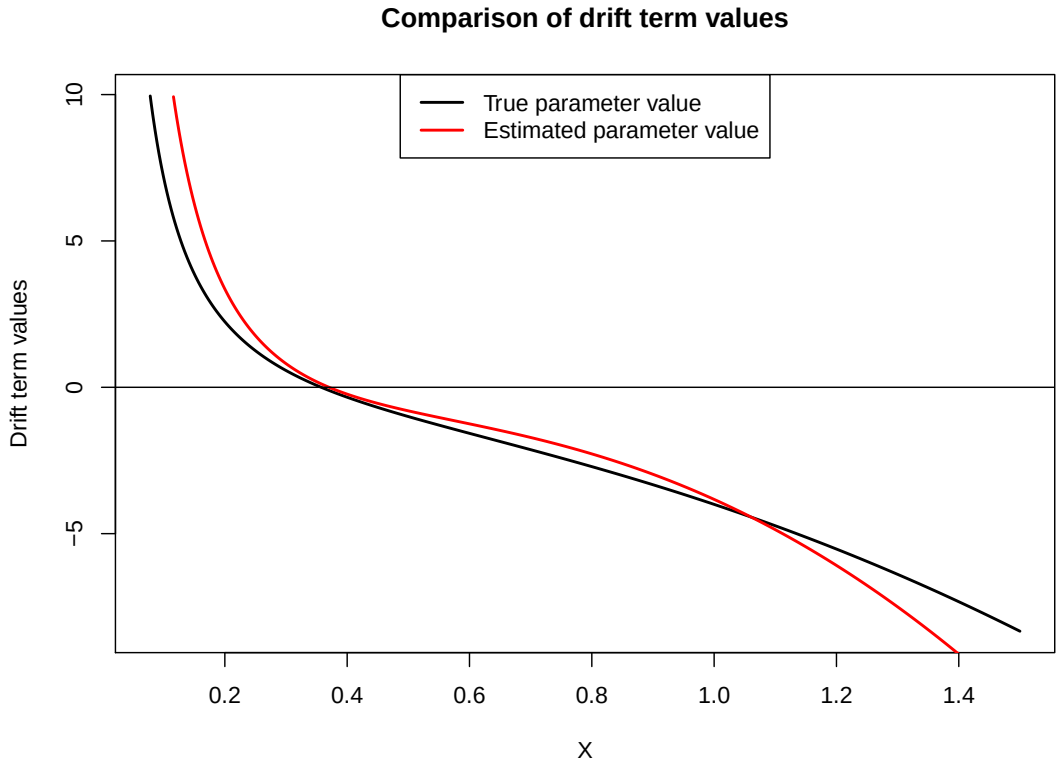

Figure 3.5: This is a plot of the drift coefficient evaluated at two different parameter values; the black line represents the drift associated with the true parameter values (the values used to generate the samples used to infer parameters) and the red line represents the drift associated with a particular set of parameter estimates obtained during the experiment. When the line is above zero, the process exhibits a drift downwards; conversly, when the line is below zero, the process drifts upwards. The intersection of the line with the x-axis (labelled X) represents the mean reversion level of the process; the value towards which the process will tend to drift over time.

ability to accurately represent the features observed in real market interest rate data. In Ait-Sähalia (1996) the author tests the various parametric models by comparing their implied parametric density to the nonparametric density derived from the market observations. Ait-Sähalia (1996) rejects most models at the 95% level; the only model not rejected was similar to (3.1)[8]. In other words, (3.1) is flexible in the sense that its parametric form is such that market data can credibly be represented by this model; more parsimonious models are not flexible enough for this purpose. While this flexibility is beneficial in the sense that the model is capable of representing the features observed in real data, it is a hinderance when it comes to actually inferring the parameter values that are most likely to have generated the observed data, precisely because the flexibility of the model results in the feature that many different choices of model parameters could credibly give rise to the observed data.

3.6 Discussion & summary

In this chapter we introduced a new method of parameter estimation for equation(3.1) based on a drift-implicit discretisation of the SDE. The new estimation method was tested in order to determine whether the advantages associated with drift-implicit discretisations in the context of numerical simulation of SDEs are carried over into the problem of parameter inference. In summary the new method of parameter estimation introduced here, namely the DI-PML approximation, fails to provide parameter estimates that are sufficiently accurate to merit the adoption of this method of parameter inference, at least in the context of drift parameter estimation. As the above discussion suggests, this is a feature that is inherent in

[8]The form of the drift coefficient in the non-rejected model was the same as in (3.1), but the diffusion coefficient posessed a more general parametric form than the diffusion coefficient of (3.1).

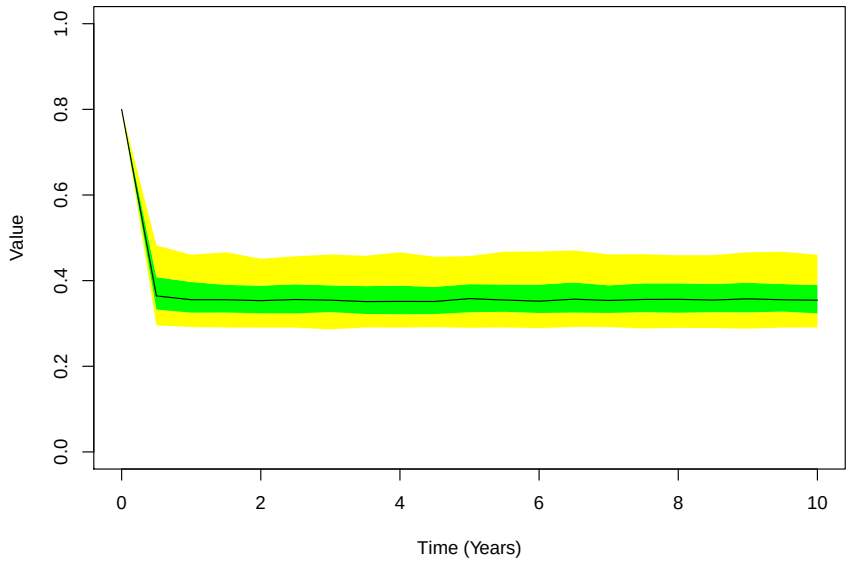

Figure 3.6: This figure plots certain percentiles of the distribution of the process, X, whose dynamics is given by (3.1). The true parameters, θ_0, were used to generate the data, and the percentiles were derived by simulating 1000 sample paths of the process and evaluating the 95th, 75th, 50th, 25th and 5th empirical percentiles of the sample paths. The black line represents the median (50th percentile), the green area represents the inter-quartile range, and the yellow regions represent the range between the 5th and 25th, and the 75th and 95th percentiles.

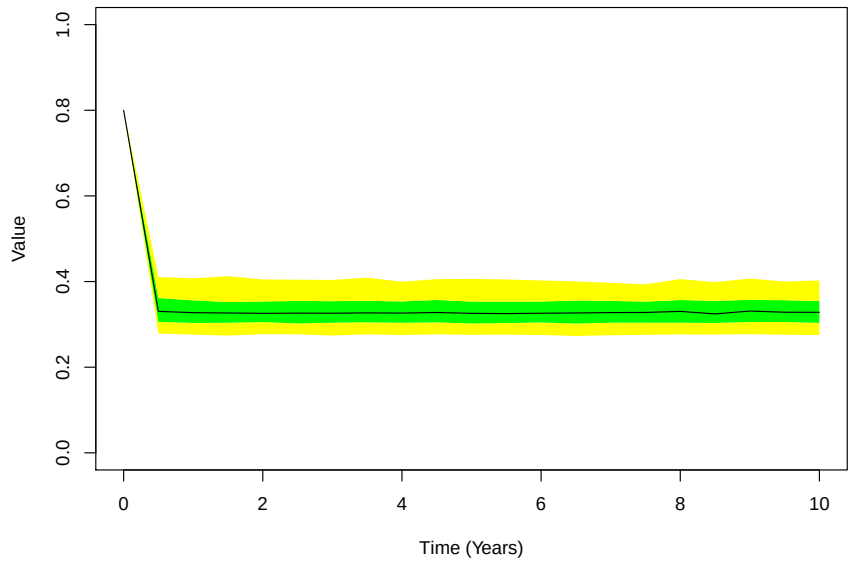

Figure 3.7: This funnel chart was constructed in a similar manner to Figure 3.6, except the sample data used was derived from (3.1) using a sample from the parameter estimates obtained during the experiment. The parameter values used were $(1.40, 4.78, 2.50, 2.94, 0.73, 1.21)$.

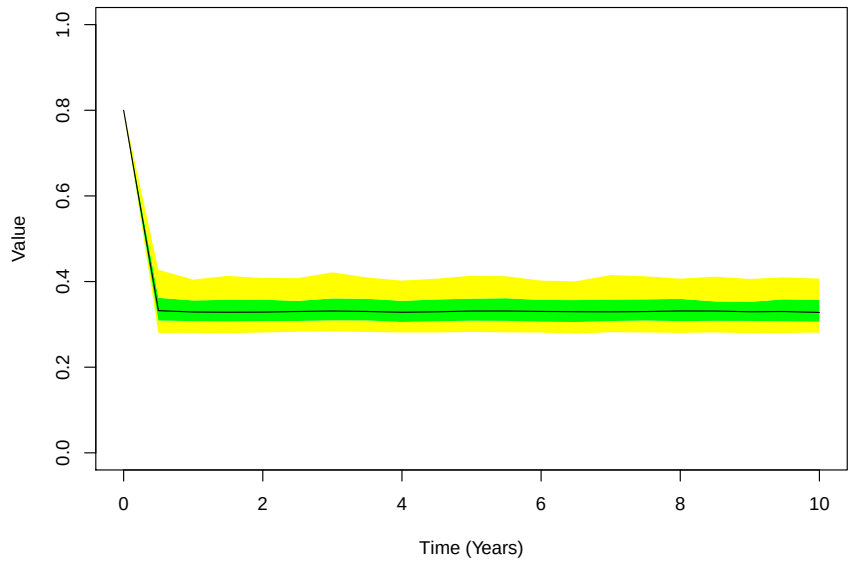

Figure 3.8: This funnel chart was produced using the same method that was used in the creation of figures 3.6 and 3.7. The sample paths used to obtain the percentiles were generated using a sample from the parameter estimates obtained during the experiment. The parameter values used were $(2.15, 9.48, 11.63, 8.08, 0.76, 1.26)$.

the structure of the AS model itself, which is popular due to its ability to capture a wide range of statistical features that appear in market interest rate data. In cases where a large number of parameter choices can credibly give rise to the observed data, traditional point estimates are not of much use—there is no reason, when taking a maximum likelihood approach to inference, to take one point estimate over another if both can describe the data equally well. One way of encapsulating this ambiguity in the parameter estimates is to take a Bayesian approach to inference, whereby one can talk of parameters in terms of distributions, without relying on asymptotic normality assumptions that are sometimes required when constructing confidence intervals, for example, in the frequentist approach to statisics. By taking such an approach, one can properly estimate the distribution of the parameters, taking into account dependencies between parameters, as well as having a robust framework for handling uncertainty in the model and an explicit method of taking additional sources of information into account in the process of parameter inference, which may be necessary if one wishes to reduce the ambiguity surrounding the 'best choice' of model parameters. In the succeeding chapters we will present some new Bayesian inference techniques that can be used to estimate financial models.

Chapter 4

Statistical inference of model parameters using Approximate Bayesian Computation

4.1 Background and prerequisites

In the previous chapter concerning the estimation of model parameters using pseudo-likelihood methods, we encountered problems associated with standard point estimates based on optimisation of the approximate loglikelihood function. We observed that a large subset of the parameter space could have credibly given rise to the observed data on which inference was based. As a result the point estimates of the drift coefficients obtained during the exercise, under both forms of approximation (PML and DI-PML), were severely biased away from the true parameter values. Construction of empirical confidence intervals did not help matters as the intervals often spanned large portions of the parameter space. Only when we implemented a MCMC sampler to build up a picture of the approximate loglikelihood surface were we able to precisely identify the parameter identifiability problem. In what follows we introduce several newly developed samplers, based on a Bayesian approach to parameter estimation, that can be used to generate

samples from distributions of interest (typically the posterior distribution of model parameters) and obtain relevant point and interval estimates of model parameters. Approximate Bayesian Computation (ABC) is a method of model inference that takes advantage of the asymmetry represented by the relative ease with which one can simulate data from a model and the difficulty in inferring parameter values given sample data from a model. Standard methods of parameter inference for SDEs usually involve the computation of the so called likelihood function associated with the model; however, obtaining this function often requires the analytic solution of the SDE under investigation to be known, a criterion which is often not satisfied. Typically, SDEs that are used in finance are intractable, in the sense that analytic solutions are frequently not available. Hence, the likelihood function is not known, which makes performing inference a difficult task. ABC methods provide a promising, and flexible, approach to the estimation of model parameters, including the parameters of SDEs used in finance. If the utility of using more complex SDEs to realistically represent the dynamics of state variables, such as interest rates or share prices, is to be realised, robust methods for reliably estimating such models must be developed. In this chapter we will provide an overview of the development of the method of inference under investigation; we will also set out the relevant results necessary for understanding the method at a fundamental level.

4.1.1 Background

There are various approaches one could take when attempting to estimate the parameters of a model; the most common approaches being likelihood based methods (which belong to the family of frequentist approaches to inference) and Bayesian methods. Both methods have pros and cons associated with them; however, the

systematic way in which information contained within observed data and external information, e.g. expert opinion, can be combined via Bayes' theorem, along with the feature that inference can be performed without the use of asymptotic results, made the Bayesian approach an attractive one. From a practical perspective, Bayesian results also enjoy the advantage of being intuitively interpretable—Bayesian credible intervals can legitimately be understood as representing the probability that a particular estimand falls in some range; in contrast, the frequentist confidence interval must be strictly understood in terms of a series of repeatable experiments, a concept that is more difficult for lay-persons to grasp (Gelman et al., 2003). The main controversy associated with Bayesian inference is that the results can be influenced by subjective opinion via the prior. Of course, this is paradoxically also seen as one of the strengths of Bayesianism—it is often the case that the statistician has additional information relevant to the problem that is not contained within the observations, for example there may be parameter ranges outwith which the model does not give realistic outputs. The relative influence of data and prior beliefs is determined partly by the amount of data available. When estimating financial models—the main practical application considered in this thesis—rich data sets are often available thanks to detailed market data being readily available from sources like Datastream or Bloomberg, and therefore the data will tend to exert a strong influence on the posterior distribution, provided the prior is not too tightly constained i.e. provided the degree of belief in the external information is not too high. For this reason, it can be argued that the Bayesian approach has the potential to be a flexible, reliable, and powerful tool for conducting parameter estimation in a financial context. In addition, by selecting the prior distribution of parameters to be uniform, the posterior density obtained via Bayes' Theorem actually coincides with the likelihood surface, and therefore the maximum a posteriori point (MAP)

(the mode of the posterior distribution) coincides with the maximum likelihood estimate, provided the range of the prior distribution contains the MAP. Although both methods are interpreted differently at a philosophical level, in a practical sense the Bayesian approach to parameter estimation can be seen as a general method of estimating parameters, the maximum likelihood method being a particular case of the Bayesian method. As mentioned in the previous paragraph, Bayes' Theorem is the fundamental relation upon which all Bayesian analysis depends. Bayes' Theorem is given below

$$\pi(\theta|D) = \frac{f(D|\theta)\pi(\theta)}{\pi(D)}, \quad \text{where } \pi(D) = \int_{\theta} f(D|\theta)\pi(\theta)\mathrm{d}\theta. \tag{4.1}$$

Here, $f(D|\theta)$ stands for the likelihood of the data conditional on a particular parameter value θ, which can be broadly interpreted as the probability of observing data D from the model, given the model is parameterised by the parameter value θ[1]. $\pi(\theta)$ represents the prior distribution of the parameter vector. The prior contains external information about the model, e.g. expert opinion about the ranges within which the components of the parameter vector are likely to lie. $\pi(D)$ serves as a normalising constant, ensuring that the LHS of (4.1) is a valid probability density; it is sometimes referred to as the marginal likelihood, model evidence, or the prior predictive density (Prior PD) (Gelman et al., 2003). The parameter vector θ is defined on the space Θ, usually a subset of $\mathbb{R}^p$, with $dim(\theta) = p$. For complex models with many parameters, evaluating the normalising constant (the Prior PD) in (4.1) becomes extremely difficult, which often means that obtaining analytic expressions for the posterior density is not possible. Owing to an improvement in computing power, computationally intensive methods such as Markov chain

[1]This parameter can be a single unknown quantity or a vector of unknown quantities.

Monte Carlo (MCMC) and sequential Monte Carlo (SMC) have been developed that bypass the need to evaluate the normalising constant; as long as the likelihood can be evaluated pointwise up to a constant, methods like MCMC and SMC are capable of generating samples from the posterior distribtion. The details behind these methods will be explained later in this chapter. As models become larger and more complex, the likelihood function becomes correspondingly more complex and difficult to compute (Beaumont, 2010), which severely restricts the applicability of popular methods like MCMC. ABC methods provide a means of circumventing this problem. What follows is an overview of the ideas behind ABC methods.

Although there has been a relatively rapid development in ABC techniques in recent years, the concept itself was first alluded to in the mid-eighties by Rubin (1984). The salient features of Rubin's algorithm are detailed below.

1. Draw $\theta_i \sim \pi(\theta)$.
2. Simulate $X_i \sim f(X|\theta)$.
3. Accept θ_i if $X_i = D$.

Repeat until N points have been sampled. D represents the observation used to infer parameter values. This rejection algorithm produces samples from the true posterior density of the parameter vector θ, and is therefore not strictly an example

of ABC. The validity of this straightforward algorithm is easily demonstrated as follows

$$\begin{aligned} f(\theta_i) &\propto \pi(\theta_i) \sum_{X \in \mathcal{D}} f(X|\theta_i) \mathbb{1}(X = D) \\ &= \pi(\theta_i) f(D|\theta_i) \\ &\propto \pi(\theta_i|D), \end{aligned}$$

where $\mathcal{D}$ represents the space on which model outputs are defined. This algorithm has limited uses. If the data are distributed on a continuous state space then the probability of simulating data that exactly coincides with the observed sample is zero; in this scenario it becomes necessary to adapt the previous algorithm by replacing step 3 with the following step

3. Accept θ if $\rho(X_i, D) \leq \epsilon$,

where $\rho(\cdot, \cdot)$ is some distance metric between simulated data and observed data. This formulation was proposed by Pritchard et al. (1999). Pritchard's algorithm produces samples from an approximation to the posterior distribution of parameters, the accuracy of which is controlled via the data mismatch parameter, ϵ. If ϵ is taken to be zero, the distribution of sampled parameters reduces once more to the true posterior. It is worth noting, however, that the data mismatch parameter is typically not chosen to be zero; although this choice produces samples from the true posterior, the algorithm becomes inefficient due to the zero probability of producing data from the model that exactly coincides with the observations. Therefore, a compromise between the accuracy of the posterior approximation and the efficiency of the rejection algorithm is usually sought by choosing a mismatch parameter that

is as small as possible, but not zero. Once the data become multidimensional[2] then it may become necessary to condense the information contained within the data into summary statistics (Beaumont, 2010), in order to avoid unacceptably high rejection rates in the above algorithm. That is, step three in the above procedure might be replaced with

3. Accept θ if $\rho(S(X_i), S(D)) \leq \epsilon$,

where $S(\cdot)$ represents, possibly a vector, of summary statistics that reduces the dimensionality of the observations and data generated from the model whose parameters are being estimated. The choice of summary statistics is of crucial importance and, despite efforts to generalise the process, is still generally done on a problem-by-problem basis. Methods for choosing summary statistics will be discussed in the following chapter.

As mentioned above, in general ρ is some metric that measures the distance between elements in the set of possible model outputs, but if one assumes that the set of model outputs form a group with addition operator, $+$, and operator $-$, defined as $a - b = a + (-b)$, where $-b$ is the inverse of b, then one is able to replace the (more general) distance metric, $\rho(S(X), S(D))$, with the algebraic subtraction operator, $S(D) - S(X)$. This, in turn, allows one to interpret ABC as sampling from a convolution of the true model likelihood with some error distribution, which we will refer to as the similarity kernel. Making this slightly more restrictive assumption about the structure of the set of model outputs allows us to attach an intuitive interpretation to ABC sampling techniques and to derive some important convergence results associated with the technique, which will be discussed in the following section. In most cases, this assumption is not terribly restrictive—this

[2]In this context, 'multidimensional' data refers to the case in which the model output consists of more than a single number, e.g. a time series of observations from a SDE.

assumption poses no problems in the context of mathematical finance, wherein data typically consists of numerical scalars, or vectors. This assumption is more relevant in situations where one is attempting to apply ABC sampling techniques to models over graphs, trees, or strings, which may occur in applications concerning population genetics, for example.

4.1.2 Prerequisites

The archetypal rejection algorithm outlined above, in particular the use of a distance metric ρ combined with a mismatch tolerance ϵ, was given a probabilistic interpretation by Wilkinson (2013). Wilkinson demonstrated that the ABC rejection algorithm generated samples from the true posterior distribution of the parameters if one assumes that there is a discrepancy between the model run at its best parameter values and the observations. In other words, if one assumes that the observed data, D, represents a realisation of the model run at its best input, $\mathcal{M}(\hat{\theta})$[3], plus an independent error term ε, distributed according to some distribution π_ε, i.e.

$$\begin{aligned} D &= \mathcal{M}(\hat{\theta}) + \varepsilon, \\ \varepsilon &\sim \pi_\varepsilon \end{aligned} \tag{4.2}$$

then the rejection algorithm detailed above gives samples from the exact posterior distribution of the parameters. The case where a $0-1$ cut-off is used, i.e. simulated parameter values are accepted if $\rho(X_i, D) \leq \epsilon$ and rejected otherwise, imposes on

[3]Here, $\mathcal{M}(\hat{\theta})$ denotes the model that we assume the observations are generated from; later in this chapter, we will also use the notation $f(\cdot|\theta)$ to denote the likelihood associated with the model $\mathcal{M}(\theta)$, especially in the context of generating pseudo data from the model.

the problem the assumption that the error term, ε, is uniformly distributed within an n-ball of radius ϵ, i.e.

$$\varepsilon = \begin{cases} \frac{1}{C_\varepsilon} & \text{if } \rho(X_i, D) \leq \epsilon, \\ 0 & \text{otherwise,} \end{cases}$$

where C_ε is a normalising constant that ensures that the PDF of the error term integrates to one. In the case above, where the error term is assumed to be uniformly distributed within an n-ball, the normalising constant is expressed in terms of the volume of an n-ball, V_n:

$$C_\varepsilon \equiv \frac{1}{V_n} = \frac{\Gamma(\frac{n}{2}+1)}{\pi^{n/2}\epsilon^n},$$

where n is the dimension of the data produced by the model. More generally, if we assume that the error term implicit in the observations is distributed according to some distribution that is centred at the origin, the dispersion of which being dependent on ϵ, i.e. $\varepsilon \sim \pi_\varepsilon(\cdot|\epsilon)$, then this corresponds to the following rejection algorithm:

1. Draw $\theta \sim \pi(\theta)$.
2. Simulate $X \sim f(X|\theta)$.
3. Accept θ with probability $\pi_\varepsilon(D - X|\epsilon)$.

A proof demonstrating that this algorithm produces samples from the posterior distribution, assuming (4.2) holds, is given in the appendix at the end of this chapter. Making the assumption that the observations consist of the sum of two independent random variables—the model output at the optimum parameter value and the error term—allows one to represent the likelihood of the observations

as a convolution of two probability densities: the model likelihood, $f(X|\theta)$, and the error distribution, $\pi_\varepsilon(D - X|\epsilon)$. This convolution will be labelled the ABC likelihood approximation. Furthermore, by letting the data mismatch tolerance, ϵ, tend to zero, the ABC likelihood approximation reduces to the true likelihood of the data assuming no discrepancy between model output and observations, i.e. the ABC likelihood approximation reduces to the model likelihood. This result is stated formally in the following lemma.

Lemma 7. *Assuming the observed data represents the sum of the model output and an error term* (4.2)*, the likelihood of the observations is given by*

$$f_{ABC}(D|\theta, \epsilon) = \int_{\mathcal{D}(X)} \pi_\varepsilon(D - X|\epsilon) f(X|\theta) \mathrm{d}X, \tag{4.3}$$

where $\mathcal{D}(X)$ represents the sample space on which the random variable X is defined. In addition,

$$\lim_{\epsilon \to 0} f_{ABC}(D|\theta, \epsilon) = f(D|\theta) \tag{4.4}$$

Proof In what follows, we derive the density function associated with the convolution of two independent random variables. Let $f(X|\theta)$ represent the unknown likelihoood function of the model evaluated at the model output X, which can be roughly interpreted as the probability of observing X given a particular parameter value θ. Furthermore, let π_ε represent the distribution of the error term that

contributes to the observations. The distribution function of the observations, $F_{D'}(D|\theta,\epsilon) = \mathbb{P}(D' \leq D|\theta,\epsilon)$, is given by

$$\begin{aligned} F_{D'}(D|\theta,\epsilon) &= F_{X+\varepsilon}(D|\theta,\epsilon) \\ &= \mathbb{P}(X+\varepsilon \leq D|\theta,\epsilon) \\ &= \mathbb{P}(\varepsilon \leq D-X|\theta,\epsilon) \\ &= \int_{\mathcal{D}(X)} \int_{\mathcal{D}(\varepsilon)\cap(v\leq D-u)} f_{X,\varepsilon}(u,v|\theta,\epsilon)\mathrm{d}v\mathrm{d}u \\ &= \int_{\mathcal{D}(X)} \int_{\mathcal{D}(\varepsilon)\cap(v\leq D-u)} f(u|\theta)\pi_{\varepsilon}(v|\epsilon)\mathrm{d}v\mathrm{d}u \\ &= \int_{\mathcal{D}(X)} f(u|\theta)F_{\varepsilon}(D-u|\epsilon)\mathrm{d}u. \end{aligned}$$

Taking the derivative of this expression with respect to D, as per the definition of a probability density function, gives us the first result

$$f_{ABC}(D|\theta,\epsilon) \equiv \frac{\mathrm{d}}{\mathrm{d}D}F_{D'}(D|\theta,\epsilon) = \int_{\mathcal{D}(X)} \pi_{\varepsilon}(D-X|\epsilon)f(X|\theta)\mathrm{d}X.$$

The second result can be demonstrated by noting that a Dirac delta function can be thought of as the limiting case of a sequence of distributions centred about origin, which become progressively more concentrated at the origin. Informally,

$$\begin{aligned} \delta(x) &= \lim_{\epsilon\to 0}\pi_{\epsilon}(x), \text{and,} \\ \lim_{\epsilon\to 0}\int_{-\infty}^{\infty} \pi_{\epsilon}(x)f(x)\mathrm{d}x &= f(0). \end{aligned}$$

The sequence of distributions π_{ϵ} can be called *nascent delta functions*. In this thesis we assume the limiting sequence of distributions are Gaussian, with standard

deviation ϵ, but in general the sequence of nascent delta functions can be chosen differently. We use the following informal notation for convenience:

$$\lim_{\epsilon \to 0} \pi_{\varepsilon}(D - X|\epsilon) \equiv \delta(D - X).$$

Therefore

$$\begin{aligned}
\lim_{\epsilon \to 0} f_{ABC}(D|\theta, \epsilon) &= \int_{\mathcal{D}(X)} \lim_{\epsilon \to 0} \pi_{\varepsilon}(D - X|\epsilon) f(X|\theta) \mathrm{d}X \\
&= \int_{\mathcal{D}(X)} \delta(D - X) f(X|\theta) \mathrm{d}X \\
&= f(D|\theta).
\end{aligned}$$

□

Wilkinson's insight allows one to interpret the acceptance probability in the above ABC rejection algorithms (referred to in step 3) as the distribution of the error assumed present in the observations, evaluated at $D - X$. Interestingly, the introduction of the error can be given a meaningful interpretation as either an error owing to some measurement error present in the process of data collection, or an error associated with the model choice (Wilkinson, 2013). The latter interpretation is an appealing one in the context of financial modelling. Financial modellers must acknowledge that the models being used to represent the dynamics of financial variables are not the true data generating processes, but are useful representations that, hopefully, capture much of the salient features one observes in the behaviour of such quantities. Viewed in this light, the ability to explicitly incorporate the possibility that the model being fitted to the data is not a perfect fit for the data could be useful. If the statistician carefully specifies an error distribution based

on knowledge about the likely form of error associated with the observations, and finds that the acceptance rates of the ABC rejection algorithm are too low, it may be necessary to use a more disperse error distribution in order to improve the efficiency of the algorithm (Wilkinson, 2013). Indeed, using the error term as a useful tool to approximate the true posterior associated with the model of interest is the most common way in which ABC methods are currently used.

The basic ABC rejection algorithms outlined above all suffer from a critical problem: if the posterior being approximated and the distribution from which the candidate parameter values are sampled differ significantly (e.g. the sampling density is a standard distribution that is easy to sample from and the posterior is complex and highly localised), the efficiency of the algorithms will be very poor as many of the points sampled from the prior will lie in regions of negligible posterior mass, leading to unacceptably high rejection rates. Given that it is precisely these situations in which ABC methods are most useful, several algorithms, based on Markov chain Monte Carlo (MCMC), sequential Monte Carlo (SMC), and sequential importance sampling (SIS) ideas, have been developed in order to overcome this problem. In what follows we will survey the existing MCMC, SMC, and SIS based methods, and introduce new variants which we have used in my applications.

4.2 Monte Carlo methods for sampling from intractible distributions

To avoid the main problem associated with rejection algorithms (tightly constrained posterior relative to the sampling distribution), three classes of algorithm have been proposed: one based on MCMC; one on SIS; and another related algorithm based on SMC. All three methods involve sampling parameters from a sequence

of sampling distributions that represent the target distribution to a progressively better extent, thereby avoiding the problems associated with sampling and target distributions that differ substantially. General MCMC methods are first introduced and then the practical considerations and difficulties with the method are discussed; then SIS and SMC techniques are introduced and discussed in a similar fashion. The ABC extensions of these algorithms will be outlined in the next section (section 2.2).

4.2.1 MCMC methods

As mentioned previously, MCMC is a powerful technique that allows one to sample from complex distributions, and consequently infer model parameter values. Additionally, expectations taken with respect to target distribution can also be estimated without knowledge of the analytic form of the distribution, which is an important application in the context of financial risk management, for example. MCMC is effectively a method of Monte Carlo integration that utilises Markov chains (Gilks et al., 1996). In order to understand why MCMC produces valid results, we will now provide a brief overview of both Monte Carlo methods and Markov chains.

Monte Carlo integration

Much of the material in this section follows the introduction to Monte Carlo methods given in Glasserman (2010). Monte Carlo methods are based on the connection between probability and volume. The probability of an event is understood as representing the volume that the event takes up in the space of possible events that could have taken place. Monte Carlo methods use this analogy in reverse; by simulating many realisations of random variables, one can approximate the

probability of an event as the ratio of outcomes that coincide with that event to the total number of simulated outcomes. This insight allows one to approximate integrals (i.e. expectations) by simulating n realisations from the distribution of interest, denoted by π_T, and calculating

$$\hat{\mathbb{E}}_n f = \frac{1}{n}\sum_{i=1}^{n} f(\theta_i), \quad \text{where } \theta_i \sim \pi_T.$$

The strong law of large numbers guarantees that this expression converges almost surely to the quantity of interest, i.e.

$$\lim_{n\to\infty} \hat{\mathbb{E}}_n f = \int_{\theta\in\Theta} f(\theta)\pi_T(\theta)\mathrm{d}\theta \equiv \mathbb{E}_{\pi_T} f \quad a.s.$$

There are a variety of methods available for approximating integrals such as this, but Monte Carlo integration comes into its own when one must approximate an integral in high dimensions. In order to appreciate this, consider the central limit theorem. If the integrand f is square integrable[4], then the central limit theorem tells us that[5]

$$\hat{\mathbb{E}}_n f \mathrel{\dot{\sim}} \mathcal{N}(\mathbb{E}_{\pi_T} f, \frac{\sigma_f^2}{n}), \quad \sigma_f^2 = \int_{\theta\in\Theta} (f(\theta) - \mathbb{E}_{\pi_T} f)^2 \pi_T(\theta)\mathrm{d}\theta.$$

In other words, the Monte carlo convergence rate is $\mathcal{O}(n^{-1/2})$ and, most importantly, this convergence rate is independent of the number of dimensions over which the integral is being taken. This feature compares favourably with other methods of approximate integration, whose convergence rate usually decreases as the number of dimensions in the integral increases. Monte Carlo integration is a powerful

[4]That is, if $\int |f(x)|^2 \mathrm{d}x < \infty$

[5]The symbol $\dot{\sim}$, that appears below, means 'approximately distributed'.

method of approximating integrals in high dimensions. The utility of the Monte Carlo method relies, however, on being able to sample points from the distribution that the integral, or expectation, is being taken with respect to. Sampling points from complex, multidimensional distributions is a non-trivial task.

Markov chains

In the interests of presenting as clearly as possible the concepts required to understand MCMC, the basic properties of Markov chains with discrete state-spaces will be outlined in this section; the results can be generalised to apply to more general Markov processes (i.e. Markov chains with uncountable state-spaces) by altering the results slightly in order to take into account the uncountable nature of the state-space. For a more rigorous treatment of Markov processes, see Meyn and Tweedie (1993). Markov chains are a type of stochastic process, consisting of a sequence of discrete-time random variables $(\theta_n)_{n\in\mathbb{Z}^+}$ defined on a common state-space S that is countable. Formally, a discrete-time stochastic process taking values on the countable set S possesses the Markov property if

$$\mathbb{P}(\theta_{n+1} = j|\theta_n = i_n,\theta_{n-1} = i_{n-1}, \ldots, \theta_0 = i_0) = \mathbb{P}(\theta_{n+1} = j|\theta_n = i_n)$$

for all $n \in \mathbb{Z}^+$ and for all $i_0, \ldots, i_n, j \in S$. Plainly, the Markov property states that, given the current state, the future state of a process is indpendent of the history of the process. A Markov chain can be fully characterised by its starting point θ_0 (or, alternatively, an initial distribution) and its one-step transition probabilities $q_{i,j}(n) = \mathbb{P}(\theta_{n+1} = j|\theta_n = i)$. If the Markov chain's transition probabilities $q_{i,j}(n)$ are not time-dependent, the Markov chain is called time-homogeneous and the one-step transition probabilities are then given by $q_{i,j} = \mathbb{P}(\theta_{n+1} = j|\theta_n = i)$, for all $n \in \mathbb{Z}^+$. In what follows, we will use the notation $q_{i,j}^{(n)}$ to represent the probability

that the Markov chain, currently in state i, will visit state j in exactly n time periods. Formally, a time-homogeneous Markov chain can be uniquely defined by:

1. A one-step transition matrix $P = \{q_{i,j}\}_{i,j \in S}$ with $q_{i,j} \geq 0$ for all i, j and $\sum_j q_{i,j} = 1$ for all i;

2. An initial distribution given by $(\lambda_i)_{i \in S}$ with $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$, such that $\mathbb{P}(\theta_0 = i) = \lambda_i$.

The Markov chains that we consider here are time-homogeneous. Note that only the one step transition probabilities are needed to fully specify a Markov chain; two-step ahead transition probabilities can be derived from the one-step ahead probabilities, i.e.

$$\mathbb{P}(\theta_{n+2} = j | \theta_n = i) = (P^2)_{i,j}.$$

And, more generally

$$\mathbb{P}(\theta_{n+m} = j | \theta_n = i) = (P^m)_{i,j}.$$

In order to understand MCMC methods, the factors determining the long run behaviour of Markov chains must be outlined, but firstly some terminology is required. The definitions in this section come from the SMSTC postgraduate lecture series on probability (Wade, 2010).

It is sometimes possible to break a Markov chain in to smaller chunks by considering so-called irreducible closed classes of the Markov chain.

Definition 1. *A non-empty subset C of the state-space S is said to be a closed class if it is not possible to leave C starting from a state within C, i.e. if $q_{i,j} = 0$ for all states $i \in C$ and $j \notin C$.*

Definition 2. *An irreducible closed class C is a closed class such that no proper subset of C is itself closed*[6].

Definition 3. *A Markov chain is called irreducible if the entire state-space S is an irreducible closed class.*

In simple terms, an irreducible Markov chain is one in which every state is reachable from any state within the space S.

Definition 4. *A state $i \in S$ is called transient if the probability of returning to that state at some future point in time is not 1, i.e. the probability of never returning to state i is non-zero. A state $i \in S$ is called recurrent if the probability of returning to that state at some point in the future is 1. More formally, if we define the first passage time for a state $i \in S$ as follows*

$$T_i = min\{n \geq 1 : \theta_n = i | \theta_0 = i\},$$

then a state is transient if

$$\mathbb{P}(T_i = \infty) > 0,$$

and recurrent if

$$\mathbb{P}(T_i = \infty) = 0.$$

Definition 5. *A positive recurrent state $j \in S$ is one for which*

$$\mathbb{E}(T_j) < \infty.$$

We require one more definition before we can discuss the conditions under which a Markov chain will settle down to a steady state in the long-run.

[6] A proper subset, S', of a set, S, is a subset that is strictly contained in S and so necessarily excludes at least one member of the set.

Definition 6. *A state $i \in S$ of a Markov chain has period d if the greatest common divisor (gcd) of the set $\{n : q_{i,i}^{(n)} > 0\}$ is d. Informally this means that the Markov chain in state i can only visit state i again at times md later, where m is some integer. Notably, if $gcd\{n : q_{i,i}^{(n)} > 0\} = 1$ for all $i \in S$ then the Markov chain is said to be aperiodic.*

A natural question to ask is: if we set a Markov chain running and let it evolve for a long period of time, will the behaviour exhibited by the chain settle down to a steady state? This long-run equilibrium concept can be captured in the following definition.

Definition 7. *A stationary distribution π of a time-homogeneous Markov chain is a probability distribution defined on S such that*

$$\pi P = \pi, \tag{4.5}$$

where P is the one-step transition matrix defined earlier. This distribution is also called the stationary or steady state distribution.

From the above definition, it is obvious that if the Markov chain's distribution is π then it will remain in that distribution forever. We next establish the conditions under which a Markov chain, if left to run for a sufficiently long time, will converge

to its stationary distribution, assuming such a distribution exists. More formally, we state the conditions under which the following holds true[7]

$$\lim_{n\to\infty} P^n = \begin{pmatrix} \pi \\ \pi \\ \vdots \\ \pi \end{pmatrix}. \tag{4.6}$$

The following lemma provides the sufficient conditions for this limiting behaviour (Norris, 1997).

Lemma 8. *An irreducible Markov chain has a unique stationary distribution if and only if the chain is positive recurrent. Furthermore, if the Markov chain is aperiodic, then the chain's limiting distribution will equal the unique stationary distribution of the chain. Such a process is sometimes referred to as an ergodic Markov chain.*

One final result from the theory of Markov chains is required in order to progress with an explanation of MCMC methods (Johannes and Polson, 2010).

Lemma 9. *Suppose f is some real valued function with $\int |f| \mathrm{d}\pi < \infty$. If $(\theta_j)_{j\in\mathbb{Z}^+}$ is an ergodic Markov chain with stationary distribution π, then for any initial starting value θ_0*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} f(\theta_j) = \int f(\theta)\pi(\theta)\mathrm{d}\theta, \quad a.s.$$

Note that the samples being used to approximate expectations with respect to the stationary density π are not independent; Markov chains are, by construction,

[7]Recall that π is a row vector, hence the RHS of (4.6) is shorthand for a matrix, with each row corresponding to the stationary distribution of the chain.

correlated random variables. With these results we can now turn our attention to the details behind MCMC.

MCMC

As mentioned previously, Monte Carlo integration is a powerful technique used to approximate expectations with respect to high dimensional distributions, but sampling from high dimensional distributions is not easy. The idea behind MCMC is to construct an ergodic Markov chain with a stationary distribution equal to the complex distribution that one requires samples from. By running such a chain for a sufficiently long period of time, the samples from the chain will (approximately) represent a series of correlated samples from the stationary distribution of the Markov chain. Determining the length of time that the chain must be run for before the statistician can be confident that the Markov chain has converged to the target distribution, usually called the *burn-in* time of the chain, is vitally important; this issue will be discussed in more detail later in this section. Assuming that the chain has converged, the samples from the Markov chain can be used in the so-called ergodic average given in (4.2.1). Hastings (1970), in 1970, developed an algorithm that produces a Markov chain with the desired stationary distribution (the algorithm is itself a generalisation of work done previously by Metropolis et al. (1953)). In what follows, the target distribution (the stationary distribution of the Markov chain) is labelled π_T. The so-called Metropolis-Hastings algorithm takes the point in the Markov chain at stage i, labelled θ_i, and proposes a new point θ' using some proposal density denoted by $q(\theta'|\theta_i)$. Note that the proposal density can be a function of the current state of the chain, and that there are no restrictions on the functional form of the proposal density (Gilks et al., 1996). It is common to use a multivariate Gaussian distribution, centred at the current point in the

process θ_i, with dimension equal to the dimension of the distribution from which one wishes to sample. The candidate point θ' is then accepted with probability

$$\alpha(\theta_i, \theta') = min\left(1, \frac{\pi(\theta')q(\theta_i|\theta')}{\pi(\theta_i)q(\theta'|\theta_i)}\right). \tag{4.7}$$

If the candidate point is accepted, the next point in the chain, θ_{i+1}, is set equal to the candidate point, θ'. If the candidate point is rejected, the current point in the Markov chain is taken as the next point, i.e. $\theta_{i+1} = \theta_i$. The full algorithm is given below:

1. Initialise the chain at some point, θ_0, within the prior's support, set $i = 0$.
2. Sample candidate from $q(\theta'|\theta_i)$.
3. Sample a uniform random variable $U \sim \mathcal{U}(0, 1)$ and set $\theta_{i+1} = \theta'$ if $U \leq \alpha(\theta_i, \theta')$, otherwise, set $\theta_{i+1} = \theta_i$.
4. Set $i = i + 1$ and return to step 2.

This algorithm is guaranteed to generate a Markov chain with stationary distribution π_T. In order to demonstrate this, it is sufficient to show that the Markov chain so constructed satisfies the detailed balance condition (Johannes and Polson, 2010), which is stated below

$$\pi(x)P(x, y) = \pi(y)P(y, x), \tag{4.8}$$

for any $x, y \in S$. As before, π represents the stationary distribution of the Markov chain and P represents the transition density of the chain. Intuitively, this means that the net probability flux between any two states of the chain is zero, or that the probability of getting to state y from x is equal to the probability of getting to state x from y. To see why detailed balance implies that π is a stationary distribution of

the Markov chain with transition density $P(x, y)$, notice that if we assume $x \sim \pi$ then integrating the LHS of (4.8) with respect to x gives the marginal distribution of the state y, which we label $\pi_y(y)$. So we have that

$$\begin{aligned}\pi_y(y) &= \int_{x \in S} \pi(y) P(y, x) \mathrm{d}x \\ &= \pi(y).\end{aligned}$$

This is the continuous state-space analogue of (4.5) which defined a stationary distribution in the discrete state-space Markov chain theory. In order to demonstrate that the Markov chain generated by the Metropolis-Hastings (MH) algorithm satisfies the detailed-balance equations, first note that the transition density of the Markov chain induced by MH is

$$P(x, y) = q(y|x)\alpha(x, y) + \mathbb{1}(x = y)\left(1 - \int_{z \in S} q(z|x)\alpha(x, z)\mathrm{d}z\right).$$

The first contribution on the RHS relates to the scenario in which the chain makes a jump from state x to y and that candidate point is accepted; the second contribution

relates to the scenario in which the chain was already in state y and the candidate point was rejected. Thus

$$\begin{aligned}
\pi(x)P(x,y) &= \pi(x)\left(q(y|x)\alpha(x,y) + \mathbb{1}(x=y)(1-\int_{z\in S} q(z|x)\alpha(x,z)\mathrm{d}z)\right) \\
&= \pi(x)q(y|x)min\left(1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right) + \\
&\qquad\qquad \pi(x)\mathbb{1}(x=y)(1-\int_{z\in S} q(z|x)\alpha(x,z)\mathrm{d}z) \\
&= min(\pi(x)q(y|x), \pi(y)q(x|y)) + \\
&\qquad\qquad \pi(y)\mathbb{1}(x=y)(1-\int_{z\in S} q(z|y)\alpha(y,z)\mathrm{d}z) \\
&= \pi(y)q(x|y)min\left(1, \frac{\pi(x)q(y|x)}{\pi(y)q(x|y)}\right) + \\
&\qquad\qquad \pi(y)\mathbb{1}(x=y)(1-\int_{z\in S} q(z|y)\alpha(y,z)\mathrm{d}z) \\
&= \pi(y)\Big(q(x|y)\alpha(y,x) + \mathbb{1}(x=y)(1-\int_{z\in S} q(z|y)\alpha(y,z)\mathrm{d}z)\Big) \\
&= \pi(y)P(y,x).
\end{aligned}$$

To reiterate, this result proves that the Markov chain induced by MH posseses a stationary distribution π. This alone does not imply that the Markov chain will necessarily converge to this distribution (recall that a Markov chain must be irreducible, positive recurrent and aperiodic in order to guarantee convergence to the stationary distribution). See Gilks et al. (1996) or Johannes and Polson (2010) for a more detailed discussion of the way in which one might demonstrate the ergodicity of Markov chains generated via the Metropolis-Hastings algorithm. Having covered the important concepts underpinning MCMC, a number of important considerations will now be discussed below.

Convergence issues

The fact that the MH algorithm converges to the target distribution, regardless of the functional form of the proposal density $q(\cdot, \cdot)$, is remarkable; however, in practice a judicious choice of proposal density can be the difference between an efficient MCMC algorithm and a horribly inefficient algorithm. Firstly, to maximise computational efficiency, the proposal density should be relatively easy to sample from—MCMC provides a means of sampling from complex distributions, but if the algorithm involves sampling from another difficult distribution then the utility of the algorithm is diminished. In addition, although the functional form of q does not affect whether the chain will converge in theory, the rate at which the chain converges is heavily influenced by the choice of q. Furthermore, even if the chain converges quickly the proposal distribution might result in the state space of the target distribution being explored very slowly, requiring the algorithm to be run for a long period of time to give the chain time to explore different regions of the target distribution in the correct proportions. A chain that explores the space of the target distribution slowly is called *slow mixing*; a chain that explores the target distribution quickly is called *fast mixing*. Therefore, the proposal density should be chosen such that: it is relatively easy to sample from; the chain converges at an acceptable rate; and such that the chain exhibits fast mixing. In practice, the choice of q often results from experimentation and craftmanship (Gilks et al., 1996). Two classes of proposal density are now briefly outlined:

Metropolis Algorithm: this variant of the MH algorithm involves using a symmetric proposal density , i.e. $q(y|x) = q(x|y)$, which simplifies the acceptance probability used in the algorithm

$$\begin{aligned}\alpha(y, x) &= min\left(1, \frac{\pi_T(y)q(x|y)}{\pi_T(x)q(y|x)}\right) \\ &= min\left(1, \frac{\pi_T(y)}{\pi_T(x)}\right).\end{aligned}$$

A special case of the Matropolis algorithm is the random-walk proposal, represented by $q(y|x) = q(|y - x|)$. A multi-dimensional Gaussian proposal density is an example of a random-walk Metropolis algorithm, and it is the approach used in the applications covered later in this thesis. When using a random-walk proposal, the efficiency of the algorithm will usually depend on some scale parameter (Gilks et al., 1996). If a multivariate Gaussian proposal is used, the scale parameter corresponds to the covariance matrix of the proposal distribution. If the scale is chosen to be too small, the Markov chain will make small jumps around the state-space; this will likely result in high acceptance rates (the ratio of accepted moves to total moves proposed) but poor mixing. A scale parameter that is too large results in large proposed jumps around the state-space, often to the tails of the target distribution, and low acceptance rates, again resulting in poor mixing. Typically, the scale parameter is scaled by trial and error to arrive at a proposal density that avoids both extremes (Gilks et al., 1996). For particularly complex or high dimensional target densities, an initial exploratory stage is usually implemented in order to get a rough idea of the location of the modes, and the covariance structure of the target distribution—this information then allows the statistician to design proposal distributions with appropriate scale

parameters, and pick promising regions of the sample space to initialise the MCMC algorithm, both of which should aid in speeding up convergence rates and improving the chain's mixing properties (Gilks et al., 1996).

Independence Sampler: this algorithm involves using a proposal density that is independent of the current point, i.e. $q(y|x) = q(y)$. This results in an acceptance probability of the form

$$\alpha(y, x) = min\left(1, \frac{w(y)}{w(x)}\right), \quad \text{where } w(x) = \frac{\pi_T(x)}{q(x)}.$$

The effectiveness of the independence sampler depends on the match between the proposal density and the target density; typically, one should look for a proposal density that is similar to the target, but with fatter tails (Gilks et al., 1996). Fatter tails reduce the chance of the Markov chain getting stuck in low probability regions of the target density, and therefore promote faster mixing of the chain.

In this section we have stated that the proposal density should be chosen with the goal of maximising the convergence rate of the algorithm in mind. A key consideration is how to determine if the chain has converged. As mentioned earlier, only once the Markov chain induced by MCMC has converged are we justified in using the samples to compute Monte Carlo estimates of expectations. The Markov chain needs to be left to run for some initial period (the *burn-in* period) in order to give the chain a chance to settle down to its stationary distribution, but how long should the burn-in be? This question is considered in more detail below.

Determining the burn-in period

After using the Metropolis-Hastings algorithm to construct a Markov chain with the correct stationary distribution, the chain should be initialised and run until it has converged after, say, m time-steps. We then run the chain for a further $n - m$ time-steps to obtain a sample $\{\theta_i\}_{i=m+1,\ldots,n}$ from the target distribution, which we can then use to evaulate Monte Carlo approximations via

$$\hat{\mathbb{E}}f = \frac{1}{n-m} \sum_{i=m+1}^{n} f(\theta_i).$$

In theory, it is sometimes possible to analytically determine the required burn-in length; however, the calculations involved are far from trivial and are not usually practical or possible in more complex examples (Gilks et al., 1996). Aside from the small number of situations in which we can analytically determine the necessary burn-in period, there is currently no way to guarantee, using only the output from the chain, that a Markov chain has converged to its stationary distribution. Despite this, there are a number of techniques commonly used in practice. One method is to visually inspect the Markov chain to see if it has converged. This approach, although intuitive, is unfortunately not adequate. Even if the Markov chain appears to settle down around some region of the state-space, there is nothing about this behaviour that guarantees convergence; the chain could have, for example, become trapped in a local mode, or it might be very slow mixing, giving the impression that it has converged when in fact it has not. This approach also quickly becomes impractical when the dimension of the target distribution increases. A more reliable approach is to design some convergence diagnostic that gives a more objective, and reliable indication of convergence. One such example is the Gelman-Rubin (GR) statistic. The GR convergence diagnostic involves running a small number of MCMC chains

simulataneously, initialised from dispersed starting positions in the state-space, and periodically measuring the similarity between each of the chains. Only when the distributions of each of the chains are similar to the combined distribution of the sample can the chains have converged to the stationary distribution (Gelman et al., 2003). Note that the GR diagnostic does not guarantee convergence, but it can be a useful statistic that assists in determining convergence. In the case where the Markov chain generated by MCMC is multidimensional (say, dimension p), the Gelman-Rubin statistic should be calculated for each element of the p-dimensional Markov chain, and the chain should be run until each statistic corresponding to each element has indicated convergence. Assume that there are m Markov chains of length n generated by MCMC and that each chain is p-dimensional. The following steps describe how the Gelman-Rubin statistic should be calculated for each element of the p-dimensional Markov chain:

1. Run m chains from starting points dispersed across the state-spate. Denote the ith sampled point from the jth chain by $\psi_{i,j}$.

2. After running each chain for a certain length of time (e.g. 1000 iterations), discard the first half of the samples from each chain - this is the burn-in.

3. With the remaining n sampled points from each of the m chains, calculate the following quantities

$$B = \frac{n}{m-1}\sum_{j=1}^{m}(\bar{\psi}_{\cdot,j} - \bar{\psi}_{\cdot,\cdot})^2, \text{ where } \bar{\psi}_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n}\psi_{i,j}, \text{ and } \bar{\psi}_{\cdot,\cdot} = \frac{1}{m}\sum_{j=1}^{m}\bar{\psi}_{\cdot,j}$$

$$W = \frac{1}{m}\sum_{j=1}^{m}s_j^2, \text{ where } s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\psi_{i,j} - \bar{\psi}_{\cdot,j})^2.$$

4. Calculate an estimate of the target distribution's variance using

$$\widehat{var}^{+}(\psi) = \frac{n-1}{n}W + \frac{1}{n}B.$$

5. Calculate $\hat{R} = \sqrt{\frac{\widehat{var}^{+}(\psi)}{W}}$.

6. If $\hat{R}$ is above some threshold (usually around 1.1 or 1.05), run the chain for another n iterations and re-calculate $\hat{R}$. Repeat until statistics for all p elements of the Markov chain are below the threshold.

Note that, assuming the m Markov chains were initialised using dispersed starting positions, the variance estimate $\widehat{var}^{+}(\psi)$ overestimates the variance of the target distribution, but is an unbiased estimator of the target variance as $n \to \infty$. The within-chain variance statistic, W, is an underestimate of the variance of the target distribution[8], but is also an unbiased estimate in the limit $n \to \infty$. Therefore, the Gelman-Rubin statistic, $\hat{R}$, should decline towards 1 as $n \to \infty$ as numerator and denominator approach the same value from above and below respectively.

Thinning

Once the Markov chain has converged sufficiently, the MCMC algorithm produces a sequence of correlated samples from the approximate target distribution. By virtue of the ergodic theorem introduced earlier, correlation among the samples does not stop us from using the samples to generate Monte Carlo estimates as described earlier; however, when the dimensionality of the Markov chain being generated is large, it can be preferrable to only store every kth draw (where k is some integer) from the MCMC output in order to mitigate the practical problems associated

[8]This is the case because the individual Markov chains have not yet had the chance to traverse the full state-space.

with storing large amounts of data. The process of only selecting every kth iterate is known as *thinning*.

4.2.2 Sequential importance sampling

Having now outlined the salient features of Markov chain Monte Carlo, we will now survey the first of two alternative but related Monte Carlo based methods for sampling from intractible distributions: sequential importance sampling (SIS). In the next section we will cover sequential Monte Carlo (SMC), which is described in detail in Del Moral et al. (2006). These methods have certain advantages over MCMC, namely they are not hampered by the problems associated with assessing convergence of a Markov chain, and they avoid the complications of MCMC algorithms becoming stuck in local modes of the target distribution. In addition, these sequential methods are readily parallelisable, i.e. the computational efficiency scales with the computing power available to the statistician. The shortfalls of both SIS and SMC will be considered after the ideas have been outlined.

SIS is a method of generating samples from intractible distributions that, as the name of the method suggests, involves using well-known importance sampling techniques to sequentially move through a series of distributions, starting with a sample from an easy to sample from distribution and moving the sample in such a way as to end up with a sample from the target distribution. SIS is based on importance sampling, the main ideas associated with which are given in the following section. In what follows, we will refer to individual sampled points as *particles*, in keeping with the terminology used by others discussing SIS and SMC.

Importance sampling

Importance sampling is a powerful technique that can be used to estimate properties of a particular distribution. In particular, the technique can be used to generate samples from distributions that are otherwise difficult to sample from. In what follows we will denote by f the (unnormalised) density function of the distribution Π_T that we wish to obtain samples from. The technique of using importance sampling to derive samples from intractible distributions is usually referred to as sequential importance resampling (SIR) (Bernardo and Smith, 2000). Assume that a sample is required from the following probability density

$$\pi_T(\theta) = \frac{f(\theta)}{\int f(\theta) \mathrm{d}\theta},$$

where, as stated above, only the functional form of f is known (the normalising constant is not available). Assuming that a sample from g, some other probability density that is easy to sample from, is available, SIR can be utilised to generate samples from π_T. There are two cases to consider:

1. There exists an identifiable constant M, such that

$$\frac{f(\theta)}{g(\theta)} \leq M, \quad \text{for all } \theta.$$

2. The bound M is not available.

In the first case, samples from π_T can be readily obtained via a simple rejection algorithm (Bernardo and Smith, 2000):

1. Consider a particle $\theta_i \sim g$.

2. Generate a uniform random variable $U \sim \mathcal{U}(0,1)$.

3. If
$$U \leq \frac{f(\theta_i)}{Mg(\theta_i)},$$
accept the particle θ_i as a sample from π_T, otherwise reject the simulated particle.

In the second case, where the upper bound M is not available, one can still derive an approximate sample from π_T via the following steps:

1. For each particle in the sample from g, calculate a weight given by
$$q_i = \frac{w_i}{\sum_{j=1}^{n} w_j} \quad \text{where } w_i = \frac{f(\theta_i)}{g(\theta_i)}. \tag{4.9}$$

2. Draw particle θ_i from the sample with probability q_i.

To see that the sampled points are (approximately) distributed according to π_T, observe that the distribution function of the particle sample is given by (Bernardo and Smith, 2000)

$$\begin{aligned}
\mathbb{P}(\theta \leq a) &= \sum_{i=1}^{n} q_i \mathbb{1}(\theta_i \leq a) \\
&= \sum_{i=1}^{n} \frac{w_i}{\sum_{j=1}^{n} w_j} \mathbb{1}(\theta_i \leq a) \\
&= \frac{n^{-1} \sum_{i=1}^{n} w_i \mathbb{1}(\theta_i \leq a)}{n^{-1} \sum_{j=1}^{n} w_j}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(\theta \leq a) &= \frac{\mathbb{E}_g\left[w(x)\mathbb{1}(\theta \leq a)\right]}{\mathbb{E}_g\left[w(x)\right]} \\ &= \frac{\int_{-\infty}^{a} f(x)\mathrm{d}x}{\int_{-\infty}^{\infty} f(x)\mathrm{d}x} \\ &= \int_{-\infty}^{a} \pi_T(x)\mathrm{d}x \\ &= \Pi_T(a),\end{aligned}$$

which is the distribution function associated with the density π_T – the target from which we wish to sample. It is also interesting to note that, when sampling particles from the sample generated from g, resampling with replacement does not jeopardise the algorithm's ability to generate samples from the target density π_T, and therefore samples from the target density can be as large as desired. However, if g is not a good match to the target, the variance of the importance weights will be large, which will adversely affect the quality of the approximation to the target density.

Alternatively, importance sampling can be used to estimate integrals. In Bayesian statistics, for example, importance sampling can be used to estimate the normalising constant that appears in the denominator of Bayes' Theorem (4.1), or to estimate expectations taken with respect to the posterior distribution. The rationale behind importance sampling can be illustrated by the following observation. If b is a function and G is some probability distribution with density g, i.e. $G(\mathrm{d}x) = g(x)\mathrm{d}x$, then

$$\begin{aligned}\int b(x)\mathrm{d}x &= \int \left[\frac{b(x)}{g(x)}\right] g(x)\mathrm{d}x \\ &= \mathbb{E}_G\left[\frac{b}{g}\right].\end{aligned}$$

This suggests that the integral can be approximated by drawing a n sample points $\{x_i\}_{i=1,\ldots,n}$ from g which we choose to be a relatively straightforward density to sample from, and computing the quantity

$$\hat{\mathbb{E}}_G\left[\frac{b}{g}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{b(x_i)}{g(x_i)}, \tag{4.10}$$

which is an unbiased estimator of the expectation under G. In many cases, the integrand b is the product of some other function h and another probability density f, i.e. the integral being approximated is an expectation taken with respect to the distribution $F(\mathrm{d}x) = f(x)\mathrm{d}x$, which is impossible or computationally impractical to sample from directly. Just as the quality of matching between the sampling density, g, and target density, f, was an important factor affecting the quality of samples derived from the SIR procedure above, the variance of the unbiased estimator (4.10), and hence its reliability, clearly depends on the choice of g; if g is similar to b then the ratio of b over g will vary very little. Therefore, importance sampling works best when a sampling distribution g can be chosen that is similar in shape to the integrand b.

Sequential Importance Sampling

As mentioned previously, the quality of importance sampling estimates depends on the degree of similarity between the sampling and target densities; a sampling density that is not a good fit to the target density will produce a large variation in importance weights, which results in estimators derived via the importance sampling particles that exhibit increased variance. Typically, posterior distributions that play the role of the target distribution in Bayesian applications are complex, and it is often difficult to pick a sampling distribution that results in particle populations with low importance weight variance. To alleviate the difficulty in choosing good

sampling distribtions, SIS algorithms introduce a sequence of distributions that vary gradually between a distribution that is easy to sample from and the intractible target distribution. In theory, if each distribution in the sequence is sufficiently similar to each of its neighbours, then it should be possible to perform importance sampling between each neighbouring pair of distributions in sequence, using the sample obtained from one distribution as the sampling density for the next target distribution in the sequence. By sequentially moving the particles around in this manner, one generates a sample from the approximate target distribution. In keeping with the notation used in Del Moral et al. (2006), let the sequence of target distributions that we wish to sample from sequentially be labelled $\{\pi_n\}_{n=1,\ldots,p}$. Define a sequence of importance distributions that we will use to generate samples from the target distributions be labelled $\{\eta_n\}_{n=1,\ldots,T}$. Each particle in the sample population will be denoted by $\theta_n^{(i)}$, for $i = 1\ldots, N$. Assume the sequence of SIS distributions, $\{\pi_n\}_{n=1,\ldots,T}$, are defined on a common measurable space (E, ε). Let $K_n : E \times \varepsilon \to [0, 1]$, $n = 1, \ldots, T$, represent a sequence of Markov kernels, each with associated density $k_n(\theta, \theta')$. Finally, note that the marginal distribution of the particle population after being perturbed by the Markov kernel is given by

$$\eta_n(\theta') = \int_E \eta_{n-1}(\theta) k_n(\theta, \theta') \mathrm{d}\theta. \tag{4.11}$$

1. Initialisation: Set $n = 1$. Generate a particle sample of size N from the initial distribution in the SIS sequence, π_1, and set each particle weight equal to 1[9]. The initial population is denoted by

$$\left(\hat{\theta}_1^{(i)}, 1\right)_{i=1,\ldots,N}.$$

[9]Assuming $\eta_1 = \pi_1$ i.e. the initial importance distribution is usually chosen to coincide with the first sampling distribution, the variance of importance weights is zero (all unnormalised weights are equal to 1).

2. Set $n = n + 1$. Perturb particles $i = 1, \ldots, N$ using a Markov kernel with density $k_n(\hat{\theta}_{n-1}^{(i)}, \cdot)$ to obtain

$$\theta_n^{(i)} \sim k_n(\hat{\theta}_{n-1}^{(i)}, \cdot) \quad i = 1, \ldots, N.$$

3. Evaluate importance weights and normalise:

$$w_n^{(i)} = \frac{\pi_n(\theta_n^{(i)})}{\eta_n(\theta_n^{(i)})}, \quad W_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}, \tag{4.12}$$

where $\eta_n(\theta_n^{(i)})$ is defined in (4.11).

4. Resample N times from the population $\left(\theta_n^{(i)}, W_n^{(i)}\right)$, sampling in proportion to the normalised importance weights $(W_n^{(i)})_{i=1,\ldots,N}$, then reset all weights to $1/N$. The new population is denoted by

$$\left(\hat{\theta}_n^{(i)}, 1\right)_{i=1,\ldots,N}.$$

5. If $n = T$ stop, otherwise return to step 2.

It is assumed that step 1 in the above algorithm is straightforward, i.e. it is easy to sample from the initial distribution in the sequence of SIS distributions; if this is the case then the initial population targets the initial distribtion in the sense that

$$\lim_{N\to\infty} \frac{\sum_{i=1}^N w_1^{(i)} \psi(\hat{\theta}_1^{(i)})}{\sum_{i=1}^N w_1^{(i)}} = \mathbb{E}_{\pi_1}\psi(\theta), \quad a.s.$$

which follows directly from the strong law of large numbers. Once the particles are perturbed using the Markov kernel in step 2, the population now targets[10] the

[10]The term 'targets' should be interpreted in the same sense as before, i.e. estimates of expectations with respect to the sampling distribution converge almost surely to the correct value.

sampling distribution denoted by η_n, the expression for which is given above. The particle weights are then corrected in step 3 so that the particle population targets the next SIS distribution in the sequence. It is a trivial exercise to demonstrate that the re-weighted particle population obtained after stage 3 targets the distribution π_n:

$$\begin{aligned}\lim_{N\to\infty} \frac{\sum_{i=1}^{N} w_n^{(i)}\psi(\theta_n^{(i)})}{\sum_{i=1}^{N} w_n^{(i)}} &= \frac{\mathbb{E}_{\eta_n}(w_n\psi(\theta_n))}{\mathbb{E}_{\eta_n}(w_n)} \\ &= \frac{\int_E w_n\psi(\theta)\eta_n(\theta)\mathrm{d}\theta}{\int_E w_n\eta_n(\theta)\mathrm{d}\theta} \\ &= \frac{\int_E \psi(\theta)\pi_n(\theta)\mathrm{d}\theta}{\int_E \pi_n(\theta)\mathrm{d}\theta} \\ &= \mathbb{E}_{\pi_n}\psi(\theta).\end{aligned} \tag{4.13}$$

Step 4 in the SIS algorithm serves to remove those particles in the population with small weights and replace them with particles that are more representative of the target distribution (particles with larger weights). There are several methods of carrying out this selection stage of the SIS algorithm; in our application we use a multinomial sampling technique which amounts to drawing each $\theta_n^{(i)}$ independently from a multinomial distribution with probability $W_n^{(i)}$ (see Chopin (2004) for further details, including a brief survey of other approaches to selection of particles).

Practical difficulties and other considerations

In the SIS algorithm outlined above, the Markov kernel used to propagate the particle population forwards through the sequence of intermediate distributions was not defined explicitly. A random-walk kernel was used in the applications considered in this thesis, i.e. we use a Gaussian transition kernel centered at the

current particle value with a kernel bandwidth (the standard deviation parameter) equal to the empirical standard deviation of the most recently generated particle population. The reason for this particular choice will become clear when the difficulties associated with the SIS method are discussed below. There are other possibilities for the Markov kernel; for example, an MCMC kernel (of the type defined in the previous section on MCMC methods) could be used in place of a Gaussian kernel. For a more detailed overview of the choices available, see Del Moral et al. (2006).

The SIS algorithm outlined in the previous section suffers from one significant drawback: when calculating the importance weights (step 3), the marginal probability density of the perturbed particles η_n has to be evaluated. Recall that the expression for this density is

$$\eta_n(\theta_n) = \int_E \eta_{n-1}(\theta) k_n(\theta, \theta_n) \mathrm{d}\theta. \tag{4.14}$$

In most cases, this integral will be impossible to evaluate analytically and therefore evaluating the importance weights, which is necessary if the particle population is to target the correct distribution, is not possible. One way to circumvent this problem is to estimate (4.14) using a Monte Carlo estimate, i.e.

$$\eta_n(\theta_n) \approx \frac{1}{N} \sum_{i=1}^{N} k_n(\hat{\theta}_{n-1}^{(i)}, \theta_n).$$

As pointed out in Del Moral et al. (2006), this approach comes with its own difficulties: firstly, it increases the computational complexity of the algorithm, requiring an extra layer of calculations in order to approximate the marginal distribution above; secondly, there are cases in which the transition kernel cannot be computed pointwise analytically, for example if the Markov kernel is an MCMC

kernel. The latter difficulty is the main motivating factor behind choosing a Gaussian kernel for the applications considered later—Gaussian kernels can be evaluated pointwise. The former difficulty is not a major problem considering the applications in which the SIS algorithm is being used in this thesis. In the applications considered in the succeeding chapter, the SIS algorithm is integrated with the ABC approximation to the likelihood function (the combination being labelled *ABC SIS*) that was outlined earlier. As explained in Beaumont et al. (2009), the main bottleneck in computer code associated with ABC sampling algorithms arises as a result of having to simulate data sets from the model under investigation; the additional computational expense incurred by approximating the marginal distribution of the perturbed particles contributes relatively little to the overall complexity of the ABC SIS algorithm and is therefore not a major issue.

One final thing to note about the SIS algorithm outlined above is that step 4—the resampling of the particle population in proportion to their importance weights—is not strictly necessary for the algorithm to target the distribution of interest (as demonstrated by (4.13)). Recall that the motivation for step 4 was to discard particles with small importance weights and replace them with particles with larger weights; the rationale being that particles with large weights represent samples from the target distribution to a greater extent than particles with small importance weights. Consider the case where no resampling is carried out, i.e. at each stage of the algorithm, each particle is perturbed using the Markov kernel K_n and then re-weighted using the importance weight formula that was made explicit earlier, before repeating the process at the next stage of the algorithm. In this case, the sampling density at stage n of the SIS algorithm is given by

$$\eta_n(\theta_n) = \int \eta_1(\theta_1) \prod_{j=2}^{n} k_j(\theta_{j-1}, \theta_j) \mathrm{d}\theta_{1:n-1},$$

which, as n increases, one would expect to diverge more and more from the target distribution π_n, leading to increasing variance of the importance weights (sometimes referred to as a degeneracy in the particle population) (Del Moral et al., 2006). One solution is to resample at every stage of the algorithm, in line with the approach outlined earlier in this subsection; however, this approach can actually have a detrimental impact on the quality of the samples derived from the algorithm. In the case where the variance of the importance weights is small, resampling is unneccessary and typically wasteful as the resampling introduces some variance to the sample, without discernibly reducing the variance of importance weights. Ideally, resampling should only be carried out when the variance of the importance weights becomes unacceptably high. One way to incorporate this optional resampling step into the SIS algorithm is to monitor the *effective sample size* (or ESS) of the particle population, and only resample if the ESS falls below some threshold value. The ESS calculates the equivalent number of (unweighted) samples from the target distribution that would give rise to the same Monte Carlo error as the weighted particle sample derived via SIS (Sisson et al., 2007). The ESS can be estimated by

$$ESS = \left[\sum_{i=1}^{N} (W_n^{(i)})^2\right]^{-1},$$

where, as before the $W_n^{(i)}$ represent the normalised importance weights in the SIS algorithm. The ESS estimate lies in the range $1 \leq ESS \leq N$. If the particle population is not degenerate (i.e. the importance weights do not have a large variance) then the ESS will be large – in the extreme case when all weights are equal, the ESS equals the SIS population size N; at the other extreme, where only one particle has mass, the ESS reduces to 1. With this measure of population degeneracy, the fourth step in the SIS algorithm can be altered to

4 If $ESS < E$, where E is some threshold[11], then resample the particle population in proportion to the importance weights. After resampling, reset all importance weights to $1/N$.

With this extra calculation, the SIS algorithm only resamples the population when the population degenerates below some level which is determined by the statistician, i.e. resampling is only carried out when the benefits of doing so (reduced sample degeneracy) outweigh the costs (additional variance associated with resampling).

4.2.3 Sequential Monte Carlo sampling

In this section we will give an overview of the ideas behind the third and final Monte Carlo based method of sampling that we consider in this thesis: sequential Monte Carlo (SMC) sampling. SMC is closely related to the SIS algorithm outlined in the previous section, and aims to avoid the major difficulty associated with SIS; namely the evaluation of the sampling density that appears in the denominator of the importance weight formula (4.12). Recall that in the previous section it was highlighted that evaluating the marginal sampling density η_n was typically not possible, and that approximating this density by Monte Carlo resulted in an increase in computational complexity - a cost that, ideally, one would like to avoid. One of the key features that differentiates SMC methods from SIS methods is that the sequence of intermediate distributions that are used to produce samples from the (approximate) target distribution have an increasing dimension in SMC, whereas in SIS the sequence of distributions have a common state space E. The key step in constructing SMC samplers, as per Del Moral et al. (2006), is to introduce a series of backwards Markov kernels $L_{n-1} : E \times \varepsilon \to [0, 1]$, for $n = 2, \ldots, T$, with associated density $L_{n-1}(\theta_n, \theta_{n-1})$, to build a sequence of distributions with a fixed

[11] $E = N/2$ is typically chosen as a threshold ESS.

marginal distribution equal to the target distribution. Defining the artificial joint target distribution by

$$\tilde{\pi}_n(\theta_{1:n}) = \frac{\tilde{\gamma}_n(\theta_{1:n})}{Z_n}, \quad \tilde{\gamma}_n(\theta_{1:n}) = \gamma_n(\theta_n) \prod_{k=1}^{n-1} L_k(\theta_{k+1}, \theta_k) \tag{4.15}$$

where Z_n is the normalising constant of the target distribution. The dimension of these new target distributions increases over time, i.e. $\tilde{\pi}_n$ is defined on E^n and therefore this sequence of distributions is amenable to sampling by SMC methods. Furthermore, the marginal densities associated with the new targets are always equal to the target density of interest i.e. $\pi_n(\theta_n)$. This is easily demonstrated by integrating out the previous particle values, i.e.

$$\begin{aligned}
\int_{E^{n-1}} \tilde{\pi}_n(\theta_{1:n}) \mathrm{d}\theta_{1:n-1} &= \frac{\gamma_n(\theta_n)}{Z_n} \int_{E^{n-1}} \prod_{k=1}^{n-1} L_k(\theta_{k+1}, \theta_k) \mathrm{d}\theta_{1:n-1} \\
&= \pi_n(\theta_n) \int_{E^{n-2}} \left[\int_E L_1(\theta_2, \theta_1) \mathrm{d}\theta_1 \right] \prod_{k=2}^{n-1} L_k(\theta_{k+1}, \theta_k) \mathrm{d}\theta_{2:n-1} \\
&= \pi_n(\theta_n) \int_{E^{n-2}} \prod_{k=2}^{n-1} L_k(\theta_{k+1}, \theta_k) \mathrm{d}\theta_{2:n-1} \\
&\vdots \\
&= \pi_n(\theta_n).
\end{aligned}$$

Similar to the importance sampling case, define the unnormalised importance weights as the ratio of target density to sampling density:

$$w_n(\theta_{1:n}) = \frac{\tilde{\gamma}_n(\theta_{1:n})}{\eta_n(\theta_{1:n})}, \tag{4.16}$$

noting that the importance and sampling densities are now associated with the path of a particle $\theta_{1:n}$ as it is moved through the sequence of importance distributions,

and not just the most recent value of the particle θ_n. At time $n-1$, assume that a particle population $\{\theta_{1:n-1}^{(i)}, W_{n-1}^{(i)}\}$ is available that targets[12] the distribution $\tilde{\pi}_{n-1}$. At time n, the path of each particle is extended via the forward Markov kernel density $k_n(\theta_{n-1}, \theta_n)$ and the importance weights recalculated to correct for the discrepancy between the sampling and target densities. The unnormalised weight functions at time n can be obtained by multiplying the unnormalised importance weights at time $n-1$ by the incremental weight function

$$\tilde{w}_n(\theta_{n-1}, \theta_n) = \frac{\gamma_n(\theta_n) L_{n-1}(\theta_n, \theta_{n-1})}{\gamma_{n-1}(\theta_{n-1}) k_n(\theta_{n-1}, x_n)}. \tag{4.17}$$

The expression for the incremental weight can be obtained from the expression for the unnormalised weight function

$$w_n(\theta_{1:n}) = w_{n-1}(\theta_{1:n-1}) \tilde{w}_n(\theta_{n-1}, \theta_n).$$

It is a straightforward exercise to demonstrate that the populations of particles generated in this manner target the artificial joint density $\tilde{\pi}_n$. For clarity, the SMC algorithm is outlined below.

1. Initialisation:

   - Set $n = 1$.
   - For $i = 1, \ldots, N$ draw $\theta_1^{(i)} \sim \eta_1$.
   - Evaluate $\left\{ w_1(\theta_1^{(i)}) \right\}$ and normalise these weights to obtain $\left\{ W_1(\theta_1^{(i)}) \right\}$.

   Iterate steps 2 and 3.

2. Resampling

[12] As before, 'targets' should be interpreted as meaning the empirical density of the particles converges to $\tilde{\pi}_{n-1}$ as the number of particles N goes to infinity.

- If $ESS < T$ (for some threshold T), resample the particles and set $W_n^{(i)} = 1/N$.

3. Sampling

- Set $n = n + 1$, if $n = T + 1$ stop.
- For $i = 1, .., N$ draw $\theta_n^{(i)} \sim K_n(\theta_{n-1}^{(i)}, \cdot)$.
- Evaluate the incremental weight function for each particle using (4.17) and normalise the particle weights

$$W_n^{(i)} = \frac{W_{n-1}^{(i)} \tilde{w}_n(\theta_{n-1}^{(i)}, \theta_n^{(i)})}{\sum_{j=1}^{N} W_{n-1}^{(j)} \tilde{w}_n(\theta_{n-1}^{(j)}, \theta_n^{(j)})}$$

Notice that the SMC algorithm also makes use of the conditional resampling step introduced in the SIS section in order to prevent the discrepancy between sampling and target densities resulting in particle degeneracy. This SMC algorithm avoids having to evaluate the marginal sampling distributions η_n and is thus an attractive alternative to the SIS algorithm. As with SIS, there are a variety of options at the statistician's disposal when it comes to the choice of the forward Markov kernel. For a detailed discussion concerning the choice of forward Markov kernel, see Del Moral et al. (2006). Applications considered in this thesis all make use of MCMC kernels for the forward kernel in the SMC algorithm; MCMC kernels are based on the Metropolis-Hastings (MH) accept-reject step that appears in MCMC algorithms outlined previously. Consider a particle $\theta_n^{(i)}$ in the population sample at time n. At the sampling stage (second bullet point in stage 3 of the SMC algorithm) the particle is perturbed by the Markov kernel $k_{n+1}(\theta_n^{(i)}, \cdot)$ in order to generate a particle that approximates the next target distribution π_{n+1}; if the Markov kernel is a MCMC kernel this involves the following steps:

1. Generate a candidate point θ' using a Markov transition kernel $q(\theta'|\theta_n^{(i)})$.

2. Evaluate the MH acceptance probability

$$\alpha(\theta_n^{(i)}, \theta') = min\left(1, \frac{\pi(\theta')q(\theta_n^{(i)}|\theta')}{\pi(\theta_n^{(i)})q(\theta'|\theta_n^{(i)})}\right)$$

3. Sample $U \sim \mathcal{U}(0,1)$ and set $\theta_{n+1}^{(i)} = \theta'$ if $U \leq \alpha(\theta_n^{(i)}, \theta')$, otherwise set $\theta_{n+1}^{(i)} = \theta_n^{(i)}$.

This series of steps can be iterated several times, or simply implemented once. As discussed earlier in the chapter, this choice of MCMC kernel generates a Markov chain that converges to the target distribution at time n, π_n, and is therefore a natural choice for the Markov kernel in SMC algorithms. In addition to the attractive convergence properties of such kernels, one can utilise the significant body of knowledge concerning the design of efficient MCMC moves that exists in the MCMC literature to design efficient sampling distributions (Del Moral et al., 2006). As in the MCMC case, we chose a symmetric (Gaussian) proposal density (labelled q above) centred at the current particle value $\theta_n^{(i)}$ with covariance matrix equal to the covariance of the particle population at time n.

The remaining feature of the SMC algorithm that requires consideration is the choice of backwards Markov kernels $L_{n-1}(\theta_n, \theta_{n-1})$. The particluar form of the backwards kernels is arbitrary—the validity of the SMC algorithm does not depend on the particular form of the kernel; however, just as the form of proposal density in MCMC algorithms should be chosen to ensure the algorithm is efficient, the backwards Markov kernels in SMC algorithms should be optimised with respect to the forward Markov kernel used in the algorithm. There exists an expression

for the optimal[13] form of the sequence of backwards kernels L_n^{opt}, $n = 1, \ldots, p$ (Del Moral et al., 2006)

$$L_{n-1}^{opt} = \frac{\eta_{n-1}(\theta_{n-1})k_n(\theta_{n-1}, \theta_n)}{\eta_n(\theta_n)}, \tag{4.18}$$

but this expression involves the evaluation of the marginal sampling densities η_n and η_{n-1} which, as already discussed, is difficult or impossible in practice, and is the main motivating factor driving the development of the SMC algorithm. With the optimal backwards kernel being unavailable, a sensible approach is to attempt to approximate (4.18) in some way. As discussed in Del Moral et al. (2006), there are a number of ways in which the optimal sequence of kernels can be approximated. If the forward Markov kernel is a MCMC kernel then Del Moral et al. (2006) suggest using the following approximation

$$L_{n-1}(\theta_n, \theta_{n-1}) = \frac{\pi_n(\theta_{n-1})K_n(\theta_{n-1}, \theta_n)}{\pi_n(\theta_n)},$$

which reduces the expression for the incremental importance weights to

$$\tilde{w}_n(\theta_{n-1}, \theta_n) = \frac{\gamma_n(\theta_{n-1})}{\gamma_{n-1}(\theta_{n-1})}.$$

Note that the expression for the incremental weights is now independent of the proposed point θ_n; in this case, the SMC algorithm outlined above should be altered so that the importance weights are calculated, and any resampling carried out, before the new candidate point is generated (Del Moral et al., 2006). Although the SMC algorithm described in this section offers attractive practical qualities, namely the avoidance of the evaluation of sampling distributions associated with the SIS algorithm, when the SMC algorithm is incorporated with the ABC likelihood approximation (called *ABC SMC*), difficulties arise that render this algorithm

[13] Optimal in the sense that the variance of the importance weights $w_n(\theta_{1:n})$ is minimised.

unsuitable for deriving samples from the ABC approximated posterior distribution. These difficulties will be discussed in detail in the next section.

4.3 Combining ABC methods with Monte Carlo methods

The overall objective that is considered in this chapter is to estimate the parameters of financial models (SDEs) by using Bayesian inference techniques. The primary obstacle that often prevents straightforward Bayesian techniques being implemented is the lack of an analytical expression for the likelihood of observations from the model being estimated. Consequently, the analytical expression for the posterior distribution of parameters (the key distribution of interest in Bayesian inference) given by (4.1) is also unavailable. We have already seen how ABC technniques can provide a means of circumventing this problem—by introducing a second random variable (typically interpreted as an error term) into the data generating process (4.2), the likelihood function can be expressed as a convolution between the true likelihood function—which is unknown—and the density of the error term, which is typically chosen by the statistician carrying out the analysis. The quality of approximation to the true likelihood is controlled via a data mismatch parameter; the smaller this parameter value, the closer the ABC liklihood approximation will be to the true likelihood function. In most ABC applications, the magnitude of the mismatch parameter is chosen to maximise the accuracy of ABC approximation, while still maintaining a minimum level of computational efficiency. In section 1.1 we gave some examples of very basic, rejection-based ABC algorithms that sampled from the ABC approximation to the posterior distribution (the ABC posterior approximation being proportional to the ABC likelihood approximation multiplied by the prior distribution), but pointed out that simple rejection-based algorithms are not suitable in the vast majority of problems due to the prior distribution

being significantly different from the posterior distribution which leads to very inefficient rejection-sampling approaches. In section 1.2 we introduced three Monte Carlo based techniques that can be used to sample from complex distributions when simple rejection-based approaches were insufficient—MCMC, SIS and SMC; despite these techniques mitigating the problems associated with diffuse sampling distributions relative to target distributions, all three methods require the target distribution to be evaluated pointwise. In the case of Bayesian inference, the target distribution is the posterior distribution of parameters which, as already pointed out, is not available—even to evaluate pointwise—in most cases. In this section we will review the ways in which ABC techniques introduced in section 1 have been combined with the Monte Carlo sampling techniques outlined in section 2; we will then go on and present the new variants of these algorithms that we have developed for the purposes of our applications.

4.3.1 ABC MCMC

MCMC is a standard approach to tackling Bayesian problems, especially in situations where there is a significant mismatch between the prior distribution and the posterior. In the most difficult problems, when the likelihood function is not able to be evaluated at all, combining the ideas relating to ABC with the MCMC algorithm can yield practical solutions. Marjoram et al. (2003) developed the first example of an ABC MCMC algorithm in the context of population genetics. In order to appreciate the methodology introduced in Marjoram et al. (2003), first note that when the target distribution in the MCMC algorithm is a posterior density, the Metropolis-Hastings acceptance probability is given by

$$\alpha(\theta_i, \theta') = \frac{\pi(\theta')f(D|\theta')q(\theta_i|\theta')}{\pi(\theta_i)f(D|\theta_i)q(\theta'|\theta_i)}, \tag{4.19}$$

where, as before, θ_i and θ' represent the point at stage i and the proposed point, respectively, of the Markov chain, D represents model observations, f represents the likelihood function, π is the prior distribution and q is the Markov proposal density. In situations where we cannot evaluate the likelihood, the likelihood ratio that appears in (4.19) cannot be evaluated. Marjoram et al. (2003) propose to approximate the likelihood ratio in (4.19) by evaluating Monte Carlo estimates of the numerator and denominator separately. The following estimate is used to approximate the likelihood

$$\hat{f}_{ABC}(D|\theta) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(X_i = D), \quad X_i \sim f(\cdot|\theta), \tag{4.20}$$

where $\mathbb{1}(\cdot)$ represents an indicator function. With this approximation in mind, the prototypical ABC MCMC algorithm is set out below

1. Set $i = 0$. Initialise the chain at a point, θ_i, within the prior's support.
2. Propose a move from the current point in the chain, θ_i to a new point θ' via a proposal density $q(\theta'|\theta_i)$.
3. Generate $X \sim f(\cdot|\theta')$.
4. If $X = D$ go to next step, otherwise discard θ' and return to step 1.
5. Calculate the MH ratio

$$MH = min\left(1, \frac{\pi(\theta')q(\theta_i|\theta')}{\pi(\theta_i)q(\theta'|\theta_i)}\right) \tag{4.21}$$

6. Accept θ' as the next point in the chain, $\theta_{i+1} = \theta'$, with probability MH, otherwise retain θ_i as the next point in the chain, $\theta_{i+1} = \theta_i$, set $i = i + 1$ and return to step 2.

It is a trivial exercise to demonstrate that the Markov chain so constructed has a stationary distribution equal to the posterior of interest (see Marjoram et al. (2003) for more details). Note that step 3 above essentially involves a Monte Carlo estimate of the likelihood of the newly generated point θ', based on one sample ($N = 1$). As was pointed out earlier in this chapter, if the data are defined on a continuous state space then it is necessary to replace the likelihood approximation (4.20) with

$$\hat{f}_{ABC}(D|\theta, \epsilon) \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\rho(D, X_i) \leq \epsilon), \quad X_i \sim f(\cdot|\theta), \tag{4.22}$$

where, as before, ρ is a measure of the similarity between observed and simulated data. In this case, step 3 of the ABC MCMC algorithm above should also be altered to

3. If $\rho(X_i, D) \leq \epsilon$ go to next step, otherwise discard θ' and return to step 1.

Now the stationary, limiting distribution of the Markov chain is equal to the ABC approximated posterior distribution, which is given by

$$\pi_{ABC}(\theta|D, \epsilon) \propto \pi(\theta)\hat{f}_{ABC}(D|\theta, \epsilon), \tag{4.23}$$

where $\hat{f}_{ABC}(D|\theta, \epsilon)$ is the ABC likelihood approximation defined in (4.22). As demonstrated previously, this posterior approximation converges to the true posterior as ϵ tends to zero. As pointed out in Beaumont (2010), a potential drawback of this ABC MCMC algorithm stems from the fact that the likelihood ratio is being crudely estimated by either a 0 or a 1. What this means is that the acceptance rate of the algorithm is proportional to the number of simulated data sets X such that $\rho(D, X) \leq \epsilon$ which is itself proportional to the likelihood, and not the likelihood

ratio, in the limit $\epsilon \to 0$. Thus, the main benefit of using MCMC (namely, that the acceptance rates of the Markov chain are based on likelihood ratios as opposed to the likelihood itself, thereby improving the efficiency of the sampler) is lost. As a result of this shortcoming, the ABC MCMC algorithm described above will tend to mix poorly—once the chain moves into the tails of the posterior distribution, the chain will tend to get stuck in one point for disproportionately long periods of time as proposed local points will rarely produce data that coincides with the observations. This problem becomes even more apparent if the ABC MCMC algorithm is initialised in the tails of the posterior (Beaumont, 2010). There have been several attempts made to avoid this problem of slow mixing; Ratmann et al. (2007) proposed to initialise the ABC MCMC algorithm with a relatively large data mismatch parameter and reduce it gradually during the burn-in stage of the algorithm so that, at convergence, the samples are from within the vicinity of the mode of the posterior (Beaumont, 2010)—a similar approach to this is also taken in certain applications of ABC SMC and ABC SIS algorithms, which will be reviewed later. Alternatively, as per Bortot et al. (2007), one can treat the model mismatch parameter ϵ as an unknown parameter and construct the ABC MCMC algorithm to produce a Markov chain on the joint space $\Theta \times [0, \infty]$ where Θ is the parameter space corresponding to the model parameters and $\epsilon \in [0, \infty]$. If ϵ is small, the samples will represent the true posterior well; if ϵ is large then much of the variability between observed data and model output can be explained by the large error term in (4.2) and thus the samples will generally not represent the true posterior well. If the prior density of ϵ is chosen such that small values of ϵ are generally favoured over large values, then the ABC MCMC algorithm will tend to produce samples broadly representative of the true posterior, while allowing for the occasional large value of ϵ to be simulated, which should have a

beneficial effect on the mixing properties of the ABC MCMC algorithm—especially when the chain strays into regions of low posterior density. As noted in Beaumont (2010), the precise interpretation of the limiting marginal distribution of the data mismatch parameter is unclear; nevertheless, the technique has some utility in that it improves the chain's mixing properties, regardless of how the distribution of the mismatch parameter is interpreted. Having outlined the basic ABC MCMC algorithm, what now follows is an explanation of the new ABC MCMC algorithm developed for the purposes of the applications considered in the next chapter. First, note that the distribution that we wish to obtain samples from is given by

$$\pi_{ABC}(\theta|D,\epsilon) \propto \pi(\theta) \int\limits_{X\sim f(\cdot|\theta)} \pi_{\varepsilon}(D-X|\epsilon) f(X|\theta)\mathrm{d}X, \tag{4.24}$$

where, as before, D is the observed data, θ represents the model parameters and X is a trace from the model run at parameter value θ. This is simply the prior distribution of parameters multiplied by the ABC approximate likelihood (the integral term in the RHS), i.e. the ABC approximated posterior density of model parameters. The error distribution $\pi_{\varepsilon}(\cdot|\epsilon)$ that determines the degree of agreement between traces generated by the model, $f(\cdot|\theta)$, and observed data, D, will be called the similarity kernel. We use a similarity kernel with Gaussian distribution, i.e.

$$\pi_{\varepsilon}(\delta|\epsilon) = \frac{1}{\sqrt{2\pi\epsilon^2}} exp\left\{\frac{-1}{2\epsilon^2}\delta^2\right\}.$$

As pointed out earlier, the distribution of the similarity kernel corresponds to the distributional assumption made with respect to the model or measurement error assumed present in the observations D. The choice of kernel distribution is essentially arbitrary in this application as we are using the ABC assumption (7) as a practical tool to aid in the approximation of intractible likelihood functions, as

opposed to using it to represent any concrete views about the distribution of errors assumed present in the data; however, if one assumes that the error term in (4.2) represents the sum of various errors associated with the data collection process and/or the model's imperfect fit, then the aggregate error can be approximated by a normal distribution, as justified by the central limit theorem. In order to incorporate the ABC approximations into the MCMC framework, note that the ABC approximation of the likelihood given by (7) can be approximated by a Monte Carlo estimate,

$$\int_X \pi_\varepsilon(D - X|\epsilon) f(X|\theta) \mathrm{d}X \approx \frac{1}{M} \sum_{i=1}^{M} \pi_\varepsilon(D - X_i|\epsilon), \quad \forall i,\, X_i \sim f(\cdot|\theta),$$

and label this Monte Carlo estimate $\hat{f}_{ABC}(D|\theta, \epsilon)$ as before. This approximation is analogous to the likelihood approximation used in (4.20), where a uniform distribution is used for the similarity kernel. Now define the MH acceptance probability (associated with moving from parameter value θ to θ') as[14]

$$\alpha_{ABC}(\theta, \theta') = min\left(1, \frac{\hat{f}_{ABC}(D|\theta')\pi(\theta')q(\theta|\theta')}{\hat{f}_{ABC}(D|\theta)\pi(\theta)q(\theta'|\theta)}\right), \tag{4.25}$$

i.e. we replace the likelihoods that appear in the numerator and denominator of the general MH ratio by Monte Carlo estimates. This is similar to the approach used in Marjoram et al. (2003), except we are now using a Gaussian kernel in place of an indicator function and we are estimating the likelihood function using $M \geq 1$ traces. Using more than one trace in the Monte Carlo estimate of the likelihood is beneficial because it offers a refinement to the crude 0–1 cutoff used in Marjoram

[14]Note that the MH ratio is also a function of the M simulated traces from the model; we omit this dependency in the MH acceptance probability for clarity.

et al. (2003) which should improve the mixing of the simulated Markov chain. The adapted ABC MCMC algorithm is as follows

1. Set $i = 0$. Initialise the chain at some point, θ_i, within the support of the prior distribution, π

2. Propose a move from the current point in the chain, θ_i to a new point θ' via a proposal density $q(\theta'|\theta_i)$.

3. Evaluate $\hat{f}_{ABC}(D|\theta', \epsilon)$ and calculate $\alpha_{ABC}(\theta_i, \theta')$ using (4.25).

4. Set $\theta_{i+1} = \theta'$ with probability α_{ABC}, otherwise set $\theta_{i+1} = \theta_i$. Return to step 2.

If the proposed point is accepted then the likelihood estimate, calculated in step 2 above, that is used in the numerator of the MH ratio should be stored for use in the denominator in the next iteration of the algorithm so as to reduce the number of computations that must be carried out on each loop. The Markov chain constructed by this algorithm is on $(\theta, \underline{X})$ where $\underline{X}$ represents the M traces from the model that are used in the Monte Carlo estimate of the ABC likelihood function $\hat{f}_{ABC}(D|\theta, \epsilon)$. The following theorem proves that the stationary marginal distribution of the Markov chain constructed by the adapted ABC MCMC algorithm is the ABC approximated posterior (4.24).

Theorem 2. *The adapted ABC MCMC algorithm produces a Markov chain with stationary limiting marginal distribution* $\pi_{ABC}(\theta|D, \epsilon)$.

Proof To prove that the adapted ABC MCMC algorithm produces a Markov chain with the required stationary distribution, it is required to show that the detailed balance equation (4.8) is satisfied. The joint target distribution is

$$\pi_T(\theta, \underline{X}|D, \epsilon) = \frac{\hat{f}_{ABC}(D|\theta, \epsilon) \prod_{i=1}^{M} f(X_i|\theta)\pi(\theta)}{\pi(D)},$$

and the transition probability of the chain is

$$p(\theta', \underline{X}'|\theta, \underline{X}) = q(\theta'|\theta) \prod_{i=1}^{M} f(X_i'|\theta')\alpha_{ABC}(\theta, \theta').$$

Therefore,

$$\begin{aligned}
\pi_T(\theta, \underline{X}|D,\epsilon)p(\theta', \underline{X}'|\theta, \underline{X}) &= q(\theta'|\theta)\prod_{i=1}^{M} f(X_i'|\theta')\frac{\hat{f}_{ABC}(D|\theta,\epsilon)\prod_{i=1}^{M} f(X_i|\theta)\pi(\theta)}{\pi(D)}\times \\
&\qquad \alpha_{ABC}(\theta,\theta') \\
&= q(\theta'|\theta)\prod_{i=1}^{M} f(X_i'|\theta')\frac{\hat{f}_{ABC}(D|\theta,\epsilon)\prod_{i=1}^{M} f(X_i|\theta)\pi(\theta)}{\pi(D)}\times \\
&\qquad min\left(1, \frac{\hat{f}_{ABC}(D|\theta',\epsilon)\pi(\theta')q(\theta|\theta')}{\hat{f}_{ABC}(D|\theta,\epsilon)\pi(\theta)q(\theta'|\theta)}\right) \\
&= q(\theta|\theta')\prod_{i=1}^{M} f(X_i|\theta)\frac{\hat{f}_{ABC}(D|\theta',\epsilon)\prod_{i=1}^{M} f(X_i'|\theta')\pi(\theta')}{\pi(D)}\times \\
&\qquad min\left(1, \frac{\hat{f}(D|\theta,\epsilon)\pi(\theta)q(\theta'|\theta)}{\hat{f}_{ABC}(D|\theta',\epsilon)\pi(\theta')q(\theta|\theta')}\right) \\
&= q(\theta|\theta')\prod_{i=1}^{M} f(X_i|\theta)\frac{\hat{f}_{ABC}(D|\theta',\epsilon)\prod_{i=1}^{M} f(X_i|\theta')\pi(\theta')}{\pi(D)}\times \\
&\qquad \alpha_{ABC}(\theta',\theta) \\
&= \pi_T(\theta', \underline{X}'|D,\epsilon)p(\theta, \underline{X}|\theta', \underline{X}').
\end{aligned}$$

This proves that the adapted ABC MCMC algorithm converges to the target distribution π_T. It is now straightforward to show that the marginal distribution of the target distribution is the ABC approximated posterior, i.e.

$$\begin{aligned}
\int_{\mathcal{D}(\underline{X})} \pi_T(\theta, \underline{X}|D,\epsilon)\mathrm{d}\underline{X} &= \frac{1}{M}\frac{\pi(\theta)}{\pi(D)}\int_{X_{1:M}} \sum_{i=1}^{M} \pi_\varepsilon(D - X_i|\epsilon)\prod_{i=1}^{M} f(X_i|D)\mathrm{d}X_{1:M} \\
&= \frac{\pi(\theta)}{\pi(D)}\int_{X} \pi_\varepsilon(D - X|\epsilon)f(X|\theta)\mathrm{d}X \\
&= \pi_{ABC}(\theta|D,\epsilon). \qquad \square
\end{aligned}$$

4.3.2 ABC SMC

Sisson et al. (2007) attempted to incorporate the ABC methodology into the generic SMC algorithm developed by Del Moral et al. (2006), which was outlined earlier in the section titled 'Sequential Monte Carlo sampling'. In this section Sisson's original algorithm will be outlined, and a variant of the ABC SMC (which will be labelled *Tempered ABC SMC*) algorithm developed with the applications considered in this thesis in mind will also be introduced. Unfortunately, Sisson *et al.*'s algorithm (and, consequently, Tempered ABC SMC) produces sample particles that are systematically biased, owing to the fundamental incompatibility between ABC and Del Moral et al. (2006)'s general SMC sampler. The reasons for this bias in the particles sampled from the ABC SMC algorithm, and the ways in which this problem has been tackled will both be discussed in this section.

Sisson *et al.*'s ABC SMC

As pointed out earlier, ABC MCMC provides a notable improvement in efficiency relative to straightforward rejection sampling because the sampling density becomes progressively more representative of the posterior density (the target density in Bayesian applications) in ABC MCMC, and therefore mitigates the problem of a diffuse sampling density relative to the target density that one suffers in rejection sampling. ABC MCMC does have its own problems—recall that Marjoram *et al.*'s acceptance probability was proportional to the likelihood (and not the likelihood ratio), which often results in extremely low acceptance rates when the Markov chain enters regions of low posterior mass. Bortot *et al.*'s solution of builing a Markov chain with auxiliary variables (namely, the similarity kernel variance) can alleviate this problem at the expense of reducing computational efficiency; only sampled points with a small similarity kernel variance can be reasonably expected

to represent the posterior density, therefore only the subset of sampled points with low kernel variance parameter are used for inference, with the remaining points discarded. Sisson et al. (2007) motivate the development of ABC SMC by noting several advantages associated with SMC over MCMC and rejection sampling, for example

1. Inefficiencies resulting from a mismatch between sampling and target densities are avoided, as in ABC MCMC.

2. Sampled points that poorly represent the posterior are discarded in favour of points that are more representative of the posterior.

3. SMC methods are better suited to sampling from complex, multi-modal posterior distributions than MCMC.

4. Particles are uncorrelated, and SMC-based methods do not require an assessment of burn-in or convergence, unlike MCMC-based methods.

The aim of ABC SMC is to obtain N particles $\{\theta^i\}_{i=1,\ldots,N}$ whos empirical distribution converges to the ABC posterior $(\pi_{ABC}(\theta|\rho(S(D), S(D)) \leq \epsilon))$as $N \to \infty$. In what follows, it is assumed that (possibly a vector of) summary statistics $S(\cdot)$that capture much of the information contained within the data are available to the statistician. In practice, the choice of such summary statistics is a difficult problem, and one that will be discussed in detail in the next chapter (section 2.3) concerning numerical experiments. Having already set out the main structure and ideas behind SMC in a previous section, all that is required to illustrate Sisson *et al.*'s algorithm is to introduce the specific choices associated with the algorithm; namely, the sequence of target distributions that the particle population will approximate and

the particular form of ABC likelihood approximation. Recall that in SMC there were a sequence of target distributions represented by

$$\pi_n(\theta) = \frac{\gamma_n(\theta)}{Z_n}, \quad n = 1, \ldots, T,$$

where Z_n is the normalising constant of the target distribution. In Bayesian applications, the target density is invariably the posterior density associated with the model parameters, which is proportional to the product of the prior distribution and the likelihood associated with the observed data. Sisson *et al.* define a sequence of target distributions as follows:

$$\pi_n(\theta|\rho(S(D), S(X)) \leq \epsilon_n) = \frac{\pi(\theta)}{B_n} \sum_{i=1}^{B_n} \mathbb{1}(\rho(S(D), S(X_i)) \leq \epsilon_n), \quad n = 1, \ldots, T, \tag{4.26}$$

where the X_i are simulated traces from the model conditioned on the parameter value θ. Note that the expression for the sequence of target distributions involves the product of the prior distribution with the ABC likelihood approximation, in which the error distribution associated with ABC (see (4.2)) is uniform, and therefore the sequence of target distributions considered here is effectively a sequence of approximations to the unnormalised posterior density of model parameters θ. ϵ_n is a monotonically decreasing sequence of error tolerances—that is, the sequence of posterior approximations becomes progressively more accurate, with ϵ_T chosen such that the end product is a particle population that approximates the true posterior up to some predetermined level of accuracy. By defining the sequence of distributions in such a way, the particle population can be smoothly moved between distributions in the sequence, resulting in a population that approximates the final target distribution in the sequence. Note that the likelihood approximation in (4.26) is of the same form used by Marjoram *et al.* in the ABC MCMC algorithm,

except the error tolerance ϵ_t and number of Monte Carlo terms B_t are allowed to vary along the sequence of target distributions. The ABC SMC sampler is as follows:

1. Initialisation

   - Specify a sequence of monotonically decreasing error tolerances $\epsilon_1, \epsilon_2, \ldots, \epsilon_T$.
   - Set population indicator $n = 1$, and particle indicator $i = 1$.

2. Particle sampling

   - If $n = 1$, sample $\theta' \sim \mu_1$, where μ_1 is the initial sampling density[15]. If $n > 1$, sample θ'' from the previous population $\{\theta_{n-1}^{(i)}\}$ with weights $\{W_{n-1}^{(i)}\}$, and perturb the particle to $\theta' \sim K_n(\theta|\theta'')$ according to a Markov transition kernel K_n. Generate a data set $X' \sim f(\cdot|\theta')$. If $\rho(S(D), S(X')) > \epsilon_n$, go to step 3.
   - Set
   $$\theta_n^{(i)} = \theta', \quad W_t^{(i)} = \begin{cases} \pi(\theta_n^{(i)})/\mu_1(\theta_n^{(i)}) & \text{if } n = 1, \\ \frac{\pi(\theta_n^{(i)}) L_{t-1}(\theta'|\theta_n^{(i)})}{\pi(\theta') K_t(\theta_n^{(i)}|\theta')} & \text{if } t > 1, \end{cases}$$
   where L_{t-1} is the backwards Markov kernel introduced earlier. If $i < N$, set $i = i + 1$ and go to step 3.

3. Resampling

   - Normalise the weights so that $\sum_{i=1}^{N} W_n^{(i)} = 1$.
   - If $ESS < E$, resample with replacement the particles in proportion to the weights $\{W_n^{(i)}\}$ and reset weights $\{W_n^{(i)} = 1/N\}$, where ESS is the effective sample size defined earlier, and E is some threshold value, typically set to half the population size, $N/2$.

[15] In practice, this density is often just the prior distribution $\pi(\theta)$.

- If $n < T$, set $n = n + 1$ and return to step 2.

Sisson et al. (2007) choose the backward Markov kernel L_{n-1} to equal the forward kernel K_n which simplifies the expression for the particle weights in step 4 above, which results in particle weights being equal, assuming a uniform prior is chosen (Beaumont et al., 2009). As demonstrated by Beaumont et al. (2009), the ABC SMC algorithm illustrated above produces biased samples from the target distributions; in particular, the empirical distribution of sampled particles tends to under-represent the tails of the target distribution. Ultimately, this bias results from the fact that the incremental weight formula (4.17) of generic SMC sampler developed by Del Moral et al. (2006) features the target density in the denominator, which is missing in the ABC SMC algorithm—the accept-reject step that is common to all ABC algorithms allows for the posterior in the numerator of the weight function to be replaced by the prior but, unfortunately, not in the denominator. For a theoretical demonstration of the biasedness of ABC SMC, see Beaumont et al. (2009).

Tempered ABC SMC

We developed an alternative ABC SMC sampler for use in this thesis that differs in several ways from the ABC SMC algorithm developed by Sisson et al. (2007):

1. The sequence of target distributions is given by

$$\pi_n(\theta) = \pi(\theta)^{1-\phi_n} \left[\pi_{ABC}(\theta|D, \epsilon)\right]^{\phi_n}, \quad n = 0, \ldots, T, \tag{4.27}$$

   where $\pi_{ABC}(\theta|D, \epsilon)$ is the ABC approximated posterior distribution and is given by (4.24), and

$$0 = \phi_0 < \phi_1 < \cdots < \phi_T = 1.$$

By substituting (4.24) into (4.27), the sequence of target distributions can also be represented by

$$\pi_n(\theta) = \pi(\theta)\left[\int_{X\sim f(\cdot|\theta)} \pi_\varepsilon(D - X|\epsilon) f(X|\theta)\mathrm{d}X\right]^{\phi_n}, \quad n = 0, \ldots, T. \quad (4.28)$$

2. The incremental weight function, the analogue of the weight formula given in step 2 of Sisson *et al.*'s ABC SMC, is given by

$$W_n^{(i)} = \hat{f}_{ABC}(D|\theta_{n-1}, \epsilon)^{\phi_n - \phi_{n-1}}, \quad \hat{f}_{ABC}(D|\theta_{n-1}, \epsilon) = \frac{1}{M}\sum_{i=1}^{M} \pi_\varepsilon(D - X|\epsilon), \quad (4.29)$$

where $X \sim f(\cdot|\theta_{n-1})$.

The sequence of distribution used in Tempered ABC SMC is inspired by the path sampling techniques developed in (Neal, 2001) and (Gelman and Meng, 1999). Rather than design a sequence of error tolerances to define the sequence of distributions, as in Sisson et al. (2007), a sequence of tempering parameters $\{\phi_n\}_{n=1,\ldots,T}$ are used to generate a sequence of distributions that move gradually from the—assumed easy to sample from—prior distribution to the complex, ABC approximated posterior distribution. Both the sequence of importance distributions defined by the decreasing error tolerance and the sequence defined by the tempering parameter ϕ_n are valid choices in the context of ABC SMC (and ABC SIS); however, the latter choice provides an elegant and easily interpretable sequence of importance distributions that moves smoothly between the prior and the intractible ABC posterior. The form of the incremental weight function follows directly from the choice of forward and backward Markov kernels used in Tempered ABC SMC; as outlined in Del Moral et al. (2006), there are a variety of options available for

choosing the forward kernels, one particular option being a MCMC kernel (see the earlier section on SMC methods for a brief description of this kernel choice). As per Del Moral et al. (2006), the suggested choice of backwards kernel associated with a MCMC forward Markov kernel results in the following expression for the incremental weight function

$$W_n^{(i)} = \frac{\pi_n(\theta_{n-1})}{\pi_{n-1}(\theta_{n-1})},$$

which, after substituting in (4.28) to the RHS of the above equation, reduces to the ABC likelihood approximation raised to the power $\phi_n - \phi_{n-1}$. In practice the ABC likelihood approximation cannot be evaluated analytically, therefore a Monte Carlo approximation is used in place of the analytic expression, which leads to the incremental weight function defined in (4.29). The results of the Tempered ABC SMC algorithm's application to some financial models will be presented in the next chapter; however, as noted previously, Tempered ABC SMC actually produces biased samples that typically fail to adequately represent the tails of the target posterior distribution of interest. This bias shares its origins with the bias present in the ABC SMC algorithm develioped by Sisson et al. (2007), which was theoretically and empirically demonstrated in Beaumont et al. (2009). What follows is a theoretical demonstration of the bias exhibited by one step of the Tempered ABC SMC algorithm. Following the approach of Beaumont et al. (2009), it is assumed that

1. $\epsilon = 0$. That is, the similarity kernel variance—the parameter in the error distribution π_ε—is zero.

2. The number of terms utilised in the Monte Carlo approximation of the ABC likelihood tends to infinity, i.e. $M \to \infty$, which, when taken with assumption 1, implies that

$$\hat{f}_{ABC}(D|\theta, \epsilon) = f(D|\theta),$$

that is, the Monte Carlo estimate of the ABC likelihood approximation equals the true likelihood function associated with the model.

3. The previous particle population is an exact sample from the target distribution at stage $n-1$, i.e.

$$\theta_{n-1} \sim \pi(\theta) f(D|\theta)^{\phi_{n-1}}.$$

By making these assumptions, the bias associated with the Tempered ABC SMC algorithm can be clearly demonstrated by considering one step of the algorithm—in particular, the expectation of some arbitrary integrable function h will be considered. Given the aforementioned assumptions, the joint density of the accepted pair of particles (θ_{n-1}, θ_n) is proportional to $\pi(\theta_{n-1}) f(D|\theta_{n-1})^{\phi_{n-1}} K_n(\theta_n|\theta_{n-1})$, therefore

$$\begin{aligned}
\mathbb{E}(h(\theta_n)W_n) &= \int\int h(\theta_n) f(D|\theta_{n-1})^{\phi_n - \phi_{n-1}} \pi(\theta_{n-1}) f(D|\theta_{n-1})^{\phi_{n-1}} K_n(\theta_n|\theta_{n-1}) \mathrm{d}\theta_{n-1} \mathrm{d}\theta_n \\
&= \int h(\theta_n) \left\{ \int \pi(\theta_{n-1}) f(D|\theta_{n-1})^{\phi_n} K_n(\theta_n|\theta_{n-1}) \mathrm{d}\theta_{n-1} \right\} \mathrm{d}\theta_n. \qquad (4.30)
\end{aligned}$$

In order for this step of the Tempered ABC SMC algorithm to yield unbiased results, the integral in parenthesis above must be proportional to the target distribution at stage n, i.e. $\pi(\theta_n) f(D|\theta_n)^{\phi_n}$, which is generally not the case. This demonstrates that the Tempered ABC SMC algorithm produces biased samples. In order to correct for this bias, the Tempered ABC SMC algorithm presented in this section will be altered in a fashion similar to the steps taken by Sisson *et al.* to correct

their ABC SMC algorithm. The corrections lead to an algorithm that is more closely aligned with standard importance sampling ideas—this corrected algorithm will be covered in section 1.3.3.

4.3.3 ABC SIS

To recap, Del Moral *et al.*'s paper (Del Moral et al., 2006) presents two generic sampling algorithms that provide the theoretical impetus for the development of two ABC based samplers: ABC SMC and ABC SIS. As pointed out in the previous section, samplers based on the generic SMC algorithm in Del Moral et al. (2006) produce biased samples when the likelihood function cannot be evaluated analytically, and are therefore not suitable for use in the ABC setting. This leaves ABC SIS as the main alternative to ABC MCMC algorithms for sampling from intractible posterior distributions. ABC SIS algorithms have been separately developed by Sisson *et al.* (Sisson et al., 2007)[16], by Toni *et al.* (Toni et al., 2009) and by Beaumont *et al.* (Beaumont et al., 2009), but all three versions of the algorithm are broadly the same. Toni *et al.*'s ABC SIS algorithm will be presented first, before introducing the new ABC SIS algorithm developed for the applications considered in this thesis.

Toni *et al.*'s ABC SIS

Recall that SIS involves selecting a sequence of intermediate target distributions, $\{\pi_n\}$, $n = 1, \ldots, T-1$, and sampling distributions $\{\eta_n\}$, $n = 1, \ldots, T-1$, and carrying out importance sampling sequentially to evolve a population of particles through the sequence of target distributions, resulting in a sample from the final target distribution π_T which, in this case, is the ABC approximated posterior

[16] As a result of the corrections to their original ABC SMC algorithm to remove the bias.

density of model parameters. The ABC SIS procedure in (Toni et al., 2009) is defined by specifying the sequence of target and sampling distributions used in the SIS setup. The target distributions are defined by

$$\pi_n(\theta) = \frac{\pi(\theta)}{B_n} \sum_{i=1}^{B_n} \mathbb{1}(\rho(D, X_i) \leq \epsilon_n),$$

where, $\pi(\cdot)$ denotes the prior density, X_i are data sets generated from the model using parameter value (or particle value, in the SIS terminology) θ, and B_n is the number of data sets utilised for the Monte Carlo approximation of the ABC likelihood. Note that the distributions making up the sequence targets are proportional to the ABC approximated posterior distribution, with varying magnitudes of error tolerance. Define $b_n = \sum_{i=1}^{B_n} \mathbb{1}(\rho(D, X_i) \leq \epsilon_n)$. The sampling distributions are defined by

$$\eta_n(\theta) = \mathbb{1}(\pi(\theta) > 0)\mathbb{1}(b_n > 0) \int \pi_{n-1}(\theta_{n-1}) K_n(\theta|\theta_{n-1}) \mathrm{d}\theta_{n-1}.$$

The indicator functions in the above definition are included to ensure that the sampling densities and target densities are equivalent (i.e. $\pi_n(\theta) > 0 \iff \eta_n(\theta) > 0$), which in turn ensures that the importance weights are well defined. The ABC SIS algorithm is as follows:

1. Initialise the sequence of error tolerances $\epsilon_1, \ldots, \epsilon_T$. Set population indicator $n = 0$.

2. Set particle indicator $i = 1$.

3. If $n = 0$, sample θ' independently from the prior, π. If $n > 0$, sample θ'' from the previous population $\left\{\theta_{n-1}^{(i)}\right\}$ in proportion to the particle weights

w_{n-1} and perturb the particle to obtain $\theta' \sim K_n(\theta|\theta'')$, where K_n is a Markov kernel.

4. If $\pi(\theta') = 0$, return to step 3, otherwise simulate B_n datasets $X_i \sim f(\cdot|\theta')$, $i = 1, \ldots, B_n$ and calculate $b_n(\theta')$.

5. If $b_n(\theta') = 0$, return to step 3, otherwise set $\theta_n^{(i)} = \theta'$ and calculate the importance weight

$$w_n^{(i)} = \begin{cases} b_n(\theta_n^{(i)}) & \text{if } n = 0, \\ \frac{\pi(\theta_n^{(i)}) b_n(\theta_n^{(i)})}{\sum_{j=1}^{N} w_{n-1}^{(j)} K_n(\theta_n^{(i)}|\theta_{n-1}^{(j)})} & \text{if } n > 0. \end{cases}$$

   If $i < N$, set $i = i + 1$ and go to step 3.

6. Normalise the weights. If $n < T$, set $n = n + 1$ and go to step 2.

Note that in step 5 above, a Monte Carlo estimate of the sampling density has been used in the denominator of the importance weight calculation. Beaumont et al. (2009) demonstrate that this ABC SIS algorithm generates sample populations that yield unbiased expectations with respect to the target distributions.

Tempered ABC SIS

The Tempered ABC SIS algorithm that we suggest in this thesis differs from existing ABC SIS algorithms in two ways: firstly, the sequence of importance distributions that are targeted is given by (4.28)—the same sequence of distributions used in the Tempered ABC SMC algorithm; secondly, the ABC likelihood approximation utilises a Gaussian distribution as opposed to the uniform distribution chosen by Toni *et al.* and the other groups involved in developing ABC SIS algorithms. To reiterate, choosing a Gaussian distribution for the similarity kernel is equivalent

to (see Wilkinson (2013)) assuming the distribution of errors assumed present in the observations is normally distributed, which, we argue, is a more realistic assumption[17] if it is assumed that the errors consist of additive errors from various sources (e.g. different types of measurement and model error). In addition to being justifiable from a theoretical perspective, the use of a Gaussian similarity kernel results in an ABC SIS algorithm that does not feature an accept-reject step, unlike in the cases where a uniform error distribution is used, which should result in some improvement in computational efficiency. In order to implement Tempered ABC SIS, replace steps 1, 4, and 5 in Toni *et al.*'s algorithm with the following:

1. Initialise the sequence of tempering parameters ϕ_n, which define the sequence of importance distributions. Set population indicator $n = 0$.

4. If $\pi(\theta') = 0$, return to step 3, otherwise simulate M datasets $X_i \sim f(\cdot|\theta')$, $i = 1, \ldots, M$ and calculate $b_n(\theta')$. In Tempered ABC SIS, $b_n(\theta')$ is defined as follows:
$$b_n(\theta') = \sum_{i=1}^{M} \pi_\varepsilon(D - X_i)|\epsilon), \quad \pi_\varepsilon(\cdot|\epsilon) \sim \mathcal{N}(0, \epsilon^2).$$

5. Set $\theta_n^{(i)} = \theta'$ and calculate the importance weight
$$w_n^{(i)} = \begin{cases} 1 & \text{if } n = 0, \\ \dfrac{\pi(\theta_n^{(i)})\left(b_n(\theta_n^{(i)})\right)^{\phi_n}}{\sum_{j=1}^{N} w_{n-1}^{(j)} K_n(\theta_n^{(i)}|\theta_{n-1}^{(j)})} & \text{if } n > 0. \end{cases}$$

The unbiasedness of this ABC SIS algorithm can be easily demonstrated by first noting that the distribtion of the particle at time n in the algorithm is given by $\hat{\pi}_n(\theta_n) \propto \sum_{i=1}^{N} w_{n-1}^{(i)} K_n(\theta_n|\theta_{n-1}^{(i)})$. All that remains is to follow the steps taken earlier

[17] By virtue of the central limit theorem.

in the chapter when we demonstrated the biasedness of the Tempered ABC SMC sampling algorithm (see equation (4.30)):

$$\begin{aligned}\mathbb{E}(w_n h(\theta_n)) &\propto \int\int h(\theta_n)\frac{\pi(\theta_n)\left(b_n(\theta_n)\right)^{\phi_n}}{\hat{\pi}_n(\theta_n)}\tilde{\pi}(\theta_{n-1})\hat{\pi}_n(\theta_n)\mathrm{d}\theta_{n-1}\mathrm{d}\theta_n \\ &= \int h(\theta_n)\pi(\theta_n)\left(b_n(\theta_n)\right)^{\phi_n}\left\{\int \tilde{\pi}(\theta_{n-1})\mathrm{d}\theta_{n-1}\right\}\mathrm{d}\theta_n,\end{aligned}$$

which does not depend on the distribution of the previously sampled points, $\tilde{\pi}(\theta_{n-1})$. This expression is proportional to the expectation of h taken with respect to the target distribution at time n, hence the Tempered ABC SIS algorithm does indeed produce unbiased samples from the sequence of importance distributions.

4.4 Discussion & summary

In this chapter we introduced the ideas associated with Approximate Bayesian Computation (ABC) and its use in parameter estimation. Firstly, the motivations behind the development of these techniques, namely the need to solve the 'reverse engineering problem' of inferring parameters from model generated observations which rapidly becomes a challenging task as the model complexity increases, were introduced, followed by a brief historical overview of the development of these ideas. Some basic results validating the basic ABC framework, i.e. approximating the true likelihood with a convolution of the likelihood function with some similarity kernel π_ε and estimating this quantity using Monte Carlo estimation, were then presented. After highlighting the limitations of basic rejection-based ABC samplers, in particular, the inefficiency of such samplers when the target distribtion differs greatly from the sampling distribution, a variety of Monte Carlo sampling methods were introduced, which are capable of sampling from complex target distributions, in each case covering the essential theory needed to understand how these procedures

produce the desired empirical samples and the practical considerations that must be taken into account when implementing the method. Lastly, a survey of Monte Carlo based ABC samplers relevant to the work in this thesis was then presented, including the ABC MCMC sampling procedure developed by Marjoram et al. (2003), the ABC SMC sampler developed by Sisson et al. (2007) and the ABC SIS sampler introduced by Toni et al. (2009). We then introduce some new importance samplers that we developed, labelled Tempered ABC SIS and Tempered ABC SMC, and a new variation of the ABC MCMC sampler introduced by Marjoram et al. (2003) that we also developed that utilises multiple sample paths and a Gaussian similarity kernel to approximate the model likelihood.

4.5 Appendix

In this appendix we provide a proof of the claim that the probabilistic approximate rejection algorithm presented in Wilkinson (2013) produces samples from the posterior distribution, assuming (4.2) holds true. This proof originally appeared in Wilkinson (2013). The proof involves demonstrating that the distribution of accepted parameter values from the algorithm is equal to the posterior distribution of parameters under assumption (4.2). Let

$$I = \begin{cases} 1 & \text{if } \theta \text{ is accepted} \\ 0 & \text{otherwise.} \end{cases}$$

Then we have that

$$\begin{aligned} Pr(I = 1|\theta) &= \int Pr(I = 1|\mathcal{M}(\theta) = X, \theta) f(X|\theta) \mathrm{d}X \\ &= \int \pi_\epsilon(D - X) f(X|\theta) \mathrm{d}X. \end{aligned}$$

Therefore, by an application of Bayes' theorem, the distribution of accepted values is given by

$$\pi(\theta|I=1) = \frac{\pi(\theta)\int \pi_\varepsilon(D-X|\epsilon)f(X|\theta)\mathrm{d}X}{\int \pi(\theta')\int \pi_\varepsilon(D-X|\epsilon)f(X|\theta')\mathrm{d}X\mathrm{d}\theta'}. \tag{4.31}$$

To complete the proof we must demonstrate that the posterior distribution of parameters, under assumption (4.2), is equal to (4.31). From (7) we know that the likelihood of the observations, under assumption (4.2), is

$$f_{ABC}(D|\theta,\epsilon) = \int \pi_\varepsilon(D-X|\epsilon)f(X|\theta)\mathrm{d}X, \quad where$$
$$X \sim f(\cdot|\theta)$$

therefore, the posterior distribution of parameters is given by

$$\pi(\theta|D) = \frac{\pi(\theta)\int \pi_\varepsilon(D-X|\epsilon)f(X|\theta)\mathrm{d}X}{\int \pi(\theta')\int \pi_\varepsilon(D-X|\epsilon)f(X|\theta')\mathrm{d}X\mathrm{d}\theta'},$$

which is equal to (4.31). □

Chapter 5

ABC-based Parameter Estimation: A Simulation Study

5.1 Introduction

In the previous chapter, the theory that underpins the ABC approach to parameter estimation was presented, along with several examples of Monte Carlo based ABC (MC ABC) samplers in the existing literature. We then proposed some new MC ABC samplers, namely Tempered ABC SIS, Tempered ABC SMC, and the adapted ABC MCMC sampler. In this chapter, these new samplers[1] will be tested against some standard models that are widely used within the field of mathematical finance, in order to assess their efficacy. In what follows we will provide an overview of the models that we have chosen for the experiments; we will then discuss the need to reduce the dimensionality of the data observations via the employment of summary statistics, followed by a survey of the various methods of choosing suitable summary statistics. We then present the specific details of the experiment set up, followed

[1]The samplers that do not produce biased samples, namely Tempered ABC SIS and adapted ABC MCMC.

by the results obtained from the analyses. The chapter concludes with a discussion of the results, and a summary of the salient points raised herein.

5.2 Overview of models studied

In order to assess the effectiveness of the ABC estimation methodology, we will apply the samplers introduced in the last chapter to two well known models, widely used in mathematical finance: Geometric Brownian Motion (GBM) and the Cox-Ingersoll-Ross (CIR) model. Both models are examples of stochastic differential equations (SDEs); which are the most common means of representing the dynamics of market variables such as share prices and interest rates. In fact, both models are examples of a subclass of SDEs known as (time-homogeneous) Itô diffusions. An n-dimensional time-homogeneous Itô diffusion process defined on the measurable space $(\Omega, \mathcal{F})$, labelled $X_t(\omega) = X(t, \omega) : [0, \infty) \times \Omega \to \mathbb{R}^n$,[2] is a stochastic process generally represented as follows

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}B_t, \quad X(0) = x_0, \tag{5.1}$$

where B_t is an m-dimensional Brownian motion and $a : \mathbb{R}^n \to \mathbb{R}^n$, $b : \mathbb{R}^n \to \mathbb{R}^{n \times m}$ are called the drift and diffusion coefficients respectively, that are assumed to satisfy certain conditions such that (5.1) possesses a unique solution. Models of this sort possess several attractive properties that make them popular within quantitative finance; diffusion processess are continuous time, stochastic processes, and as such, analysts have a rich toolbox of analytic results and methods, available from the general body of knowledge concerning stochastic processes, at their disposal. In

[2]For clarity, we supress the $\omega \in \Omega$ dependence of $X(t, \omega)$ in what follows, i.e. we will write $X(t, \omega) = X(t) = X_t$.

addition, Itô diffusion processes are Markov processes by construction, that is, for some Borel measurable function f,

$$\mathbb{E}^x \left[f(X_{t+h}) | \mathcal{F}_t^m \right] = \mathbb{E}^{X_t} \left[f(X_h) \right],$$

where $\mathbb{E}^x$ denotes the expectation taken w.r.t. the probability law, $\mathbb{Q}^x$, of the process $\{X_t\}_{t \geq 0}$, and $\mathcal{F}_t^m$ is the filtration, or σ-algebra, generated by the m-dimensional Brownian motion $\{B_r; r \leq t\}$. This feature of diffusion processes will be utilised later, when we consider the form of the likelihood associated with each model. Both models considered here, GBM and CIR, are univariate models, that is $n, m = 1$. What follows is a brief summary of the main features of each model.

5.2.1 Geometric Brownian motion

Geometric Brownian motion is arguably the most well-known SDE within the mathematical finance community, mainly down to its central role in the seminal work by Black and Scholes (1973) in deriving the fair price for a derivative contract based on an underlying asset, the dynamics of which are assumed to follow GBM. The model is as follows:

$$\mathrm{d}S_t = \mu S_t \mathrm{d}t + \sigma S_t \mathrm{d}B_t, \quad S_0 = s_0 > 0, \tag{5.2}$$

where $\theta = (\mu, \sigma)$ are the constant model parameters, μ being labelled the 'drift' coefficient and σ the 'diffusion' coefficient. This model has traditionally been used as the standard approach to modelling share prices, given the qualitative properties exhibited by processes with GBM dynamics:

- The process is non-negative, which is obviously a desirable characteristic of a model describing share prices.

- Expected returns from the model are independent of price level.
- The volatility of the process is proportional to the level of the process, which is a quality observed in real markets.

In addition to the attractive qualitative properties mentioned, GBM is one of the few SDEs that can be solved analytically, which makes working with the model relatively easy.

Applying the ABC samplers to the estimation of the drift and diffusion parameters of the GBM model will provide a basic test of the effectiveness of the ABC methodology when applied to the type of models often considered in finance. If the techniques outlined in the previous chapter are to be of any practical use in finance, they must be able to deal with relatively simple models like the GBM model, before being considered for the estimation of far more complex SDEs that are now common place in industry. Aside from this model being well-known, its analytical tractibility means that the transition density, and therefore the likelihood function, is known explicity, making it a good candidate for assessing the extent to which the ABC samplers can reproduce the analytic posterior density that we are trying to sample from. In addition to being able to derive the analytic posterior, which we can use as a yardstick to assess the performance of the samplers, we can also derive the sufficient summary statistics for this model, which will allow us to test the efficacy of the samplers using both sufficient and non-sufficient statistics. A more detailed discussion regarding the choice of summary statistics will be given in a separate section later in the chapter.

5.2.2 The CIR model

This model, first introduced by Cox et al. (1985), has been frequently used to model the dynamics of nominal (instantaneous) short term interest rates. The model is specified as follows:

$$\mathrm{d}R_t = \alpha(\beta - R_t)\mathrm{d}t + \nu R_t^{1/2}\mathrm{d}B_t, \quad R_0 = r_0 > 0, \tag{5.3}$$

where $\theta = (\alpha, \beta, \nu)$ are strictly positive, constant parameters. Notably, the model produces times series that

- Exhibit 'mean reversion'. Mean reversion is an empirical feature of market interest rate data and it relates to the tendency of nominal interest rates to be pulled back towards some long term average level (most commonly referred to as the 'mean reversion level') over time. If the current interest rate is below the mean reversion level, the process tends to exhibit a positive drift; when rates are above the mean reversion level, the process tends to exhibit a negative drift. In addition to this behaviour being observed in the data, there are compelling economic arguments that support the inclusion of mean reverting behaviour in models of nominal interest rates. When interest rates are high borrowing becomes expensive, which leads to a drop in demand for funds for investment; this leads to a fall in interest rates. Conversly, when rates are low, borrowing is relatively cheap which drives up demand for funds for investment; this leads to an increase in interest rates.

- Exhibit positivity. Standard economic theory states that nominal interest rates cannot drop below zero. Some models of short term interest rates (for example, the so-called Vasicek model) are able to produce mean reverting

behaviour, but do not exclude the possibility that the modelled process drops below zero. In the CIR model the inclusion of the square root term in the diffusion coefficient, $\sqrt{R_t}$, prevents the process from dropping below zero. Provided that the constraints on the model parameters that were outlined above hold, as the process approaches zero the diffusion term tends to zero and the deterministic drift component of the model dominates the behaviour of the process, pulling the interest rate up towards the positive mean reversion level. This model ensures that the process does not drop below zero; however, the process can still occasionally hit zero unless the model parameters satisfy a further constraint, commonly referred to as the Feller condition:

$$2\alpha\beta > \nu^2 \tag{5.4}$$

The Feller condition is typically satisfied for parameter values corresponding to realistic market data. In this thesis we will assume the Feller condition holds.

The model is characterised by a non-central chi-squared transition density, i.e. the conditional distribution, $R_t|R_s$, $s \leq t$ is non-central chi-squared. In addition to the qualitative features that the model possesses, the model also admits analytic expressions for the price of bonds and options on bonds, which is a highly desirable feature that financial engineers look for in pricing models.

Although still relatively tractable, estimation of the CIR process represents a more challenging test, relative to the Black-Scholes model of share prices, of the ABC estimation techniques developed in the previous chapter. Firstly, the model has three parameters, unlike the Black-Scholes model that possesses only two parameters; secondly, the dynamics of the process are significantly more difficult

to deal with than the log-normal dynamics of the Black-Scholes model—sufficient summary statistics cannot be derived for this particluar model, and therefore we must rely on the identification of suitable, non-sufficient, summary statistics in order to test the efficacy of the ABC parameter estimation techniques presented earlier. As such, this model estimation exercise is closer to the sort of problem that one might encounter in industry, in which the model is not sufficiently tractable as to allow sufficient summary statistics and analytic likelihood functions to be derived.

5.3 Choosing summary statistics

In the last chapter it was noted that there are certain times when it is necessary to construct summary statistics that effectively reduce the dimensionality of the observations; in this section, we explore this concept in more detail. We will discuss the rationale for using summary statistics, explain the concept of sufficiency, and survey the methods of constructing statistics that have been proposed elsewhere in the literature. We also discuss some alternative summary statistic choices, devised for the models being investigated, that will be tested in the course of conducting the numerical experiments.

5.3.1 The need for summary statistics

Recall that in the prerequisites section of the previous chapter (Section 4.1.1), it was stated that in situations where the model being estimated produces high dimensional data, it is usually necessary to use summary statistics, rather than the full data set, to estimate model parameters in order to avoid inefficiencies in the sampling algorithm. This can be readily seen in the case of a simple rejection-based ABC sampler, e.g. the sampler discussed on page 62 that was proposed by Pritchard

et al. (1999). In this basic rejection ABC sampler, a (potentially vector-valued) parameter is simulated from the prior density, and a data set is then generated from the model using the simulated parameter, the simulated data is then compared with the observed data and the generated parameter is accepted as a sample from the ABC posterior provided the generated data is sufficiently close to the observation data, i.e. provided

$$\rho(D, X_i) \leq \epsilon,$$

where ρ measures the degree of similarity between generated and observed data. In Pritchard et al. (1999) the similarity measure, ρ, is given by

$$\rho(D, X_i) \equiv \max_j \frac{|D_j - X_{i,j}|}{D_j}, \quad \text{where } D = \{D_j\}_{j=1,\ldots,n}, X_j = \{X_{i,j}\}_{j=1,\ldots,n}.$$

If the dimension of the data, n, is very large, then the probability that each component of the difference between simluated and observed data will be less than ϵ becomes very small, which results in extremely poor acceptance rates for the generated parameters. For SDEs, where the model data is typically a time-series of data points, this problem is very pronounced; the same parameter value can yield traces from the model that differ significantly, due to the randomness inherent in the system. Figure 5.1 illustrates the problem in the case of the GBM model. Despite being generated with the same parameter value, the sample paths from the GBM model all differ significantly due to the stochastic component of the SDE. If one were to try and implement the rejection ABC sampler for the GBM model by directly comparing time series generated from the model with the observed time series, it is clear that in order to maintain any sort of reasonable acceptance rate, the tolerance parameter, ϵ, would have to be chosen to be very large indeed, thereby drastically reducing the quality of the posterior approximation yielded

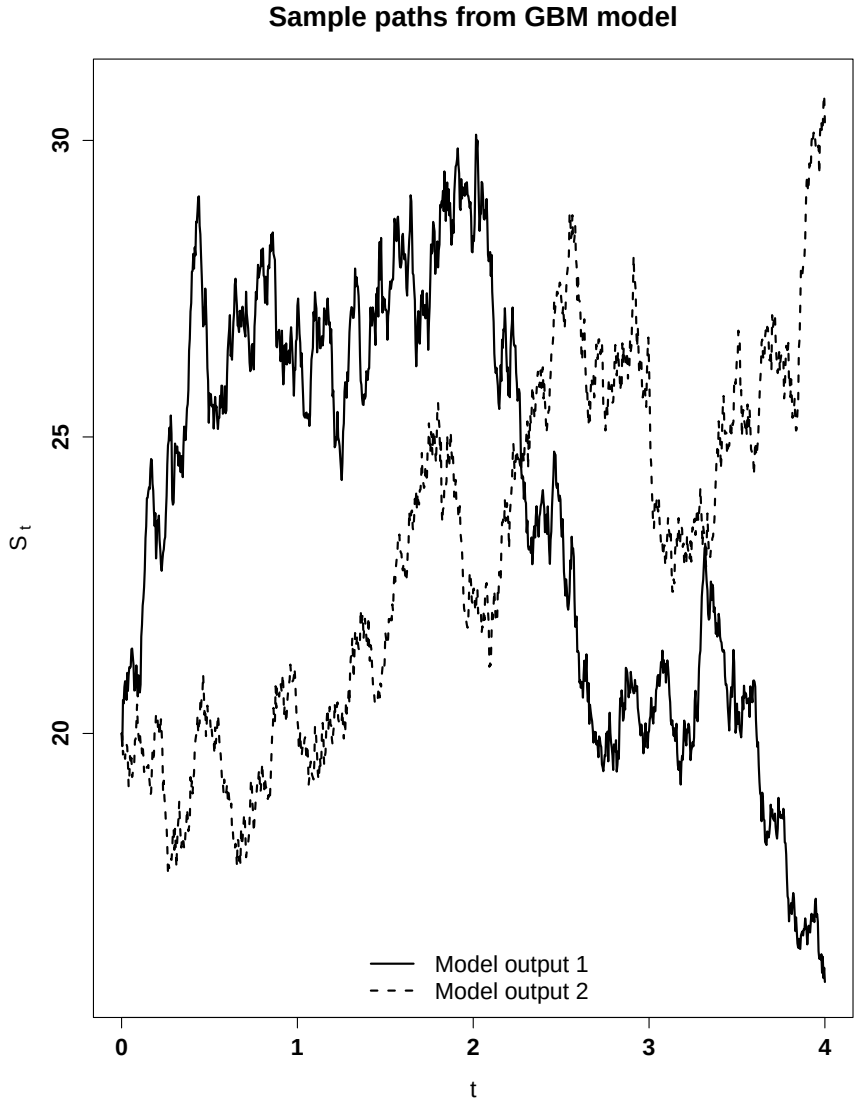

Figure 5.1: This figure contains two sample paths from the GBM model (5.2) using the same parameter value, $\theta = (0.07, 0.20)$, and initialised at the same starting point, $S_0 = 20.0$, demonstrating that traces generated using the same parameter values can result in significantly different sample paths.

by the sampler. In such situations, comparing each element of the observed and simulated data sets is not practical, and some means of capturing the information contained within the model output in a lower dimensional object must be found. This problem is considered further in the following subsections.

5.3.2 Summary statistics: some background theory

Broadly speaking, a statistic S is some (potentially vector-valued) function of a set of observations, assumed to have been generated by some parametric model $\mathcal{M}(\theta)$. Some basic examples are the sample mean, sample variance and interquartile range of a data set D. In the context of this thesis, we are interested in finding particular combinations of statistics that:

- Reduce the dimension of the model observations down to a manageable size. Ideally the number of statistics should be equal to the dimension of the unknown parameter θ.
- Capture as much of the information contained within the full data-set, D, as possible.

In statistics, a summary statistic S_{suff} is said to be sufficient if, for a given model and associated, unknown, parameter θ, no other statistic, S, can be calculated from a sample of observations that contains information about the parameter not already included within S_{suff} (Fisher, 1922). Stated in mathematical terms, sufficiency can be defined as follows:

Definition 8 (***sufficient summary statistics***)**.** *A statistic $S = r(D)$ is sufficient if, for each s, the conditional distribution of $D = \{D_1, D_2, \ldots, D_N\}$, given $S = s$ and parameter θ, is independent of θ,*

$$\mathbb{P}(D|S, \theta) = \mathbb{P}(D|S).$$

In practice, this definition is not particularly helpful as it is difficult to determine whether a particular set of statistics are sufficient using this criterion. Additionally, this definition is of no help in finding sufficient statistics. Fortunately, the above definition of sufficiency implies a more practically useful definition, which is given below:

Definition 9 (***factorisation criterion***)**.** *Let $D = \{D_1, D_2, \ldots, D_N\}$ be a random sample of observations with joint density given by $f(D|\theta)$. A statistic $S(D)$ is said to be sufficient iff the joint density of observations can be factorised as follows*

$$f(D|\theta) = g(S(D)|\theta) \cdot h(D),$$

where g and h are non-negative functions.

Thus, if a set of sufficient statistics S_{suff} can be found, inference with respect to the unknown parameter can be conducted by considering the observations only via consideration of the sufficient statistics, i.e. the sufficient statistics provide the same amount of information concerning θ as the full data-set D. Trivially, the collection of statistics $\{N, D_1, D_2, \ldots, D_N\}$ are always sufficient (Bernardo and Smith, 2000); however, the benefit of using summary statistics is in the ability to reduce the dimension of the model output to a more manageable size, and therefore we are

interested in finding the smallest number of sufficient statistics that represent the information in the model observations D. This motivates the following definition:

Definition 10 (***minimally sufficient statistics***)**.** *If $D = \{D_1, D_2, \ldots, D_N\}$ is a sequence of observations from some parameterised model $\mathcal{M}(\theta)$, and $S_{suff}(D)$ are a set of sufficient statistics, then $S_{suff}(D)$ are minimally sufficient iff, given any other set of sufficient statistics $T_{suff}(D)$, there exists a function $g(\cdot)$ such that*

$$S_{suff}(D) = g(T_{suff}(D)).$$

Intuitively, minimally sufficient statistics convey all information contained within the full sequence of observations D in the least number of statistics, i.e. minimally sufficient statistics convey information regarding the model parameters θ most efficiently. To illustrate the concept of sufficient statistics and demonstrate the means by which one might derive sufficient statistics using the factorisation criterion given above, consider the following example.

Example 1. *Let X be a sequence of n i.i.d. observations from a normal distribution with unknown mean and variance parameters,*

$$X = \{x_i\}_{i=1,\ldots,n}, \quad \forall i, x_i \sim \mathcal{N}(\mu, \sigma^2).$$

Then the likelihood of the observations is given by

$$f(X|\mu, \sigma) = C \cdot \sigma^{-n} \cdot \exp\left(\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{n} x_i^2 + 2\mu \sum_{i=1}^{n} x_i + n\mu^2\right]\right).$$

Hence, the joint distribution of the observations (the likelihood function) is a function of the data, X, only through the two functions $S_{suff}(X) = (\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2)$ which,

by the factorisation criterion given above, implies that S_{suff} are sufficient statistics for the normal model with unknown mean and variance parameters.

One can also demonstrate that the sufficient statistics given in the example above are minimally sufficient. In the applications for which the ABC approach to inference is valuable, sufficient summary statistics are not available. Typically, one identifies summary statistics for a model by first writing down the likelihood of the model observations and then factorising it as per the factorisation criterion above. If the analytic form of the likelihood function is not known, as is assumed to be the case in this thesis, then one cannot derive summary statistics in this manner, and therefore the statistician must find other, non-sufficient, summary statistics that capture as much of the information contained within the model observations as possible. Existing literature concerning the construction of such non-sufficient summary statistics, as well as the new methods developed in this thesis for the purposes of estimation of diffusion processes, will be presented in the next subsection.

5.3.3 Methods of constructing statistics

One of the most challenging areas in the design of efficient ABC samplers is the choice of suitable summary statistics, their purpose being to reduce the dimensionality of the observed data. Currently, most summary statistics are chosen on an ad hoc, model by model basis, and work on the development of generalisable, robust approaches to constructing summary statistics has been lacking in the literature. Joyce and Marjoram (2008) developed a method for choosing between a given list of approximately sufficient summary statistics; however, this work does not address the problem of how to actually choose the set of candidate summary statistics; instead, the exstence of such a list of candidate statistics is assumed, and focus

is on choosing good combinations of these candidates. For this reason, we do not consider this work any further. Fearnhead and Prangle (2012) developed a so-called semi-automatic method of constructing summary statistics for ABC inference that attempts to construct summary statistics that are as accurate as possible, with 'accurate' being defined in a specific way. What follows is an overview of their approach.

Semi-automatic summary statistics

Fearnhead and Prangle (2012) take a slightly different approach to ABC estimation than other researchers in the field; rather than using the ABC approximation to generate global approximations to the true posterior density of interest, they focus on generating approximations that yield accurate parameter estimates, where accuracy is defined in terms of a loss function for estimating parameters. By considering the class of quadratic loss functions, that is loss functions of the form

$$L(\theta_0, \hat{\theta}; A) = (\theta_0 - \hat{\theta})^T A(\theta_0 - \hat{\theta}),$$

where A is a positive definite matrix, θ_0 is the true parameter and $\hat{\theta}$ is an estimate of the parameter, Fearnhead and Prangle (2012) demonstrate that the optimal choice of summary statistic (i.e. the choice of statistic that leads to minimum quadratic loss), in the limit $\epsilon \to 0$[3], is the posterior mean of the parameter θ

$$S(D) = \mathbb{E}(\theta|D). \tag{5.5}$$

[3] ϵ represents the similarity kernel variance parameter that determines the degree of approximation between the ABC approximated likelihood and the true model likelihood.

In other words, if we choose our summary statistic to be equal to the posterior mean, then minimum quadratic loss is achieved by taking parameter estimates of the form

$$\hat{\theta} = \mathbb{E}_{ABC}(\theta|S(D)),$$

where $S(D)$ is given by (5.5). Using a quadratic loss function leads to ABC posteriors that attempt to match the mean of the true posterior. Using different choices of loss function will lead to ABC approximations that match other features of the true posterior, e.g. using an absolute error loss function will lead to ABC posteriors that match the median of the true posterior (Fearnhead and Prangle, 2012). In practice, of course, we cannot choose our summary statistics to be equal to the posterior mean—deriving approximations to the unknown, true posterior density is the aim of the analysis—therefore it appears that this result is academic. However, Fearnhead and Prangle (2012) produce estimates of the appropriate summary statistics by running an additional simulation step before running the full ABC sampler. Their approach is summarised as follows

- Run a pilot ABC sampler to determine regions of non-negligible posterior mass.
- Simulate sets of parameter values and generate data with each simulated parameter.
- Use the simulated parameters and model output to derive estimates of the summary statistics.
- Run ABC using the summary statistics derived during stage 3.

Step one is optional, and should be implemented if the priors being used are uninformative, its purpose being to focus on a particular training region of the

parameter space from which parameters are simulated. Once the training region of the parameter space has been determined, M parameter values are simulated from the prior[4] and M corresponding data sets generated from the model (step 2 above). In step 3, the simluated parameters and data are used to construct estimates of the appropriate summary statistics. Fearnhead *et al.* use linear regression for this stage of the analysis but mention that other approaches, for example the lasso, might also be used (see Tiribshani (1996) for more details concerning lasso regression). For the regression stage, the simulated parameter values generated in stage 2 are treated as response variables. The authors then introduce a (possibly vector valued) function of the simulated data, $f(X)$, and use this as the explanatory variable in the regression. The simplest choice of function is $f(X) = X$; however, the authors note that more complicated functions of the simulated data may yield better estimates of the summary statistics. For example, Fearnhead and Prangle (2012) used $f(X) = (X, X^2, X^3, X^4)$ in one application and found that this choice yielded superior summary statistics. Note that in this chapter we test the semi-automatic approach to summary statistic construction by using as explanatory variables the simulated data alone (i.e. $f(X) = X$) in the case of the GBM model, and both the simulated data and the squares of the data (i.e. $f(X) = (X, X^2)$), in the case of the CIR model. For the ith summary statistic, the following linear model is fitted using least squares:

$$\theta_i = \mathbb{E}(\theta_i|X) + \varepsilon_i = \beta_0^{(i)} + \beta^{(i)} f(X) + \varepsilon_i, \tag{5.6}$$

where ε_i is zero-mean noise. The linear function fitted during stage 3 serves as an estimate of the mean of the posterior, and is used as a summary statistic in the final stage of the estimation procedure. One advantage of this approach is that

[4]That is, the prior truncated to the training region of the parameter space.

one can use many, potentially hundreds, of explanatory variables in the regression, without directly affecting the efficiency of the ABC sampler—this is in contrast to the standard approach, in which using more summary statistics increases the dimensionality of the statistics and therefore reduces the efficiency of ABC, as per the discussion in section 5.3.1. In the applications considered herein, model output data are time-series, typically consisting of a large number of points. Assuming the statistician is fitting a model to daily observations, of which there are roughly 250 per year[5], of some financial quantity, it is not at all uncommon to be dealing with time series that are several thousand elements long; in this case, constructing summary statistics based on regressing the full path against the parameter value, as in the case of the linear model given by (5.6), can be impractical, especially if one considers vector valued functions of the full data set rather than the observations alone. Aside from the semi-automatic approach to constructing summary statistics developed by Fearnhead and Prangle (2012), no other generalised approach to selcting summary statistics has been presented (to our knowledge). Fearnhead *et al.*'s approach of constructing summary statistics via an additional simulation stage will be tested against our applications, alongside some ad hoc choices of summary statistics, developed with SDE parameter estimation in mind, which will now be presented.

Among the most straightforward approaches to parametric inference of SDEs is the so-called 'pseudo-likelihood' method, which involves assuming that the model of interest possesses a Gaussian transition density and then deriving the pseudo-likelihood function using this approximation. This method of parameter inference for SDEs is known to be ineffective for all but the simplest SDEs, due to the bias in the parameter estimates derived via this method. This bias can be significant, especially

[5]The number of trading days in a year is often approximated as being around 250.

in estimates of parameters appearing in the drift coefficient of the SDE under investigation. Despite this approach being unreliable due to the aforementioned problem with biased estimators, the pseudo-likelihood approximation at the core of the method might be of some use in deriving informative statistics (which will be referred to as Euler-Maruyama (EM) based statistics) of the data which can then be used in conjunction with an ABC sampler to produce samples from the ABC posterior. We now outline the ideas underlying the construction of EM based statistics.

EM based summary statistics

Earlier in this chapter, it was noted that diffusion processes, by construction, are Markov processes. The Markovianity of diffusion processes is a useful feature that allows such processes to be characterised by their initial states plus their transition densities, which makes simulating solutions to SDEs, as well as deriving model likelihoods, an easier task provided one can determine what the true transition density of the process is. Unfortunately, the transition density of a diffusion process is difficult to determine in all but the most straightforward cases. One way of deriving approximations to the transition density is to discretise the SDE using a crude approximation that essentially amounts to assuming that the transition density, i.e. the increments of the process, are normally distributed. For example, suppose we have sample observations, $X = \{X_k\}_{k=0,\ldots,N}$, where $X_k = X_{k\Delta t} = X(k\Delta t)$, from (5.1) at equally spaced points in time[6]. The likelihood of these observations, due to

[6]This assumption is not necessary for the purposes of constructing Gaussian approximations to the transition density of a diffusion process, but in the majority of applications in finance observations from processes of interest occur at regularly spaced intervals, e.g. daily closing share prices, therefore we will make this assumption to simplify the analysis.

the Markovian nature of the underlying process, can be represented as the product of the transition densities between neighbouring observations:

$$f(X|\theta) = \prod_{k=1}^{N} p(X_k|X_{k-1}, \theta), \tag{5.7}$$

where $p(X_k|X_{k-1}, \theta)$ is the transition density of the process. If the transition density is unknown, then one can discretise (5.1) as follows

$$\Delta X_{k-1} \equiv X_k - X_{k-1} = \mu(X_{k-1})\Delta t + \sigma(X_{k-1})\Delta B_{k-1}, \quad k = 1, \ldots, N,$$

where $\Delta B_{k-1} = B_k - B_{k-1}$ is an increment of Brownian motion which is normally distributed by definition ($\Delta B_{k-1} \sim \mathcal{N}(0, \Delta t)$). Thus, each increment of the process is assumed to be distributed as follows

$$\Delta X_{k-1}|X_{k-1} \sim \mathcal{N}(\mu(X_{k-1})\Delta t, \sigma^2(X_{k-1})\Delta t).$$

This crude approximation to the true SDE (5.1) is called the Euler-Maruyama (EM) approximation, and is frequently used when generating numerical solutions to SDEs. Substituting the EM approximation of the transition density into (5.7) yields the pseudo-likelihood approximation for (5.1)

$$\begin{aligned} f_{EM}(X|\theta) = (2\pi\Delta t)^{-N/2} \left[\prod_{k=1}^{N} \sigma^{-1}(X_{k-1})\right] \times \\ \exp\left\{\frac{-1}{2\Delta t}\sum_{k=1}^{N}\left(\frac{X_k - X_{k-1} - \mu(X_{k-1})\Delta t}{\sigma(X_{k-1})}\right)^2\right\}. \end{aligned} \tag{5.8}$$

Maximising this quantity with respect to the parameters, θ, would produce the pseudo-maximum likelihood parameter estimates (PMLEs) which, as already mentioned, are typically biased estimates of the 'true' parameters. Although PMLEs

are not suitable parameter estimates, it is possible to derive summary statistics from the expression for the pseudo-likelihood—the idea is that despite the parameter estimates that result from this approximation being unreliable, the summary statistics that one can derive from this approximation should contain information about the parameters and therefore can be used to differentiate between 'good' and 'bad' candidate parameter values. In the case of the GBM model introduced earlier, the drift and diffusion coefficients are given by

$$\mu(S_t) = \mu S_t, \quad \text{and } \sigma(S_t) = \sigma S_t$$

respectively. Substituting these expressions into (5.8), we obtain the pseudo-likelihood function for S_t

$$f_{EM}(S|\theta) = (2\pi\Delta t)^{-N/2} \left[\prod_{k=1}^{N} S_{k-1}^{-1}\right] \sigma^{-N} \times \\ \exp\left\{\frac{-1}{2\Delta t\sigma^2}\left(\sum_{k=1}^{N}\left(\frac{S_k}{S_{k-1}}\right)^2 - 2(1+\mu\Delta t)\sum_{k=1}^{N}\frac{S_k}{S_{k-1}} + N(1+\mu\Delta t)^2\right)\right\}.$$

Notice that we can factorise this pseudo-likelihood function into the product of two functions, one depending solely on the data, and one that depends on the model parameters, $\theta = (\mu, \sigma)$ and the data, but only via two functions of the data:

$$T_1(S) = \sum_{k=1}^{N}\frac{S_k}{S_{k-1}}, \quad T_2(S) = \sum_{k=1}^{N}\left(\frac{S_k}{S_{k-1}}\right)^2.$$

By virue of the factorisation theorem 9, we know that these two functions of the data must be sufficient statistics for the EM discretised SDE and capture information relating to the drift and diffusion coefficients, which we would like to estimate from observed data. We propose to combine these approximately sufficient

statistics with the ABC samplers introduced previously in order to infer the GBM model parameters from a time series of model observations. Similar calculations for the CIR model can be found in Appendix 5.7 at the end of this chapter.
Despite this technique of obtaining informative summary statistics working well, for some parameters, in the examples studied in this chapter, it is not always possible to obtain summary statistics in this manner. If the coefficients of the diffusion process exhibit non-linear dependence on the model parameters, it may be impossible to derive summary statistics from the pseudo-likelihood function due to the inability to separate out the parameters from the data. As such, this technique of constructing summary statistics would only be practical in applications in which it is possible to separate data summaries from the model parameters.

Moving average based summary statistics

In addition to the semi-automatic summary statistics and EM based statistics, we also trial an ad hoc summary statistic designed to capture information about parameters appearing in the diffusion coefficients of (5.1) and (5.3). Before explaining the rationale behind the formulation of this summary statistic, we outline the steps required to construct the statistic from a time series of observed data $D = \{D_i\}_{i=0,\ldots,N}$.

- Choose a smoothing window length and construct a smoothed (moving average) time series, D^s, from the observations, D.

- Subtract the smoothed path from the observations, and scale the resulting time series of differences by the smoothed series to obtain a time series of scaled residuals, i.e. evaluate

$$r_i^s = \frac{D_i - D_i^s}{D_i^s}, \quad i = 0, \ldots, N_s{}^7,$$

- Evaluate the standard deviation of the scaled residuals and use this as a summary statistic.

The motivation for the construction of this statistic can be appreciated by studying, for example, the form of the SDE describing GBM (equation (5.2)). First, note that in the absence of the drift coefficient of the SDE, the GBM process is essentially a Gaussian process with state-dependent variance. Thus, by constructing a moving average process and subtracting this smoothed path from the observations, we should, to a first approximation, obtain a Gaussian time series with time inhomogeneous variance. By looking at the form of the diffusion coefficient, we can deduce that the variance of the unscaled differences should depend on the level of the underlying process S_t; therefore, scaling the differences by S_t should result in a Gaussian time series with constant standard deviation equal to the diffusion parameter, σ. A graphical illustration of the data manipulation involved in the construction of the summary statistics is given in Figure 5.2. The effectiveness of this technique of construct summary statistics is limited to certain models, in a similar fashion to the EM based summary statistics introduced earlier. In particular, one must know the way in which the diffusion coefficient of the underlying diffusion process depends on the state variable in order to be able to scale the differences between the raw and smoothed time series appropriately. Therefore, models such

[7] $N_s < N$ is the length of the moving average process, constructed from the observations.

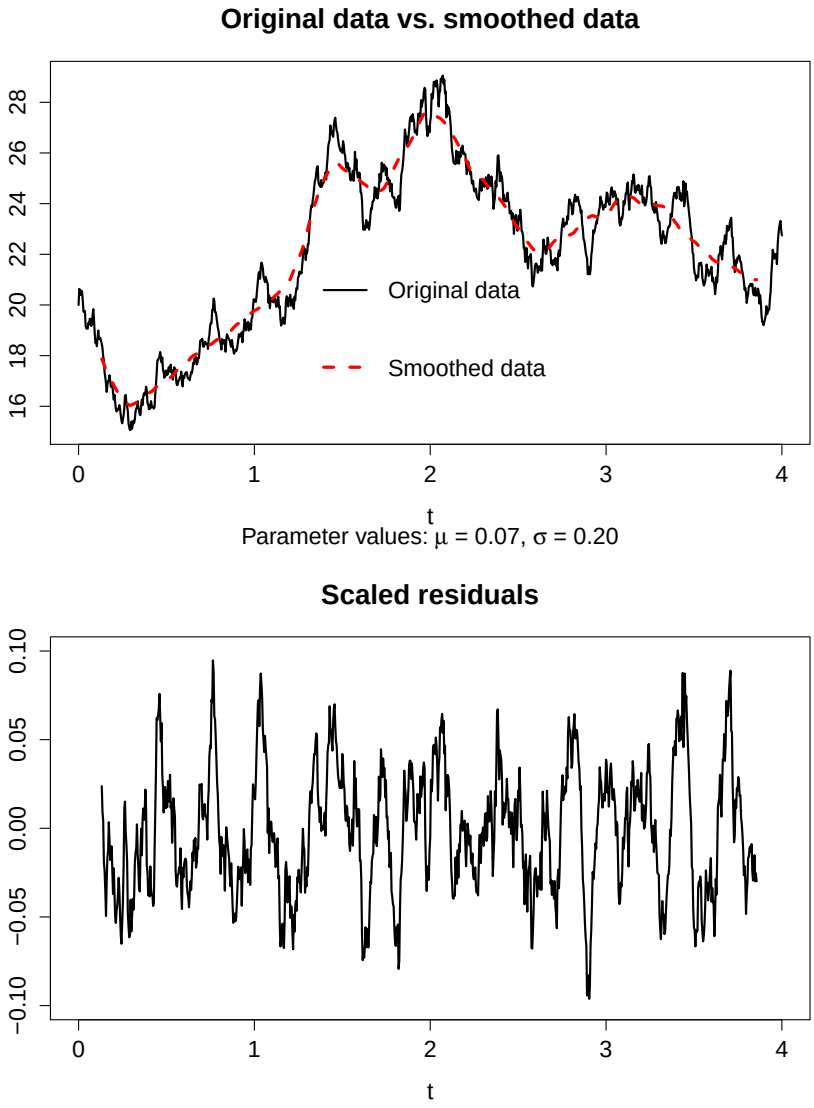

Figure 5.2: The top plot illustrates the raw observations from the model (the solid black line), and the smoothed observations that are derived from the raw observations (the dashed red line). The (scaled) residuals that result from taking the difference between the two series in the first plot, and then scaling the resulting series by the process value, are drawn on the bottom plot; the standard deviation of which is used as a summary statistic, which should contain information regarding the constant diffusion coefficient in the GBM model.

as the constant elasticity of variance (CEV) diffusion process, whose diffusion coefficient depends on the state variable raised to some unknown power, to be inferred from the observations, would not be amenable to this sort of summary statistic construction. As is generally the case at this stage in the development of ABC methodology, other ad hoc methods of constructing informative summary statistics would have to be determined for more complex models, such as the CEV process. The CIR model, given by (5.3), also possesses a diffusion coefficient that renders the construction of moving average (MA) based statistics possible, the only difference arising in the construction of the statistic for the CIR model is in the scaling of the differences between the smoothed path and the observations; in the GBM model the variance of the unscaled differences should be proportional to the level of the process S_t, whereas the differences in the CIR model should be proportional to the square root of the process, $\sqrt{R_t}$, which can be deduced from the form of the diffusion coefficient in (5.3)

Mean-gradient summary statistic

In order to capture information about the mean reversion rate parameter in the CIR model (5.3), we developed an ad hoc summary statistic based on the observation that the mean reversion rate determines the pace at which the process is pulled back towards the mean reversion level. Given this observation, one might expect a numerical approximation of the slope of the process, when drifting back towards the mean reversion level, to capture some information about the magnitude of the mean reversion rate. Larger values of the mean reversion rate should result in the process reverting relatively rapidly back to the long term level, implying steep gradients of the process; conversly, smaller mean reversion rates should translate into less steep

gradients as the process is pulled less aggressively towards the long term mean level. Henceforth we will label this summary statistic the 'mean-gradient' summary statistic. To understand the rationale behind the construction of the mean-gradient summary statistic, consider a deterministic process with dynamics given by

$$\mathrm{d}R_t = \alpha(\beta - R_t)\mathrm{d}t, \quad R_0 = r_0. \tag{5.9}$$

Note that this process is essentially the CIR model without the stochastic component of the SDE. If one wanted to produce an estimate of the mean reversion rate parameter, α, in (5.9) above, a simple approach might be to first carry out a simple Euler discretisation of the process, and then rearrange the resulting process in order to obtain $\hat{\alpha}$, the estimate of the mean reversion rate based on the observed data, $R = \{R_k\}_{k=0,\ldots,N}$, where $R_k = R_{k\Delta t} = R(k\Delta t)$, and the mean reversion level, β.

$$\begin{aligned} R_{k+1} - R_k &= \alpha(\beta - R_k)\Delta t \\ \implies \alpha &= \frac{R_{k+1} - R_k}{(\beta - R_k)\Delta t}. \end{aligned}$$

The sample average of this estimator for α, taken over all $N+1$ observed points $R = \{R_k\}_{k=0,\ldots,N}$ should yield useful information about the magnitude of the mean reversion rate parameter.

$$\hat{\alpha} = \frac{1}{N}\sum_{i=1}^{N} \phi(R_i, R_{i-1}, \hat{\beta}, \Delta t), \quad \phi(R_i, R_{i-1}, \hat{\beta}, \Delta t) = \frac{R_{i+1} - R_i}{(\hat{\beta} - R_i)\Delta t},$$

where $\hat{\beta}$ represents an estimate of the true mean reversion level parameter, β[8]. Of course, the CIR model differs from (5.9) due to the additional stochastic component of the CIR dynamics, and this stochastic component changes the nature of the dynamics of the process. The solution of (5.9) is represented graphically below. From Figure 5.3 one can see that the approximated process (5.10) monotonically approaches the mean reversion level, represented by the solid black horizontal line at $R_t = 0.10$. Due to the stochasticity of the full CIR process (5.3), the true solution of this process does not monotonically drift towards the mean reversion level, but instead exhibits volatility around the mean reversion level while gradually being pulled towards it. Due to this difference between the approximate process and the full CIR process, when evaluating the mean-gradient summary statistic we only use segments of the time series of observations in which the process is moving towards the mean reversion level. This is required so that the mean-gradient statistic always gives strictly positive estimates of the mean reversion rate parameter, α.

OLS-based summary statistics

In this subsection we outline another approach for constrcuing non-sufficient summary statistics that we utilised during the estimation of the CIR model parameters. If (5.3) is discretised using an Euler-Maruyama approximation, the resulting difference equation can be rearranged to give the following relationship:

$$Z_k = \alpha\beta X_{k,1} - \alpha X_{k,2} + \nu U_k, \quad k = 0, \ldots, N-1$$

[8]In practice the true mean reversion rate of the process is unknown, so we use the sample mean as an estimate of the mean reversion level parameter when constructing the mean-gradient summary statistic. The sample mean performed well as an estimator for the mean reversion level during our preliminary testing of the summary statistics—see section 5.3.4 for more details.

where $Z_k = \frac{R_{k+1}-R_k}{\sqrt{R_k \Delta t}}$, $X_{k,1} = \sqrt{\frac{\Delta t}{R_k}}$, $X_{k,2} = \sqrt{R_k \Delta t}$, and $U_k \sim \mathcal{N}(0,1)$, which can be used to construct the following linear regression model:

$$\begin{aligned}
\boldsymbol{Z} &= \boldsymbol{X} \cdot \boldsymbol{\Theta} + \nu \boldsymbol{U}, \\
\boldsymbol{Z} &= (Z_0, \ldots, Z_{N-1})^T, \\
\boldsymbol{\Theta} &= (\alpha\beta, -\alpha)^T, \\
\boldsymbol{U} &= (U_0, \ldots, U_{N-1})^T, \\
\boldsymbol{X} &= \begin{pmatrix} X_{0,1} & X_{0,2} \\ \vdots & \vdots \\ X_{N-1,1} & X_{N-1,2} \end{pmatrix}.
\end{aligned}$$

This equation can be solved analytically in order to obtain the OLS estimators of the model parameters, which we use as non-sufficient summary statistics in the ABC samplers that we will test later in this chapter.

Sample mean statistic

One final summary statistic we use to capture information about the mean reversion level parameter, β, in (5.3) is the sample mean of the time series of observations. On an intuituve level, one would expect the mean reversion level to strongly influence the average level of the process over time, i.e. a higher mean reversion level would produce model traces with a higher sample mean, and, conversly, a lower mean reversion level should be reflected by sample traces possessing a smaller sample mean.

Table 5.1 sets out the different combinations of summary statistics that will be used in the numerical experiments associated with the GBM model. Table 5.2

sets out equivalent information relating to the numerical experiments associated with the CIR model. For clarity, the summary statistic descriptions in these tables have been condensed: *Semi-automatic: regression*, *Semi-automatic: lasso*, and *Semi-automatic: EM regression* refer to the semi-automatic summary statistics of Fearnhead and Prangle (2012), constructed using linear regression against the full sample path, regularised regression (lasso) against the sample path, and linear regression against the EM summary statistics outlined in section 5.3.3, respectively. The statistic descriptions in Table 5.2 follow the same format as those used in Table 5.1, the only difference being that the *Semi-automatic: regression* statistics were constructed by regressing the model parameters against the sample path and squares of the sample path. The non-sufficient EM summary statistic used to estimate the mean reversion level parameter, β, in the CIR model was statistic 4 in Appendix 5.7.

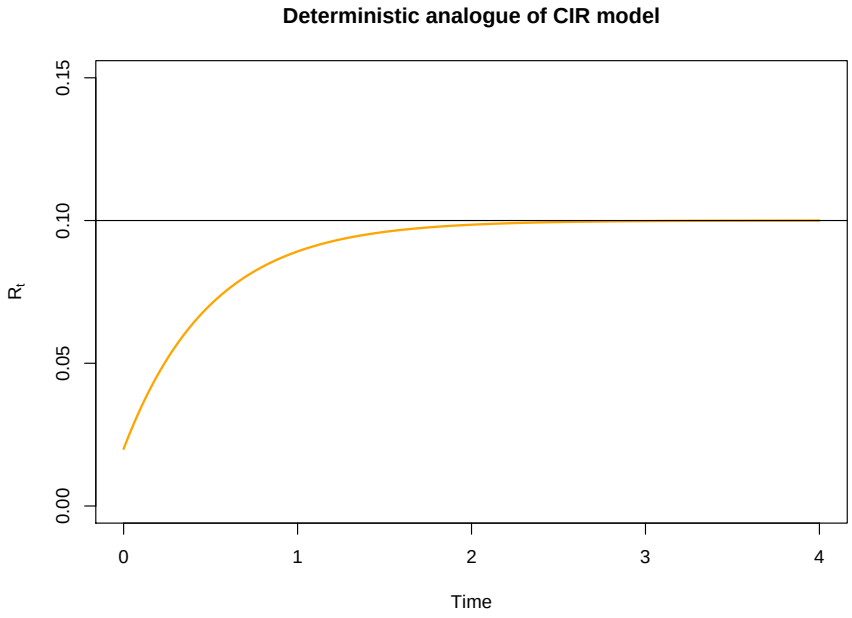

Figure 5.3: The orange line represents the solution of equation (5.9), conditional on $R_0 = 0.02$. The solid black line represents the mean reversion level, which was chosen to be 0.1. The mean reversion rate used to produce this plot was 2.

Table 5.1: Summary statistic combinations for GBM model

Experiment	Drift parameter	Diffusion parameter
1	Sufficient statistic	Sufficient statistic
2	Semi-automatic: regression	Semi-automatic: regression
3	Semi-automatic: lasso	Semi-automatic: lasso
4	Semi-automatic: EM regression	Semi-automatic: EM regression
5	EM statistic	EM statistic
6	Semi-automatic: regression	MA based statistic
7	Semi-automatic: lasso	MA based statistic
8	Semi-automatic: EM regression	MA based statistic
9	EM statistic	MA based statistic

Table 5.2: Summary statistic combinations for CIR model

Experiment	Rate parameter	Level parameter	Volatilty parameter
1	OLS statistic	Sample mean statistic	MA statistic
2	Mean-gradient statistic	Sample mean statistic	MA statistic
3	Semi-automatic: regression	Sample mean statistic	MA statistic
4	OLS statistic	OLS statistic	MA statistic
5	Mean-gradient statistic	OLS statistic	MA statistic
6	Semi-automatic: regression	OLS statistic	MA statistic
7	OLS statistic	EM statistic	MA statistic
8	Semi-automatic: regression	EM statistic	MA statistic
9	OLS statistic	Semi-automatic: regression	MA statistic
10	Mean-gradient statistic	Semi-automatic: regression	MA statistic
11	Semi-automatic: regression	Semi-automatic: regression	MA statistic
12	OLS statistic	Semi-automatic: lasso	MA statistic
13	Mean-gradient statistic	Semi-automatic: lasso	MA statistic
14	Semi-automatic: regression	Semi-automatic: lasso	MA statistic

In the following subsection we test each of the summary statistics outlined above.

5.3.4 Assessing choices of summary statistics

Having presented the various ways in which one might construct summary statistics for the parameter estimation of SDEs, we now focus attention on the problem of determining the extent to which a particular choice of summary statistic is informative. Obviously, the statistician can always pick some form of summary statistic and run the sampler to obtain empirical posterior distributions, thereby gaining some information about how well the chosen statistics capture information pertaining to the model parameters; however, if the underlying posterior is unknown, as is assumed to be the case, then assessing the quality of a particular choice of summary statistic becomes difficult. In addition, it is not practical to have to run the sampler for each choice of statistic to assess their effectiveness, especially if there are time constraints on the analysis. Before explicitly testing each type of summary statistic by estimating the GBM and CIR models, we ran some preliminary diagnostics to try and determine which summary statistics were likely to perform best. The steps involved in this diagnostic test were as follows:

- We simulated a sequence of parameter values over the support of the prior distribution of the model parameters.
- For each simulated parameter, we generated a sample path from the underlying model.
- For each simulated data set, we then evaluated the candidate summary statistic and plotted the statistic values obtained for each data set and parameter value against the sequence of parameter values.

The idea behind this diagnostic is that the summary statistics, if informative about the model parameters, should be able to map the simulated data back to the parameter used to generate the data. Hence, if the choice of summary statistic is good, the resulting plot should exhibit a rough one-to-one relationship between the parameter value and the summary statistic obtained from the model data generated by the parameter value. Figure 5.4 illustrates the results of the diagnostic test for the various statistics used to estimate the GBM model. As expected, the sufficient statistics for the GBM model are able to differentiate between data generated by different parameter values. All the other, non-sufficient, statistics for the drift parameter (plots in the left hand column of 5.2) appear to be able to differentiate between statistic values at least to some extent, with the EM based statistic and the semi automatic construction using the EM statistics yielding the best results from the diagnostic test. With the exception of the moving average based statistic, all non-sufficient summary statistics relating to the diffusion parameter (right column of plots) exhibit a very weak relationship to the parameter value, indicating that these statistics do not do a good job of mapping observed data back to the underlying parameter used to generate the observed data. From this preliminary analysis, we can conclude that some combination of either the EM based statistic or the semi-automatic EM regression statistic for the drift parameter, coupled with the MA based statistic for the diffusion parameter is most likely to yield the best results out of all combinations of non-sufficient summary statistics.

Similar diagnostic tests were carried out for the candidate summary statistics related to the CIR model. The results of the diagnostic test for the most promising statistic candidates[9] are illustrated in figures 5.5, 5.6, and 5.7. One can see that

[9]Due to the large number of possible combinations of summary statistics that could be used to estimate the parameters of the CIR model, we first ran a pilot ABC SIS sampler for each possible combination of summary statistics. Combinations of summary statistics that did not produce promising results were discarded.

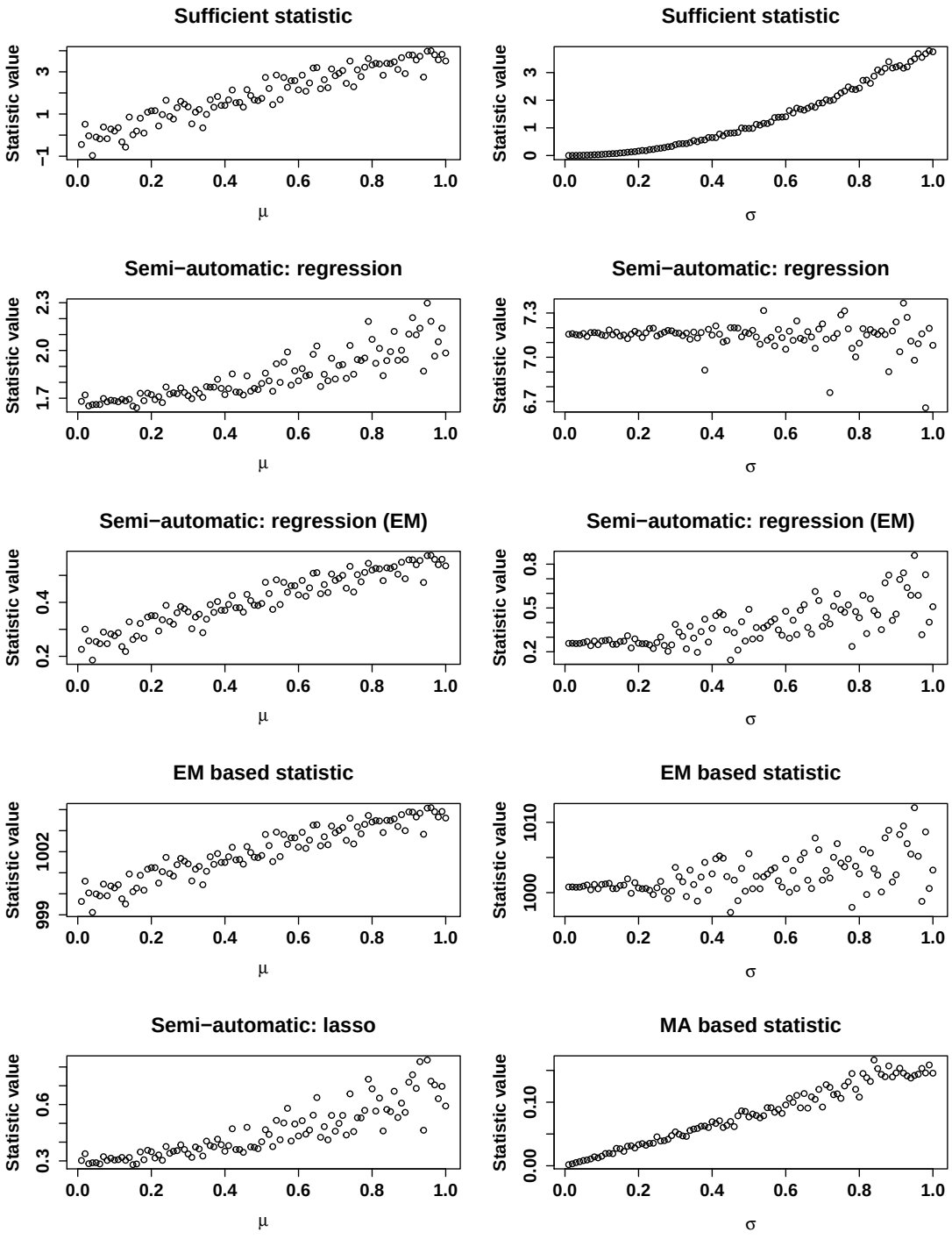

Figure 5.4: Results of the summary statistic diagnostic test for all types of summary statistic utilised for estimation of the GBM model. Good statistic choices should lead to clear one-to-one relationships between the parameter and the statistic value, as is illustrated in the top two plots of the sufficient statistics.

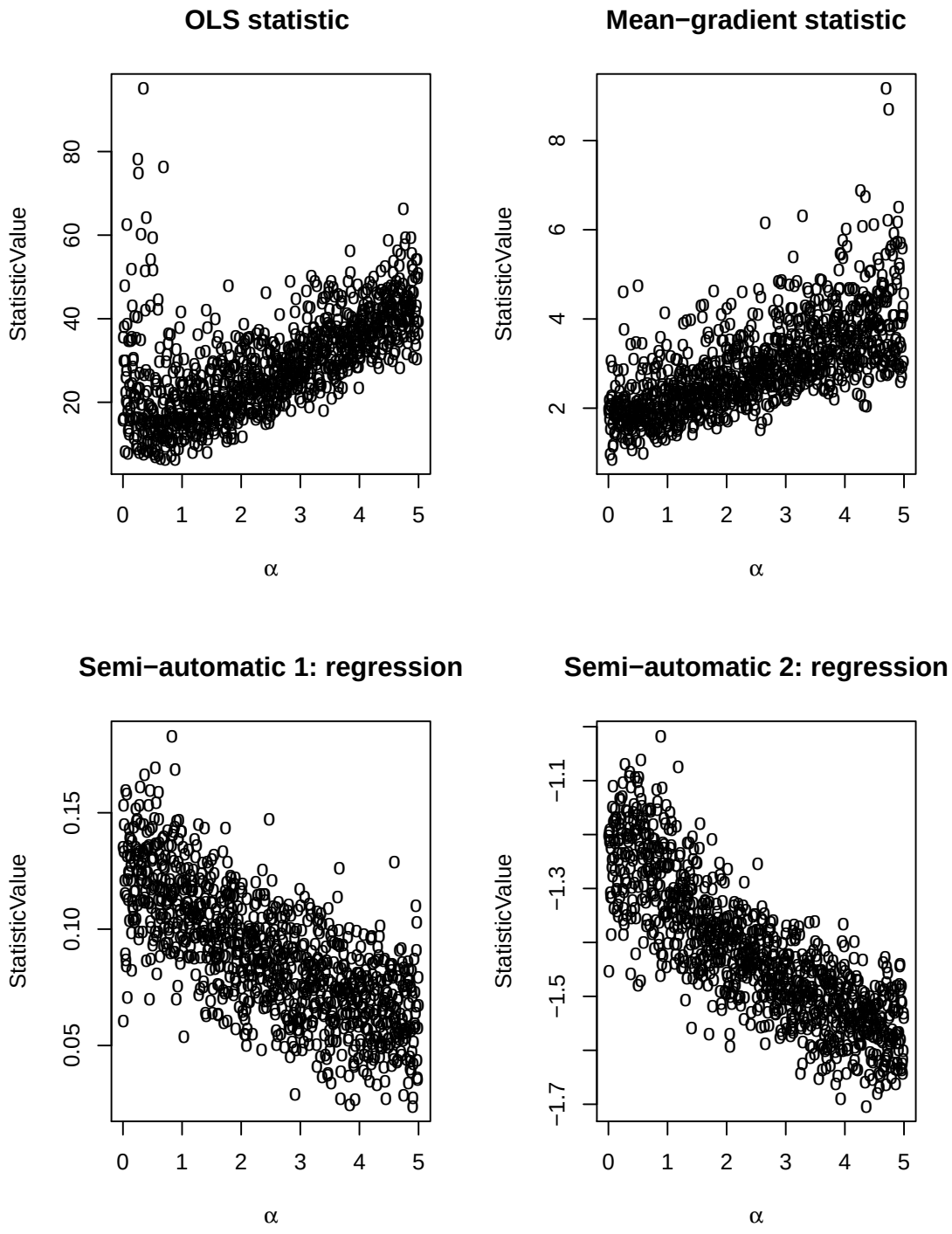

Figure 5.5: Results of the summary statistic diagnostic test for all types of summary statistic utilised for estimation of the mean reversion rate parameter in the CIR model. Good statistic choices should lead to clear one-to-one relationships between the parameter and the statistic value.

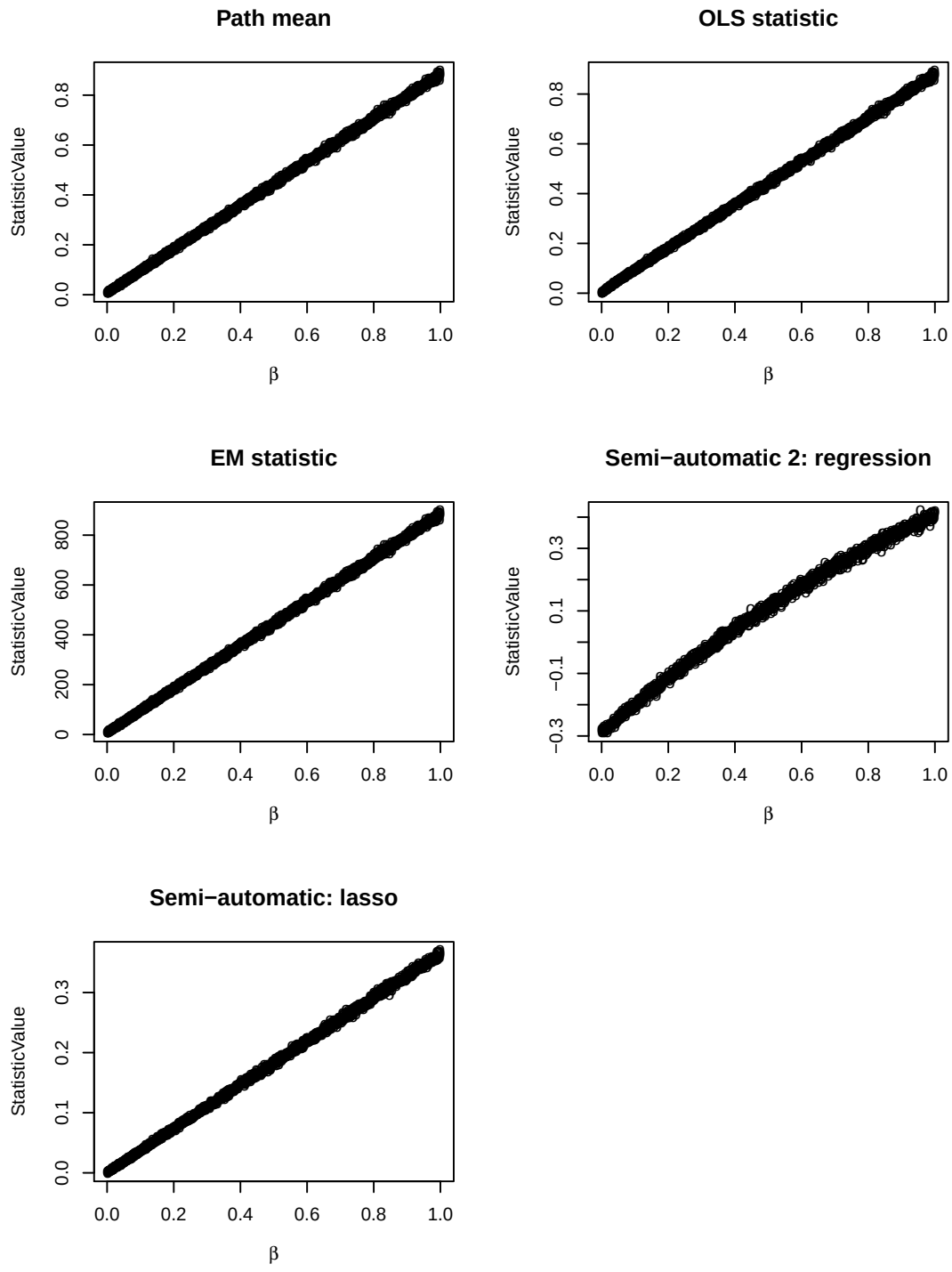

Figure 5.6: Results of the summary statistic diagnostic test for all types of summary statistic utilised for estimation of the mean reversion level parameter in the CIR model. Good statistic choices should lead to clear one-to-one relationships between the parameter and the statistic value.

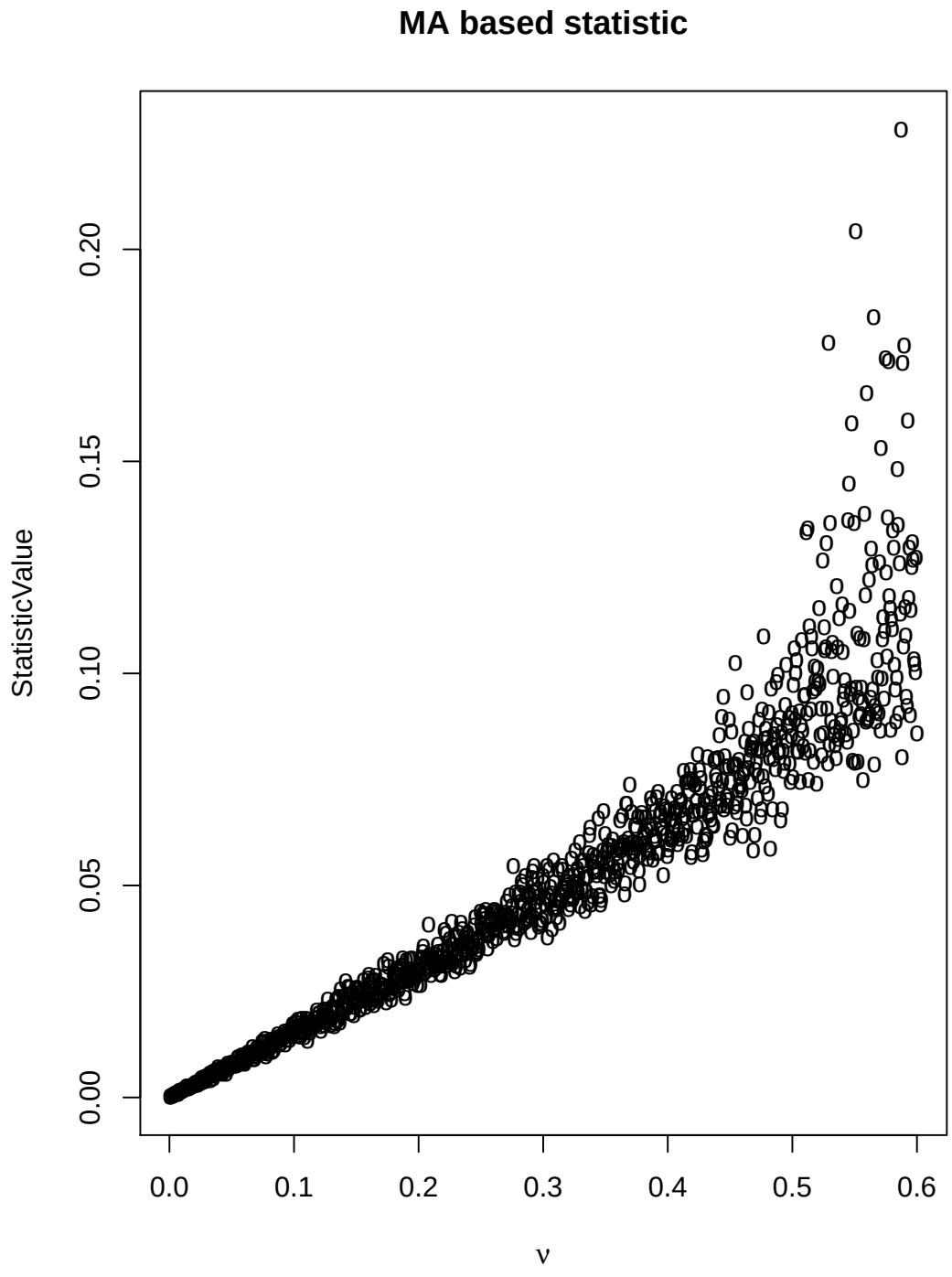

Figure 5.7: Results of the summary statistic diagnostic test of the moving average based summary statistic used to estimate the volatility parameter in the CIR model. Good statistic choices should lead to clear one-to-one relationships between the parameter and the statistic value.

the summary statistics chosen for the estimation exercise all exhibit a rough one to one relationship with the parameter value used to generate the data, indicating that these statistics are capable of capturing information about the data generating parameters. Having now introduced the various methods used to construct summary statistics, we now summarise the experimental set up that was chosen to test the two Monte Carlo based ABC (MC ABC) samplers introduced in the last chapter.

5.4 Simulation study design

In this section we outline the details relating to the numerical experiments conducted while testing the samplers.

5.4.1 Practicalities

In the numerical experiments conducted here, we used toy data generated from the underlying process with fixed, known parameter values[10]. In theory, real market data could have been used to test the samplers; however, given that the purpose of the experiments is to test the ability of the samplers to recover the true parameter values that generated the data (or, in Bayesian terms, to recover the true underlying posterior distribution of the model parameters), it made most sense to conduct the inference using data generated from the underlying process with known parameters. Daily closing prices are very typical of the data that one has access to when fitting a model to market observations, therefore when constructing the toy data used in this thesis, we assumed that our data consisted of 4 years worth of daily observations, of which there are roughly 250 per year, giving an observed time series of length 1000. The data were generated in each experiment by simulating a sample path from the underlying SDE (equations (5.2) and (5.3)

[10]The parameters chosen for generation of the underlying data from the GBM model were $\theta = (0.07, 0.20)$, the true CIR model parameters were $\theta = (2.0, 0.1, 0.06)$

respectively) using the Euler-Maruyama discretisation with a very fine step size[11]. Both Monte Carlo ABC samplers tested were implemented in the C++ language and compiled using the GCC compiler. As discussed previously, the similarity kernel variance parameter ϵ plays a crucial role in ABC, controlling the trade off between the accuracy of the ABC posterior approximation to the true posterior and the efficiency of the sampling algorithm. In practice, this parameter was chosen by trial and error, with as small a number as possible being chosen such that the samplers still produced acceptable populations[12].

5.4.2 Methodology

We will test the efficacy of two MC ABC samplers: the Tempered ABC SIS sampler and the new ABC MCMC sampler introduced in the previous chapter. Due to the biased nature of the Tempered ABC SMC sampler, which was demonstrated earlier, we will not provide results for this procedure.

For each sampler, we assess the quality of the empirical ABC posterior distributions obtained using various combinations of summary statistics by comparing the empirical densities (both the marginal and full densities) against the analytic posterior for the underlying models. For the estimation of the GBM model parameters, all prior densities were assumed to be uniform over a suitable range of values[13], which represents a reasonable window within which the parameters are likely to lie. The prior densities specified for the CIR model were slightly more complicated; model parameters were sampled uniformly from a constrained parameter space such that the Feller condition (5.4) was always satisfied. This was achieved by first uniformly

[11]The step size used to simulate model observations was $1,000,000$ per time period.

[12]For Tempered ABC SIS, this equates to the sampler producing a particle population with sufficiently large effective sample size; for ABC MCMC, this meant choosing the parameter to be as small as possible while maintaining a reasonable acceptance rate for the sampler.

[13]The parameter range was $(0, 1)$ for both drift and diffusion parameters.

sampling the mean reversion rate and level parameters from a suitable range of values[14], and then using a simple rejection sampling step to sample the diffusion parameter such that (5.4) was satisfied. When implementing the Tempered ABC SIS sampler, we used a linear sequence of tempering parameters in both sets of experiments to gradually evolve the target population from the prior to the posterior density, i.e.

$$\phi_t = \frac{t}{T}, \quad t = 0, \ldots, T.$$

If applying Tempered ABC SIS to other, more complicated, models, it may be necessary to investigate different tempering sequences to determine which sequence gives the best samples (i.e. samples with the least particle degeneracy), but for the relatively straightforward models estimated here, a simple linear sequence proved to be sufficient. For a more detailed discussion of the choice of tempering sequence, see Calderhead and Girolami (2003).

5.5 Simulation study results

5.5.1 GBM model

Due to the number of experiments being run (both the Tempered ABC SIS and new ABC MCMC samplers will be run using each of the summary statistic combinations set out in Tables 5.1, giving a total of 18 experiments), we will label each setup with a number and use the number to refer to the particular experiment. There are nine different combinations of summary statistics, therefore we will label each combination from one to nine, in the order they are given in table 5.1. So, for example, the label *experiment 3* refers to the results of the Tempered ABC SIS

[14]The mean reversion rate parameter was sampled from the range $(0, 5)$, and the mean reversion level parameter was sampled from the range $(0, 1)$.

sampler using semi-automatic summary statistics constructed using the lasso, and *experiment 5 (MCMC)* refers to the results of the new ABC MCMC sampler using EM based statistics. We first present a series of plots of the full joint posterior sample obtained from each experiment, plotted alongside the contours of the analytic posterior for comparison. Both experiment 1 and experiment 1 (MCMC) (figures 5.8 and 5.9) produced samples that matched the analytic posterior very closely, demonstrating that when the summary statistics are sufficient, both MC ABC samplers are capable of producing good approximations to the model posterior. All other choices of (non-sufficient) summary statistic produced joint posteriors that were inferior when compared with the posterior generated using sufficient statistics, indicating that the non-sufficient summaries used did not capture as much parametric information contained within the model observations as the sufficient statistics did. This is particularly evident in the case of the diffusion parameter, σ—all semi-automatic approaches to constructing data summaries, as well as EM based summary statistics failed to produce accurate inferences about this parameter. The MA based summary statistics were significantly more informative, leading to joint posteriors that more closely resembled the analytic posterior, although it appears that there may be a small upward bias in estimates of the diffusion parameter obtained with this choice of statistic.

To further examine the quality of the samples derived via the two MC ABC samplers, figures 5.26 and 5.27 compare the marginal posterior densities produced via the two MC ABC samplers with the analytic marginal posteriors. Once again, the experiments utilising sufficient statistics produces extremely good approximations to the true marginal posteriors, with very little discrepancy between the approximate and exact densities. In each of the other experiments utilising non-sufficient summary statistics, the densities derived via Tempered ABC SIS appear to yield

better approximations to the true posterior than the ABC MCMC sampler, particularly when semi-automatic summary statistics were used (experiments two, three, six, seven and eight). Figure 5.27 clearly demonstrates that all methods of constructing summary statistics for the diffusion parameter, with the exception of the MA based statistics, are ineffective. The MA based summary statistics are able to capture some information relating to the diffusion parameter, σ; however, as mentioned previously, there seems to be a small bias present in the empirical posterior densities derived using this statistic, with the mode of the empirical posterior (the maximum a posteriori (MAP) estimate) located slightly to the right of the analytic MAP in all densities produced via this statistic.

Having generated our approximations to the model's posterior distribution, we now demonstrate how one might go about constructing predictions about the future state of the process, given the historical observations used in the parameter estimation stage of the analysis. The *posterior predictive density (PPD)*, which we will denote by $f(\tilde{S}|S)$, where S represents historical observations from the model, is the predictive density of a new independent set of observables, $\tilde{S}$, generated from the model (5.2), conditional on information contained in the actual observations Gilks et al. (1996). In frequentist statistics, one might first obtain the optimal parameter estimate (via maximum-likelihood, or via some other technique such as moment matching) and then generate many traces from the model using these parameter estimates in order to build up a picture of how the process might evolve in the future; however, this approach fails to take into account the uncertainty in the parameter estimates themselves—this approach generates an empirical sample from the distribution of the process conditioned on one particular parameter value. In contrast, the PPD can be used to evaluate the distribution of new observations

that takes into account the uncertainty in the model parameters in addition to the stochasticity of the model itself. The PPD is computed as:

$$\begin{aligned} f(\tilde{S}|S) &= \int_{\mathcal{D}(\theta)} f(\tilde{S}, \theta|S) \mathrm{d}\theta \\ &= \int_{\mathcal{D}(\theta)} f(\tilde{S}|S)\pi(\theta|S)\mathrm{d}\theta, \end{aligned}$$

where $\pi(\theta|S)$ represents the posterior density of parameters. From this expression, one sees that the PPD averages the conditional distribution of new observations against the posterior knowledge about the observables, which is encapsulated in the posterior distribution of parameters. In Figure 5.28 we illustrate the credible intervals of the PPD of four years' worth of new data generated from the GBM model, using the best approximation to the model posterior obtained during the numerical experiments[15]. We also overlay the observations used to generate our posterior density approximations, denoted by the red line. One clearly observes that the model observations, S, used to generate the model posterior (the red line in Figure 5.28) are well within the 95 percent PPD credible intervals, indicating that the model fits the observed data well[16].

[15]The results obtained from Tempered ABC SIS experiment 8 were used to generate this data.

[16]Note that in this experiment the observed data were generated from (5.2), and so one would expect the model to fit the data well. This sort of model checking becomes useful when one attempts to fit a model to extraneous data, e.g. observations of some market variable such as a share price quoted on an exchange

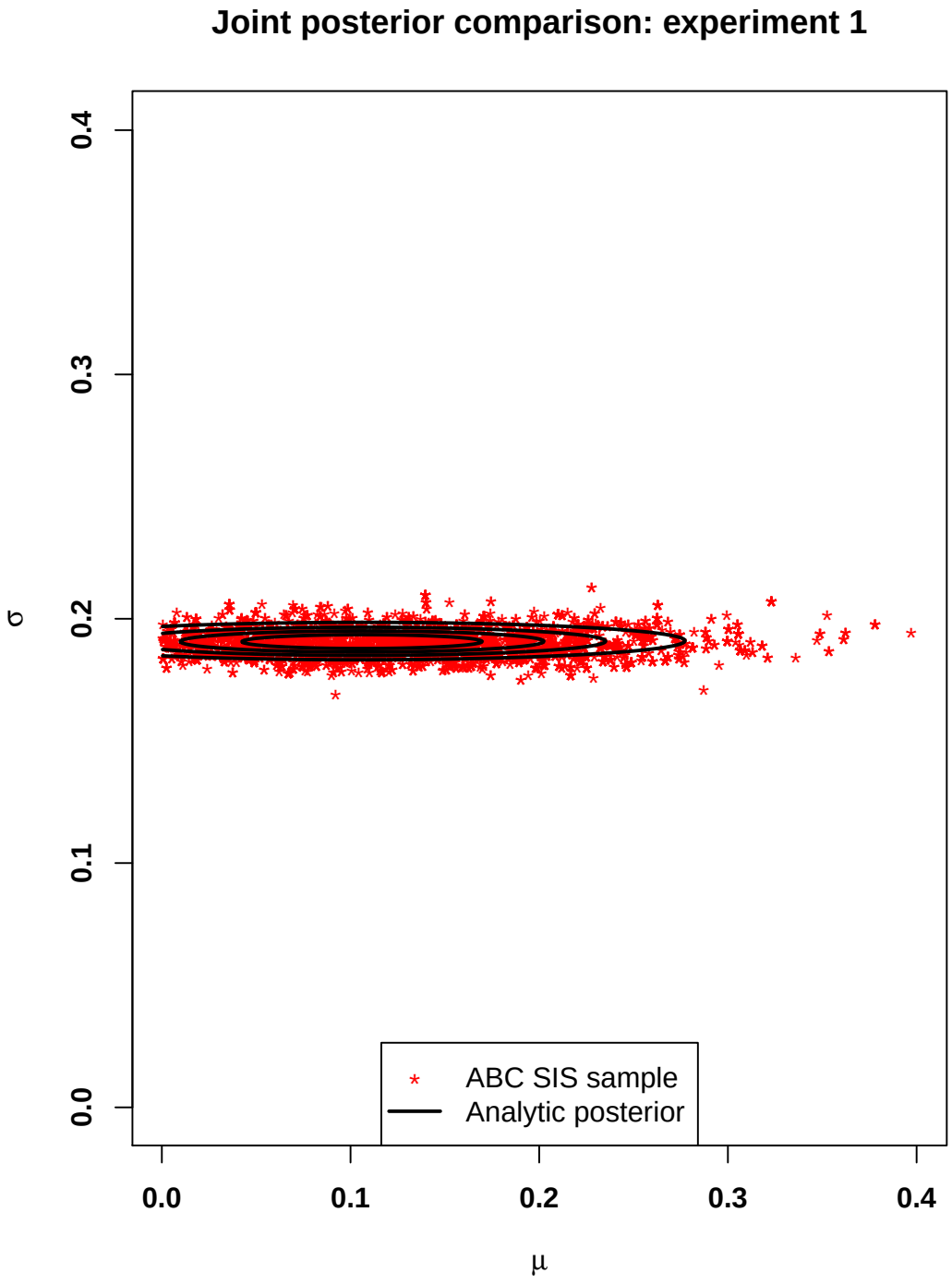

Figure 5.8: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with sufficient summary statistics. The correlation structure of the ABC posterior matches the true posterior's correlation structure well, indicating that the sampler is effective provided good summary statistics can be found for the model.

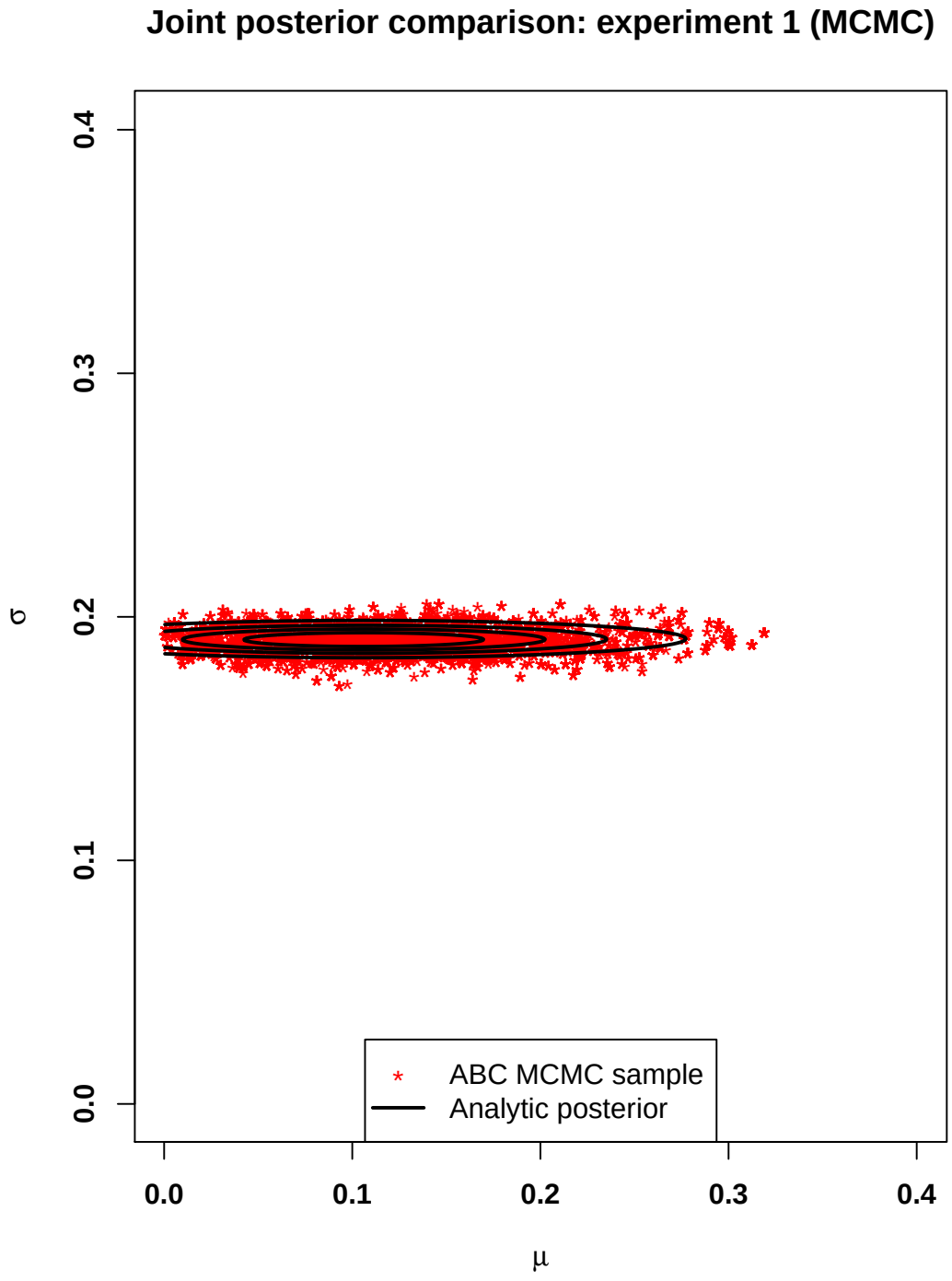

Figure 5.9: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with sufficient summary statistics. As in figure 5.8, the empirical distribution matches the analytic distribution closely.

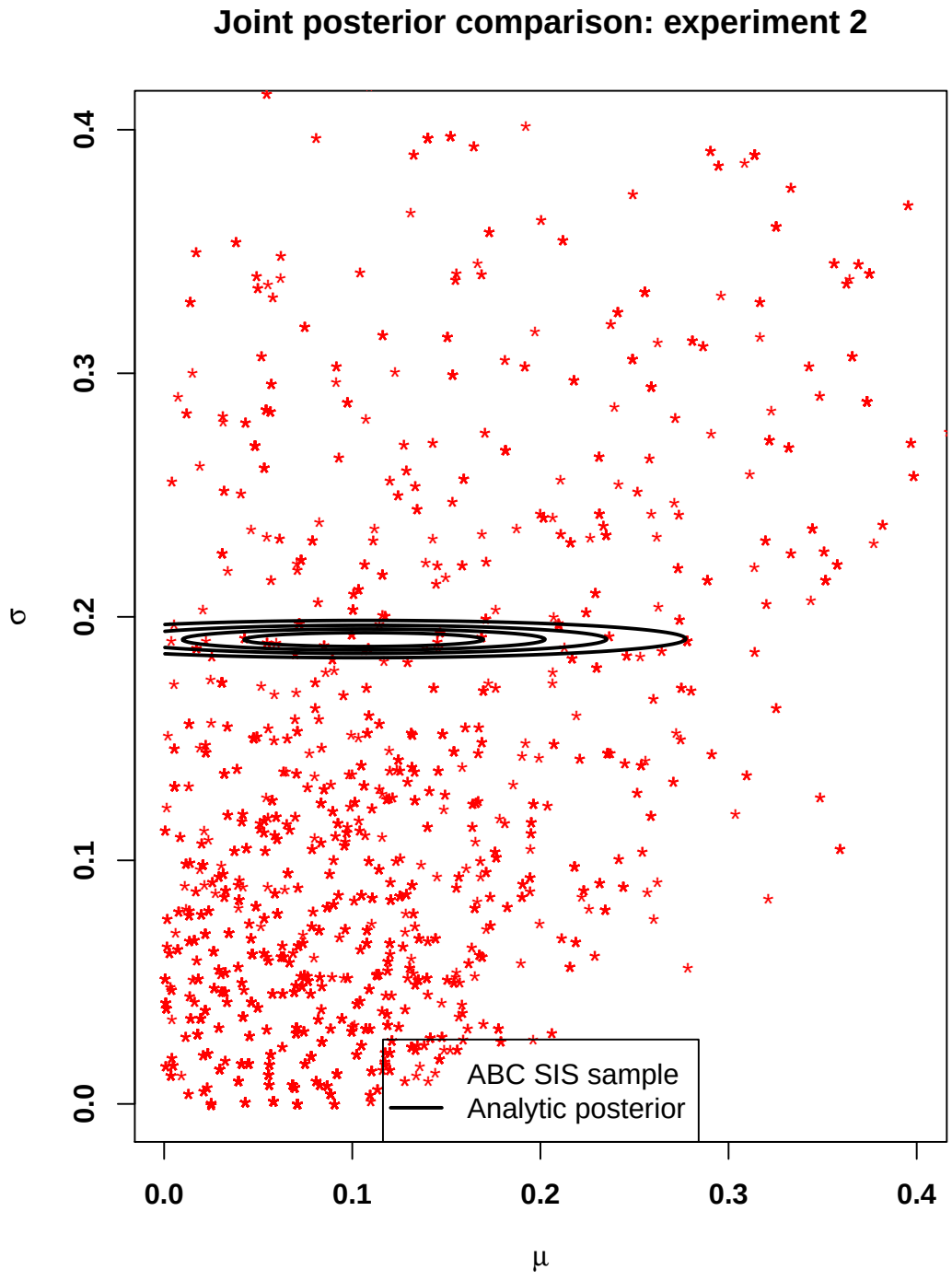

Figure 5.10: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with semi-automatic summary statistics derived using least squares regression.

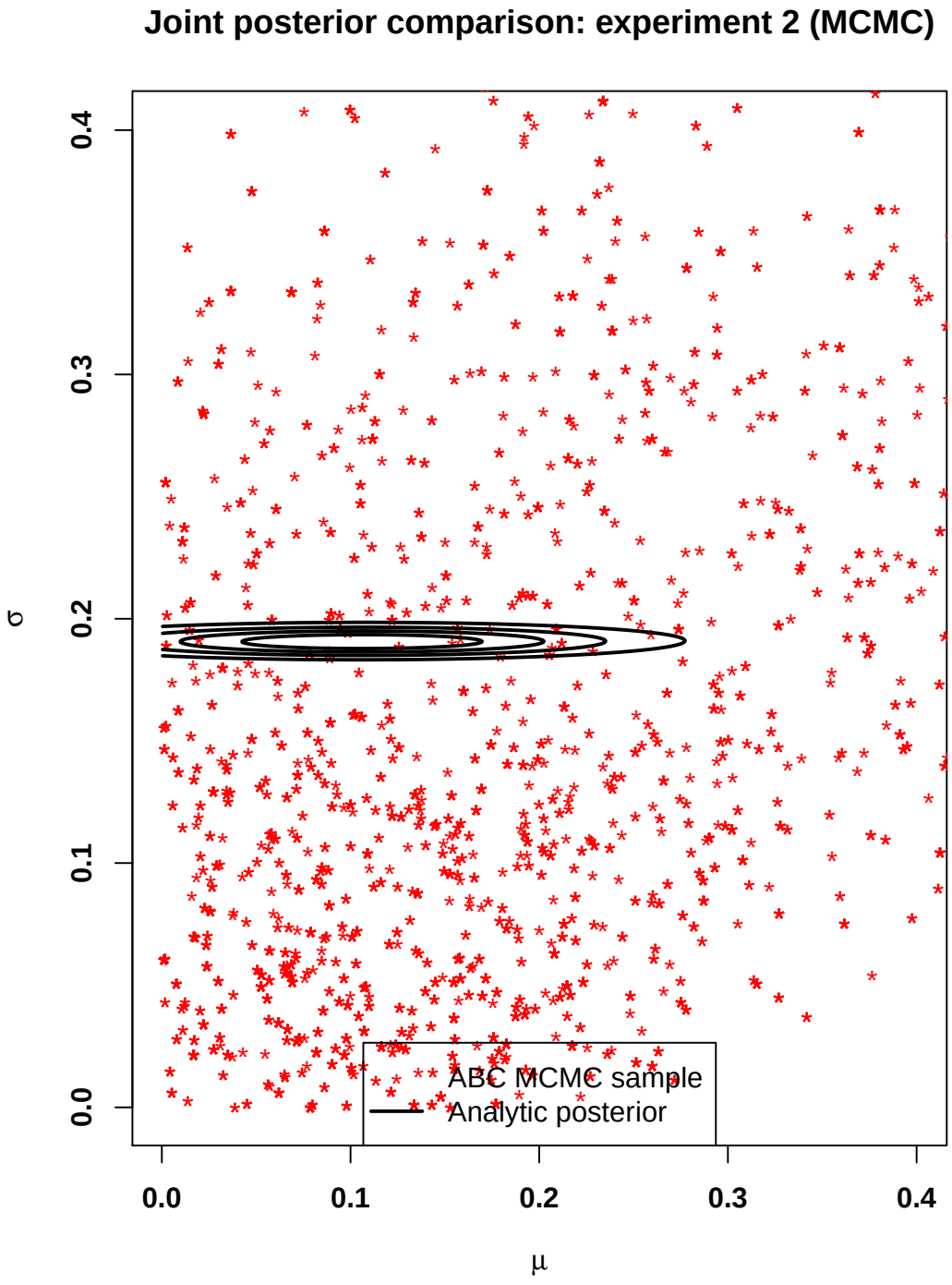

Figure 5.11: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with semi-automatic summary statistics derived using least squares regression.

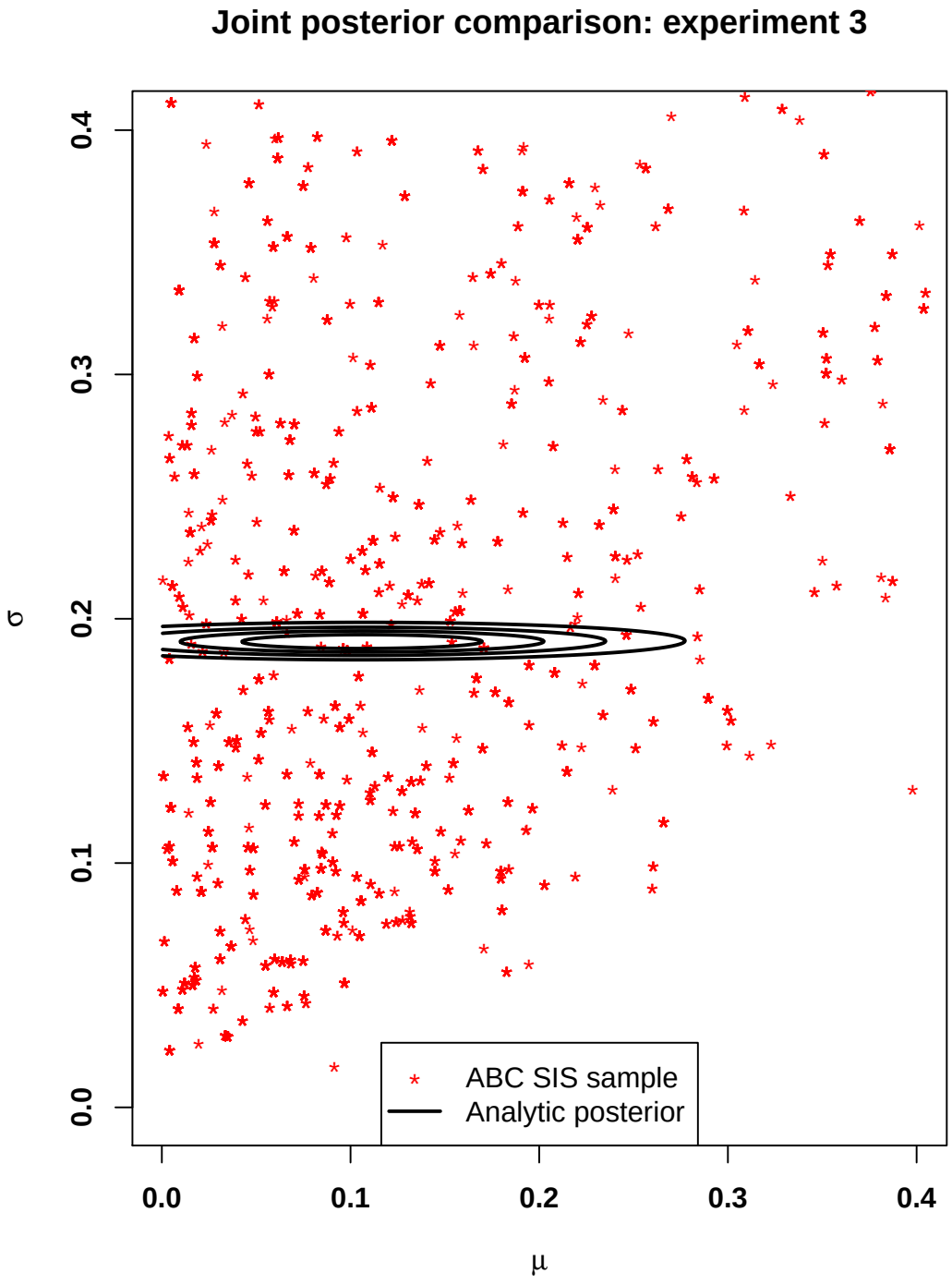

Figure 5.12: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with semi-automatic summary statistics derived using the lasso.

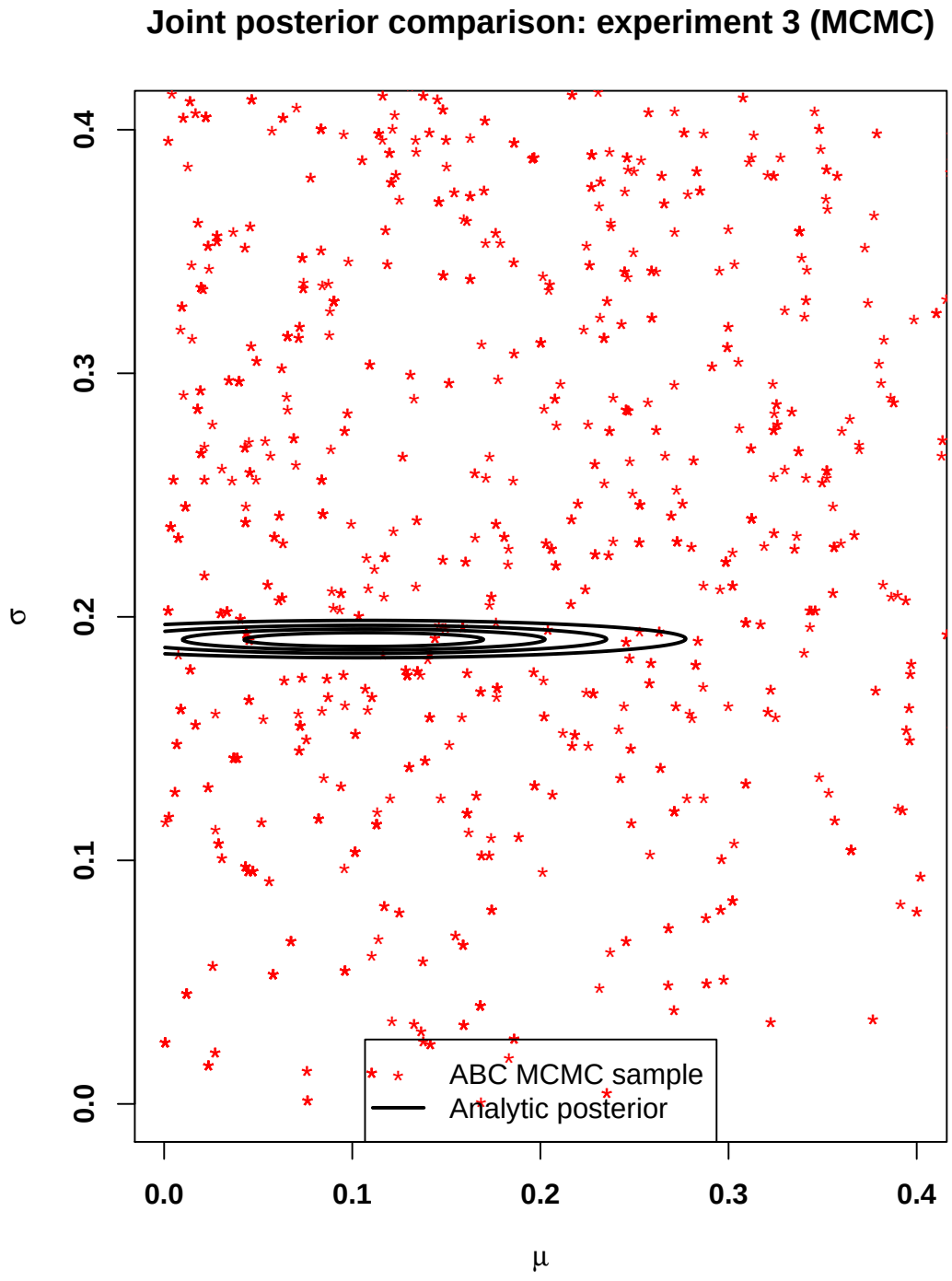

Figure 5.13: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with semi-automatic summary statistics derived using the lasso.

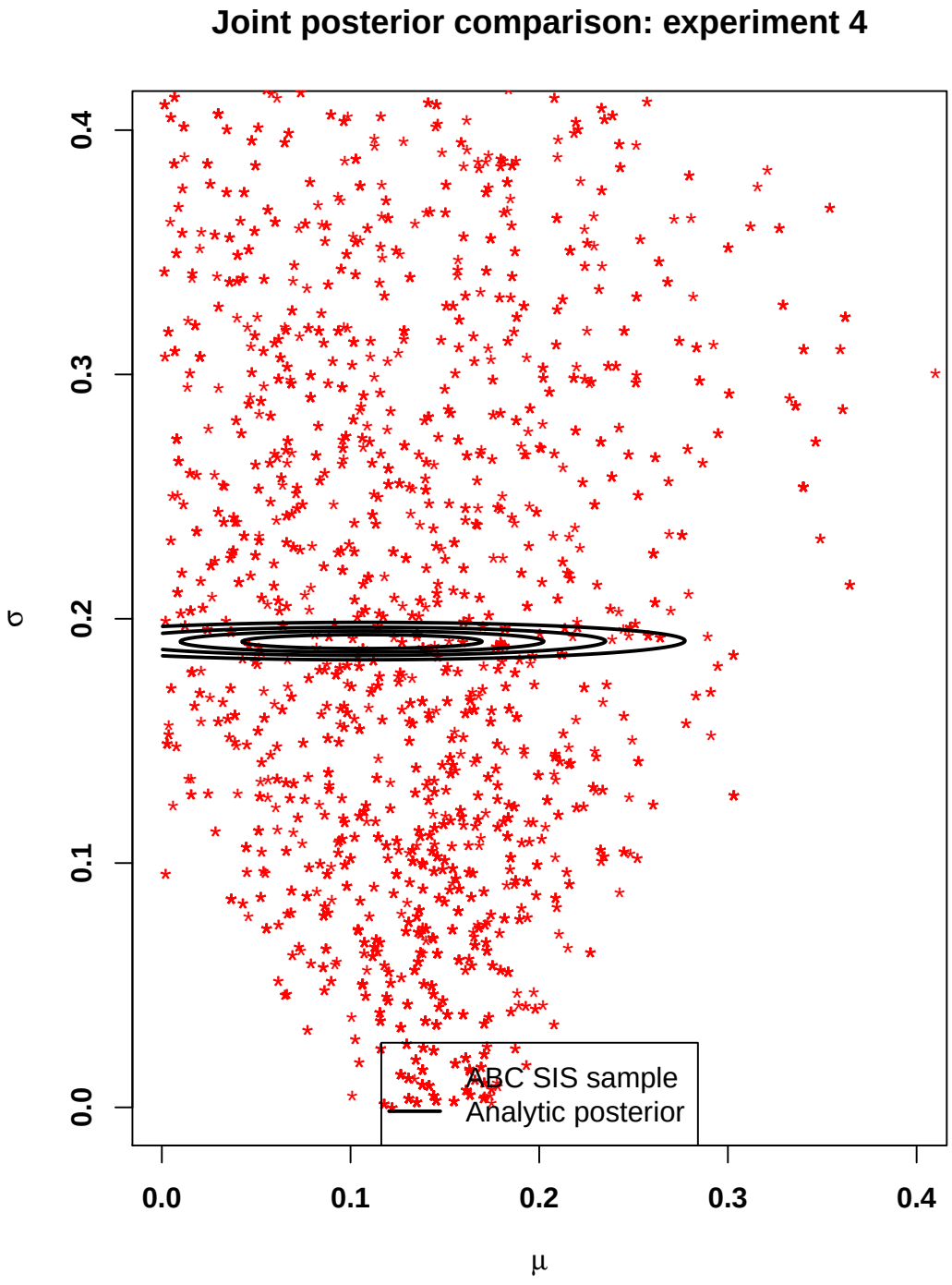

Figure 5.14: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with semi-automatic summary statistics derived by linear regression using EM based summary statistics as explanatory variables.

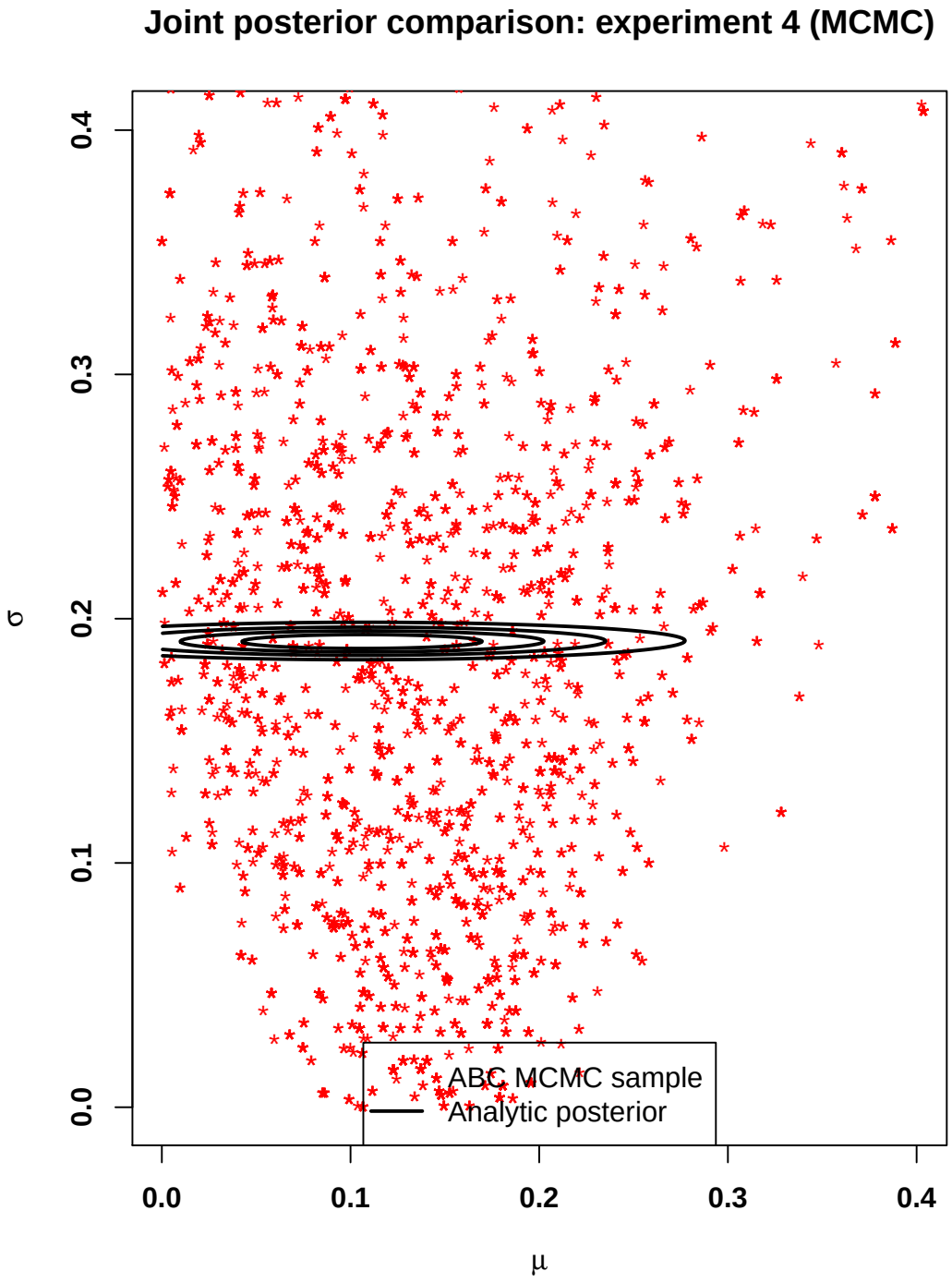

Figure 5.15: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with semi-automatic summary statistics derived by linear regression using EM based summary statistics as explanatory variables.

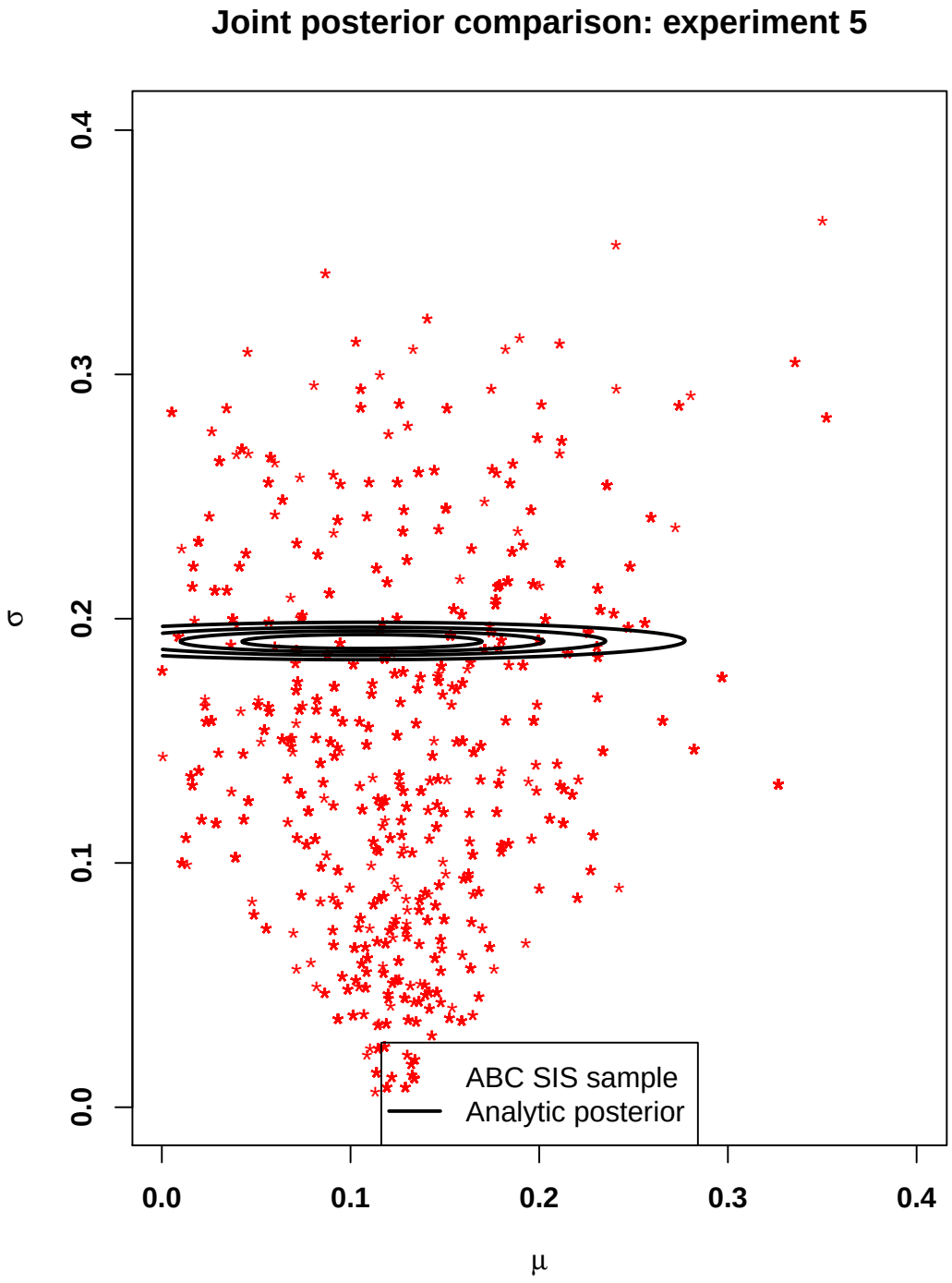

Figure 5.16: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with EM based summary statistics.

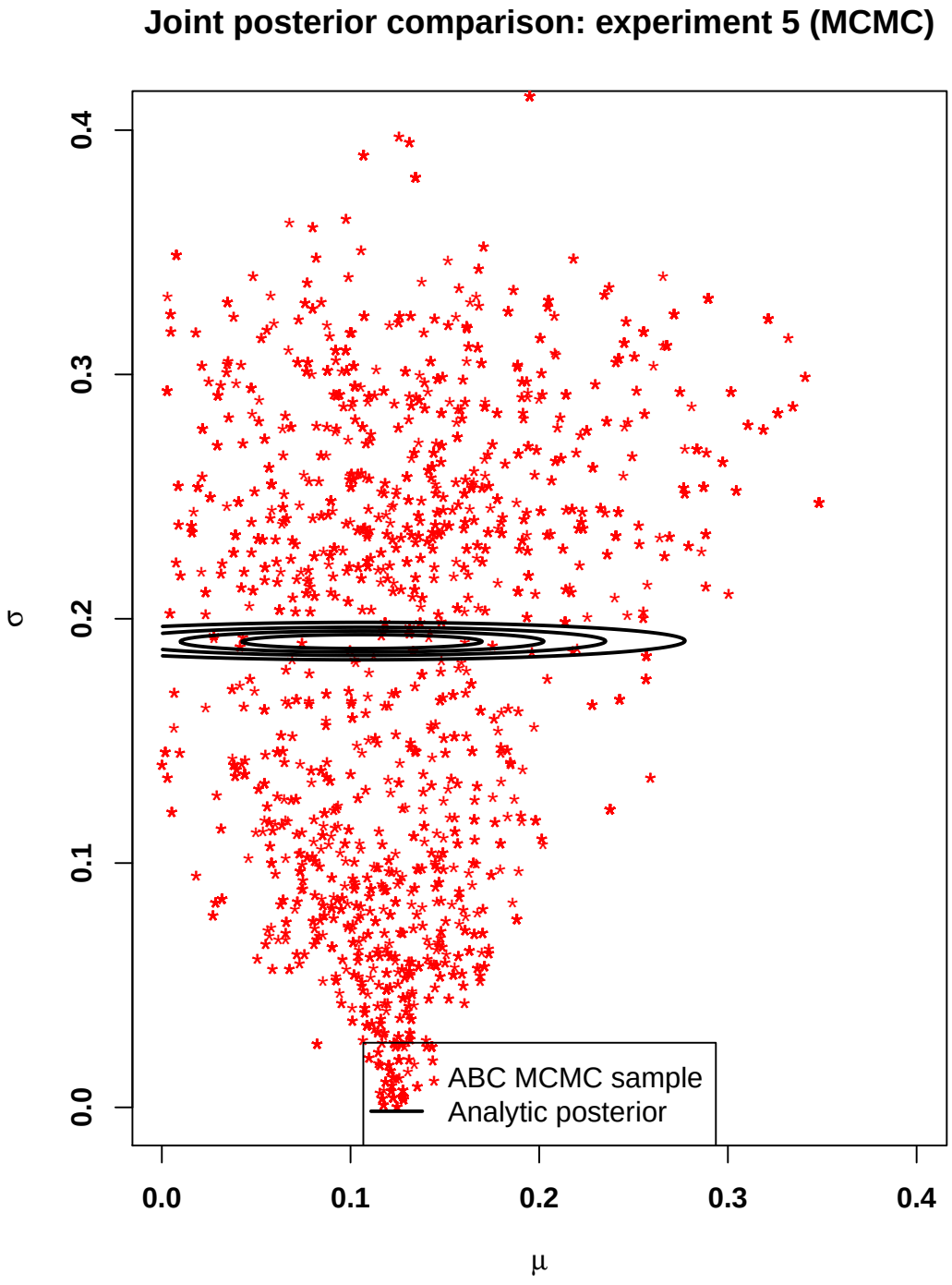

Figure 5.17: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with EM based summary statistics.

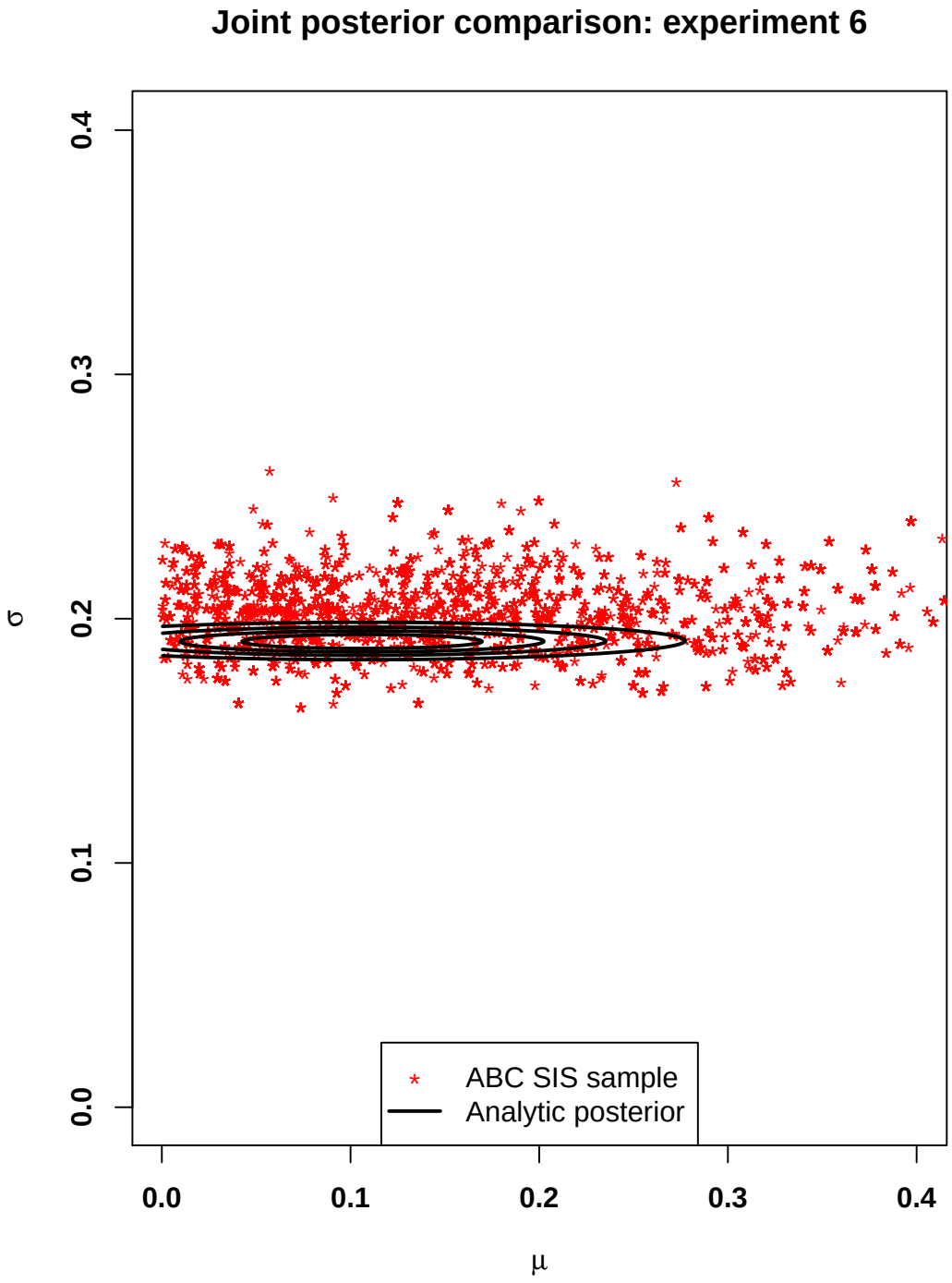

Figure 5.18: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with semi-automatic (regression) summary statistic for the drift parameter and a MA based summary statistic for the diffusion coefficient.

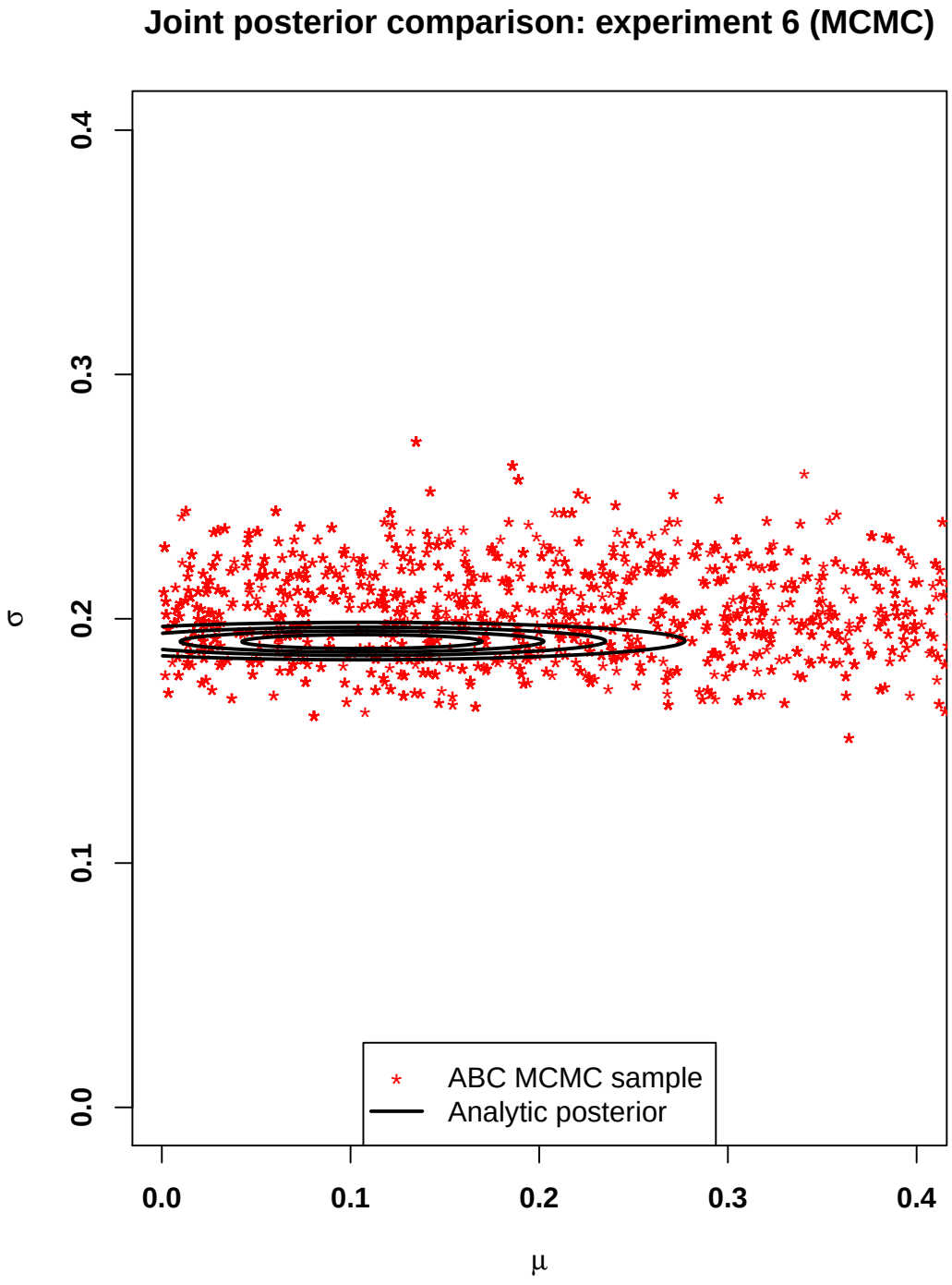

Figure 5.19: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with semi-automatic (regression) summary statistic for the drift parameter and a MA based summary statistic for the diffusion coefficient.

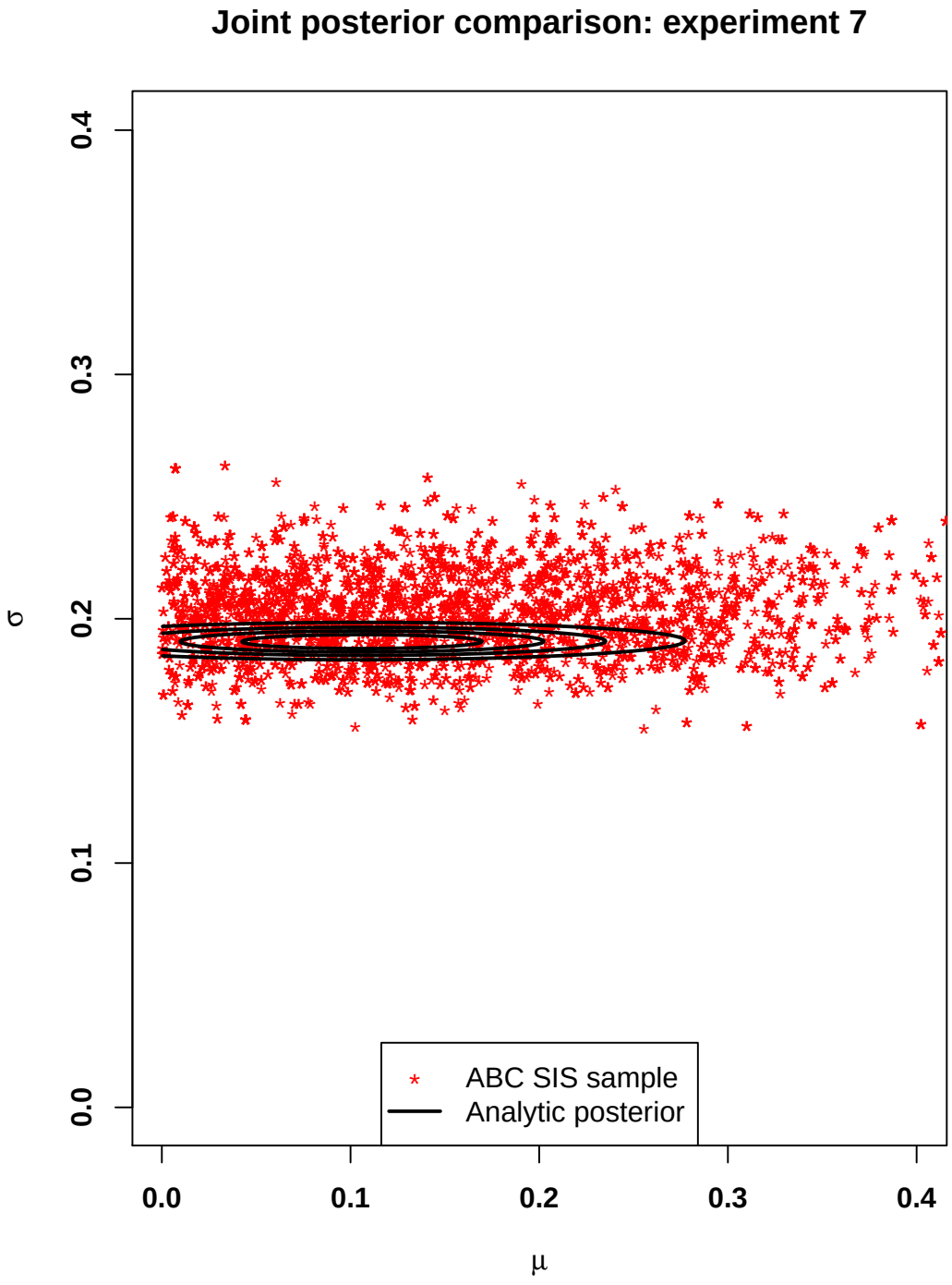

Figure 5.20: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with semi-automatic (lasso) summary statistic for the drift parameter and a MA based summary statistic for the diffusion coefficient.

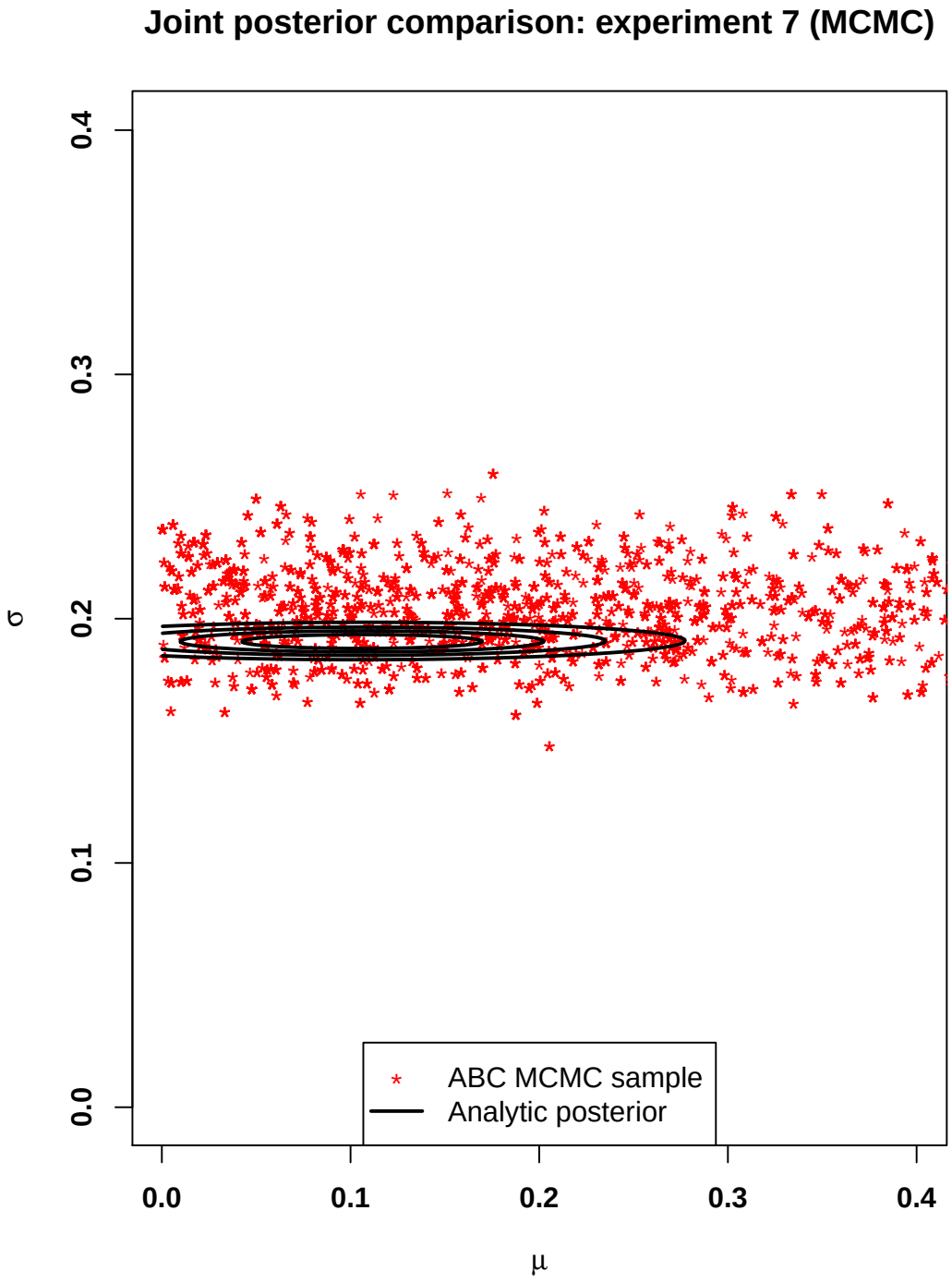

Figure 5.21: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with semi-automatic (lasso) summary statistic for the drift parameter and a MA based summary statistic for the diffusion coefficient.

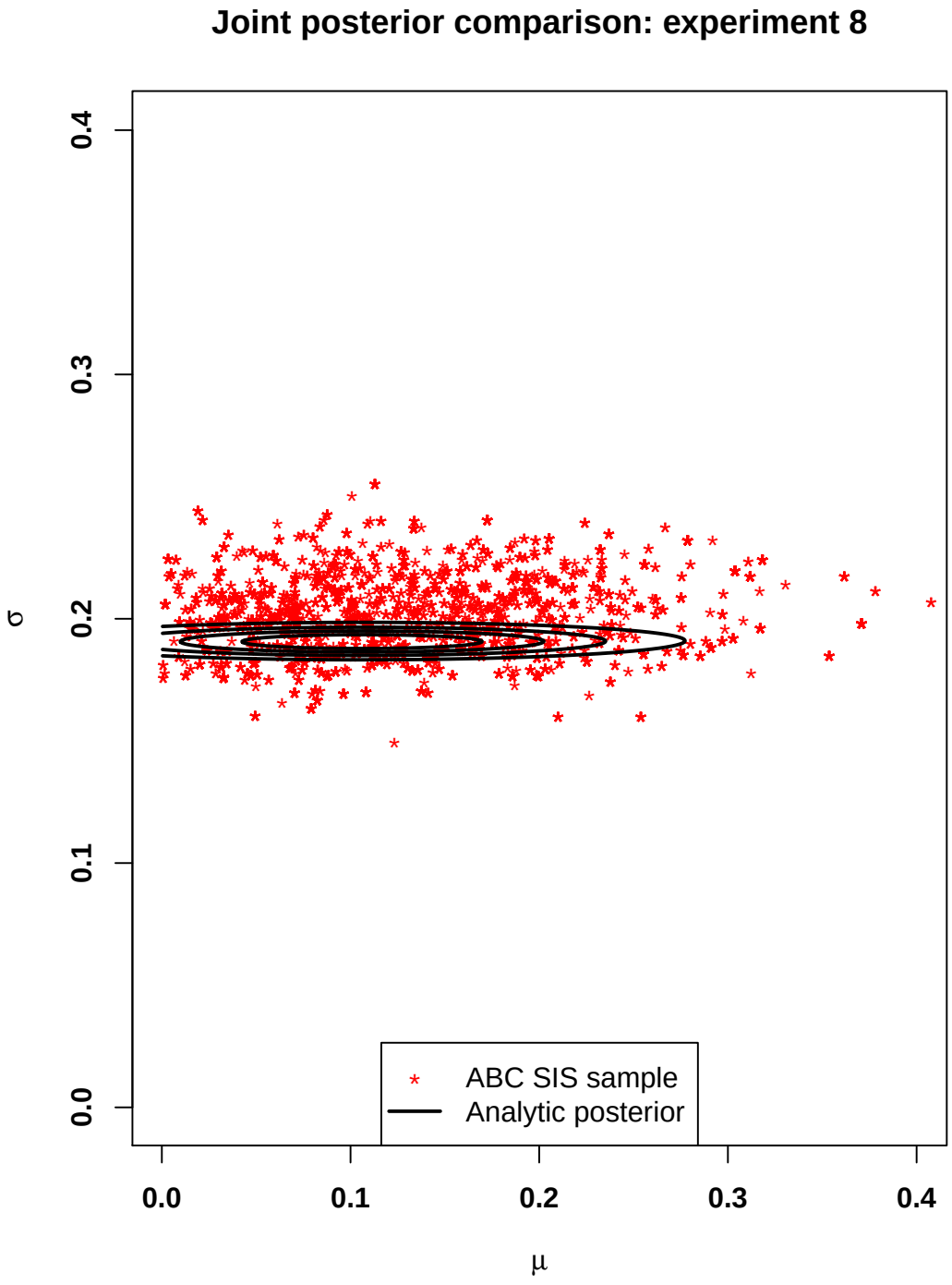

Figure 5.22: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with semi-automatic (regression) summary statistic, with EM based summary statistics being used as explanatory variables, for the drift parameter and a MA based summary statistic for the diffusion coefficient.

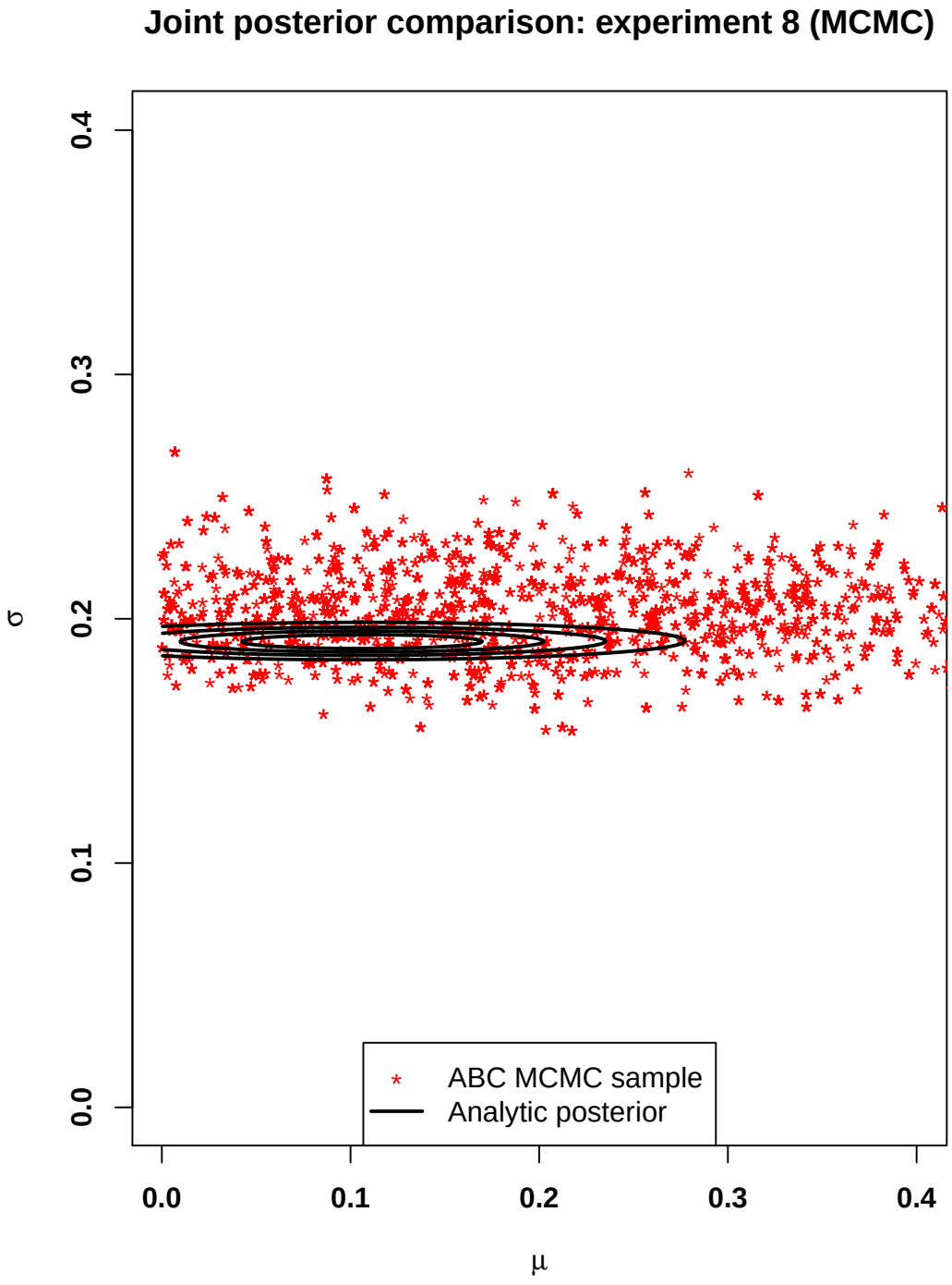

Figure 5.23: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with semi-automatic (regression) summary statistic, with EM based summary statistics being used as explanatory variables, for the drift parameter and a MA based summary statistic for the diffusion coefficient.

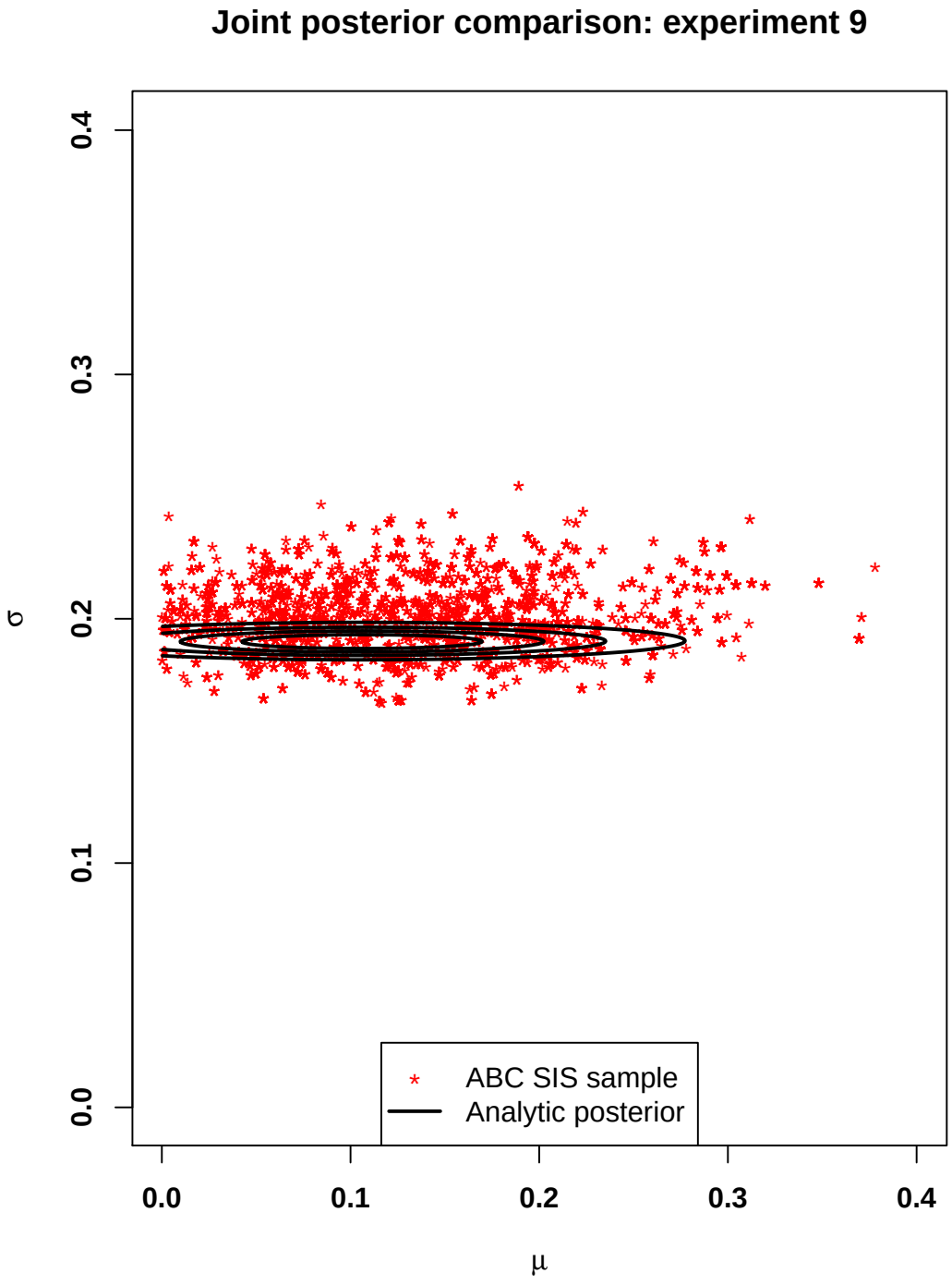

Figure 5.24: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS with an EM based summary statistic for the drift parameter and a MA based summary statistic for the diffusion coefficient.

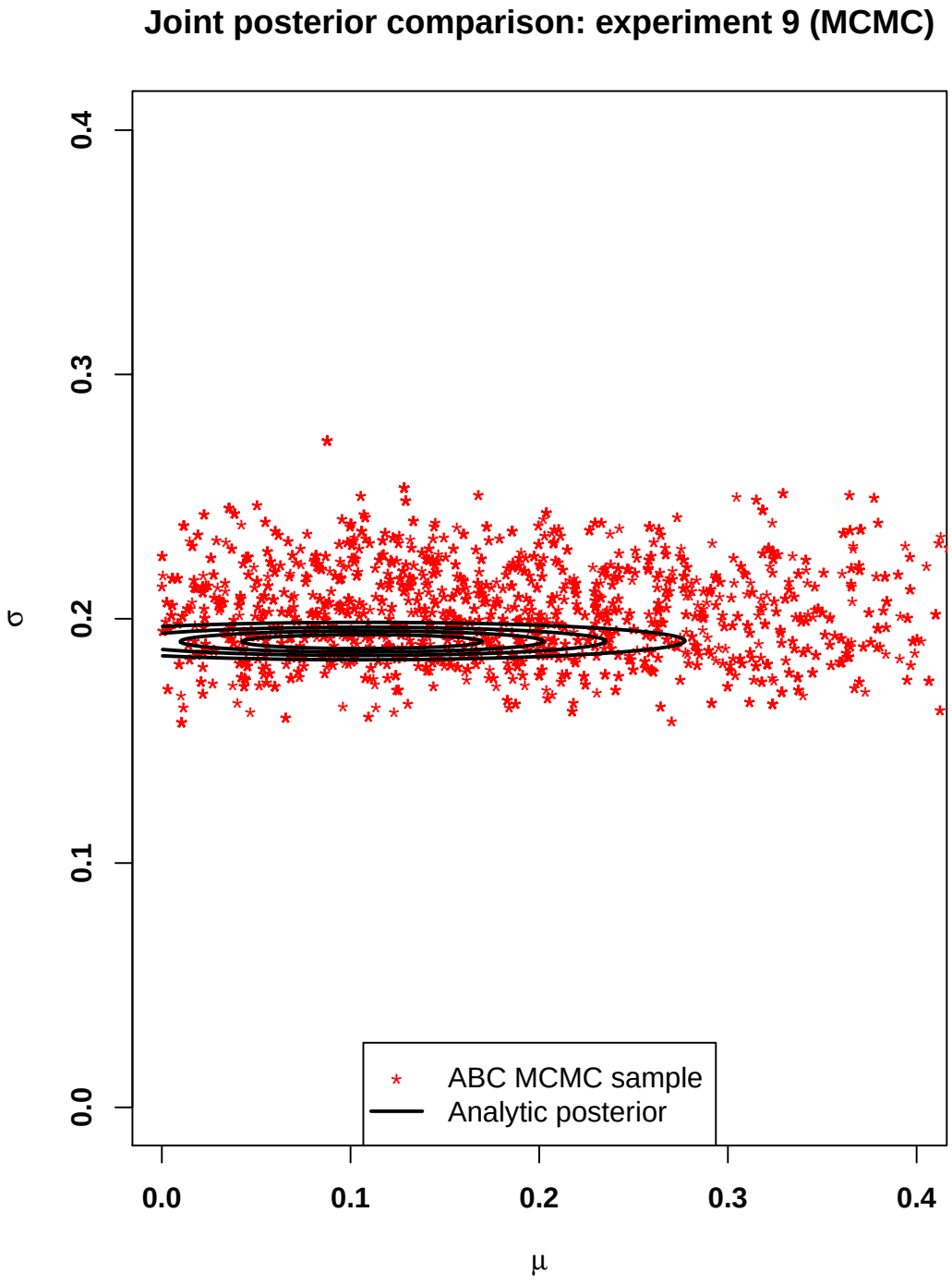

Figure 5.25: Contour plot of the analytic posterior (solid black line) overlaid with the empirical samples from the posterior derived via ABC MCMC with an EM based summary statistic for the drift parameter and a MA based summary statistic for the diffusion coefficient.

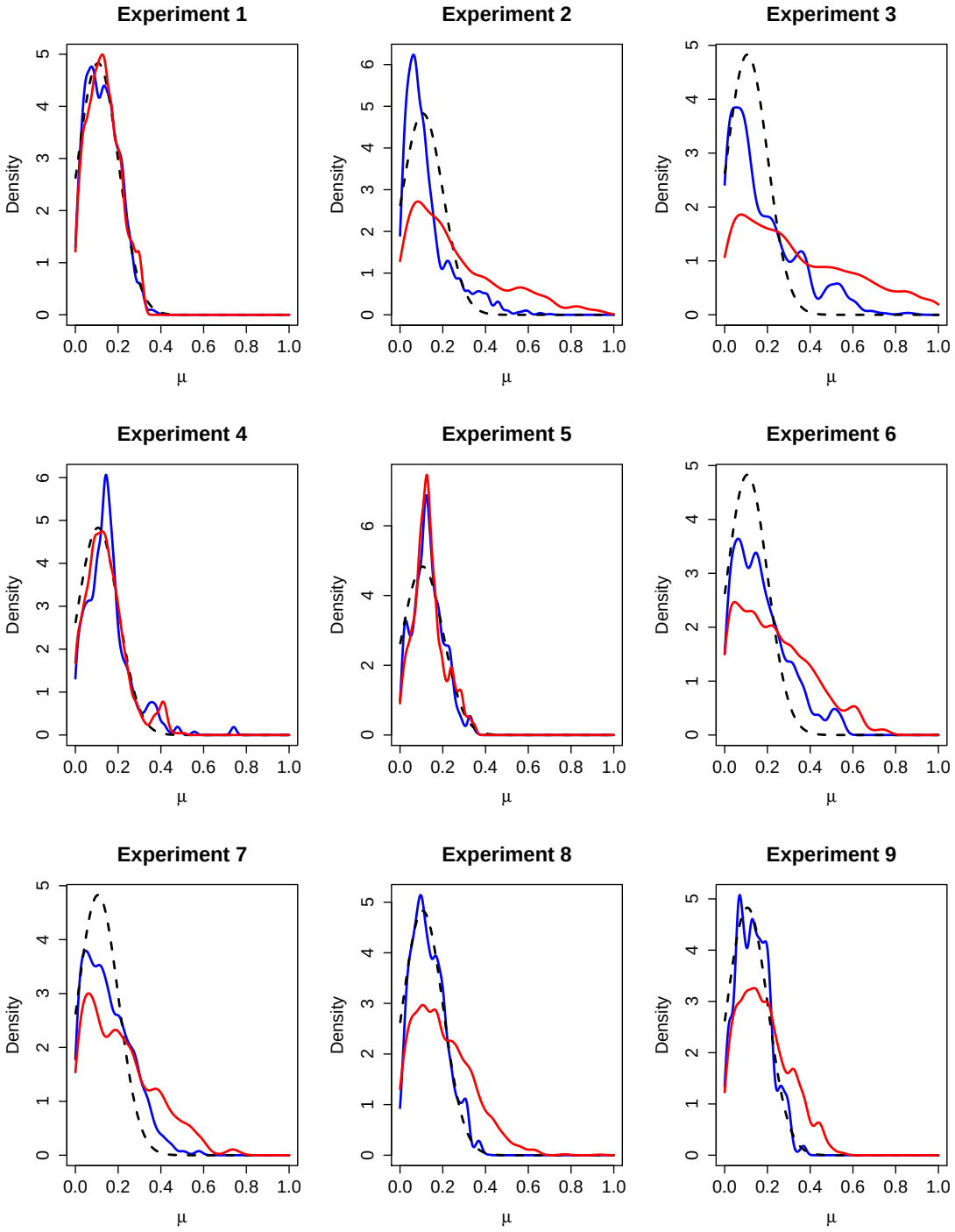

Figure 5.26: Plots of the marginal drift densities derived via Tempered ABC SIS (blue) and ABC MCMC (red) and the analytic marginal posterior (black dotted line).

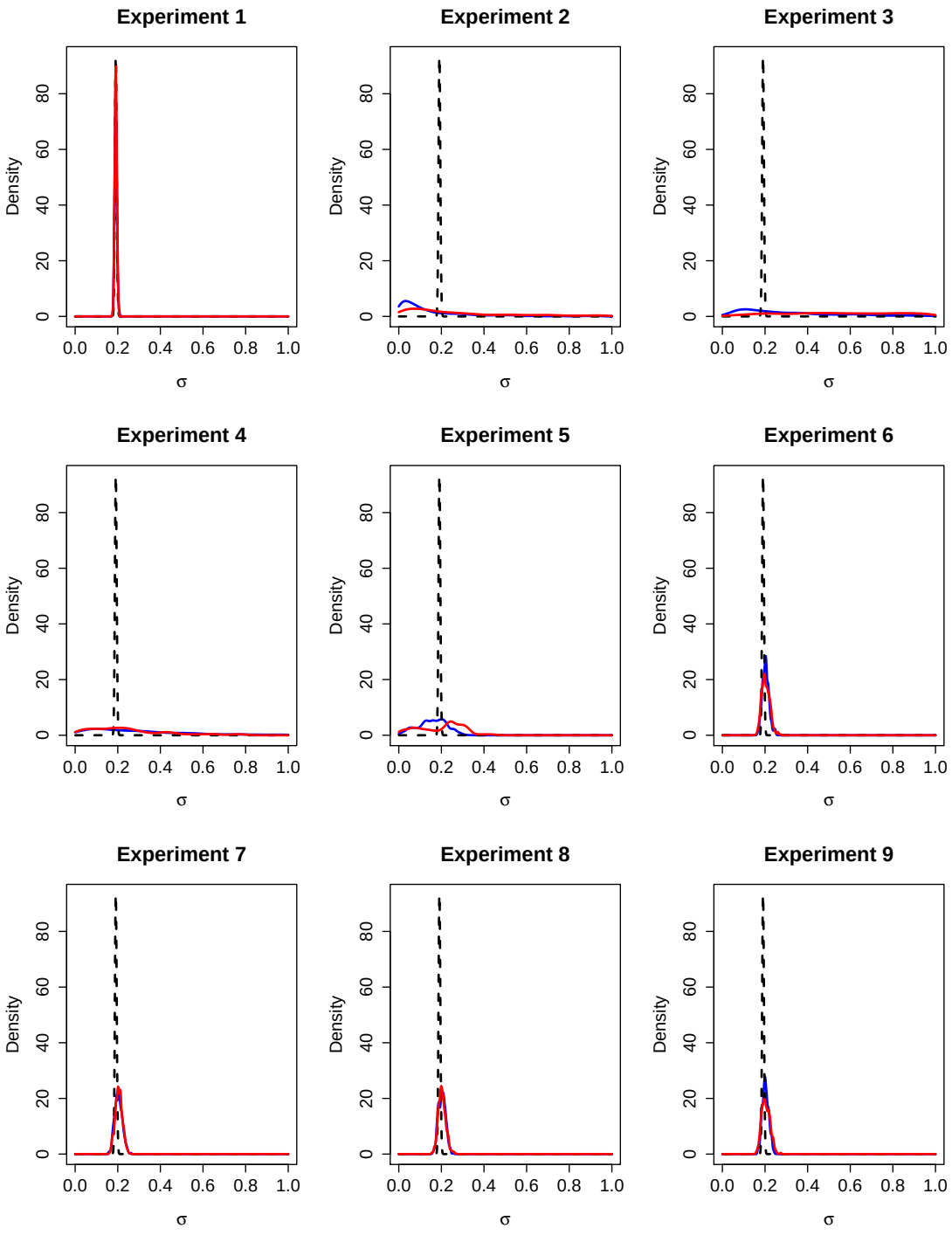

Figure 5.27: Plots of the marginal diffusion densities derived via Tempered ABC SIS (blue) and ABC MCMC (red) and the analytic marginal posterior (black dotted line).

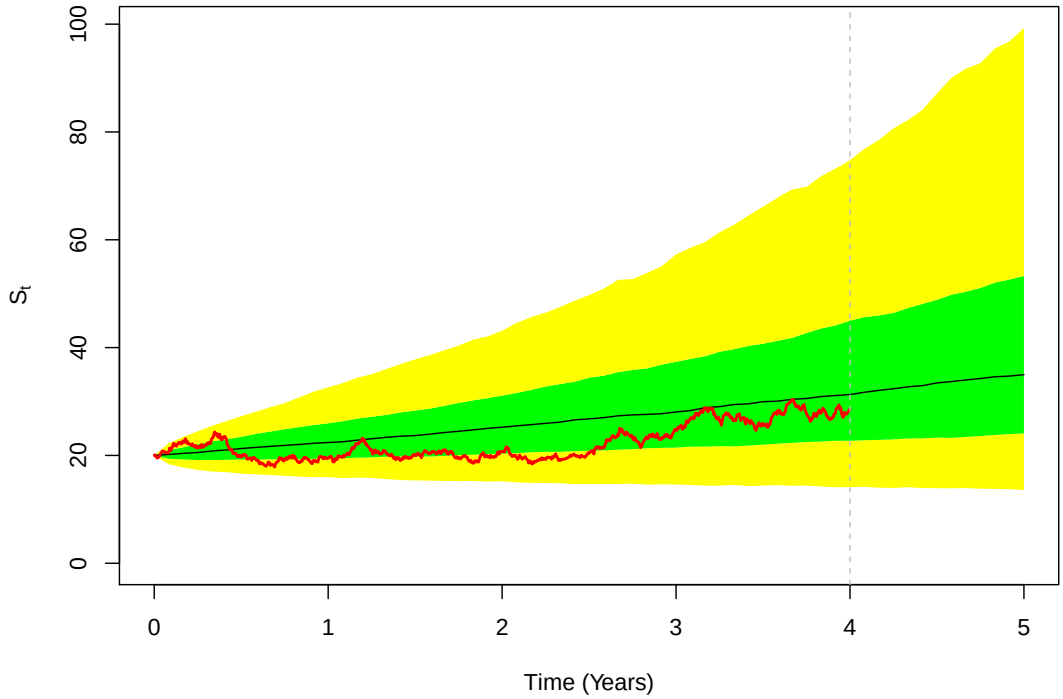

Figure 5.28: Posterior predictive density of five years' worth of new data, conditional on the observations used for inference. The black line represents the median of the PPD, the green area represents the inter-quartile range, and the yellow regions represent the ranges between the 5th and 25th, and the 75th and 95th percentiles. The red line represents the observations that were used to generate our samples from the posterior distribution of parameters via Temepered ABC SIS and ABC MCMC.

5.5.2 CIR model

In this section we present the results of 28 different numerical experiments: 14 different combinations of summary statistics, given in Table 5.2, were used to derive samples from the ABC approximated posterior using the Tempered ABC SIS and the adapted ABC MCMC samplers that were introduced in the previous chapter. We label each experiment using the same format that was used to label the GBM model experiments, e.g. *experiment 3* refers to the results associated with the

Tempered ABC SIS sampler, using a semi-automatic summary statistic for the mean reversion rate, the sample mean statistic for the mean reversion level, and the moving average based statistic for the diffusion parameter. We first present the two dimensional marginal posterior samples obtained from our ABC experiments, overlaid with contours of the analytic two dimensional posterior distribution, in order to assess the quality of the sampling algorithms. Due to the large number of marginal posterior plots (84 in total: 3 two dimensional marginal plots per experiment and 28 different experiments), we will present the marginal posterior plots for the best results obtained via the Tempered ABC SIS and adapted ABC MCMC samplers, and include the remaining plots in Appendix 5.8 at the end of this chapter.

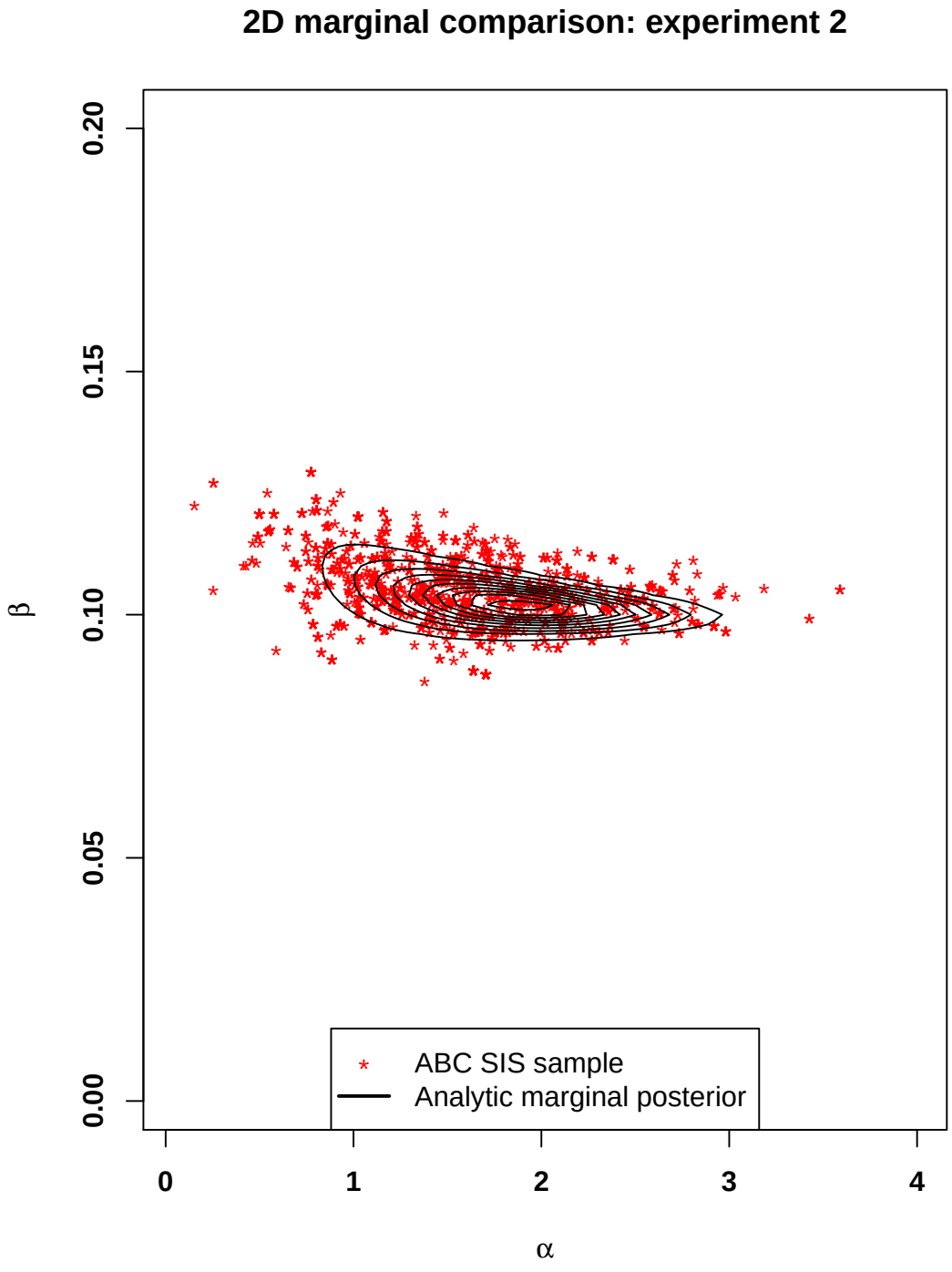

Figure 5.29: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS (red data points) for the mean reversion rate and mean reversion level parameters. The mean-gradient, sample mean, and MA statistics were used for the mean reversion rate, mean reversion level, and diffusion parameters respectively.

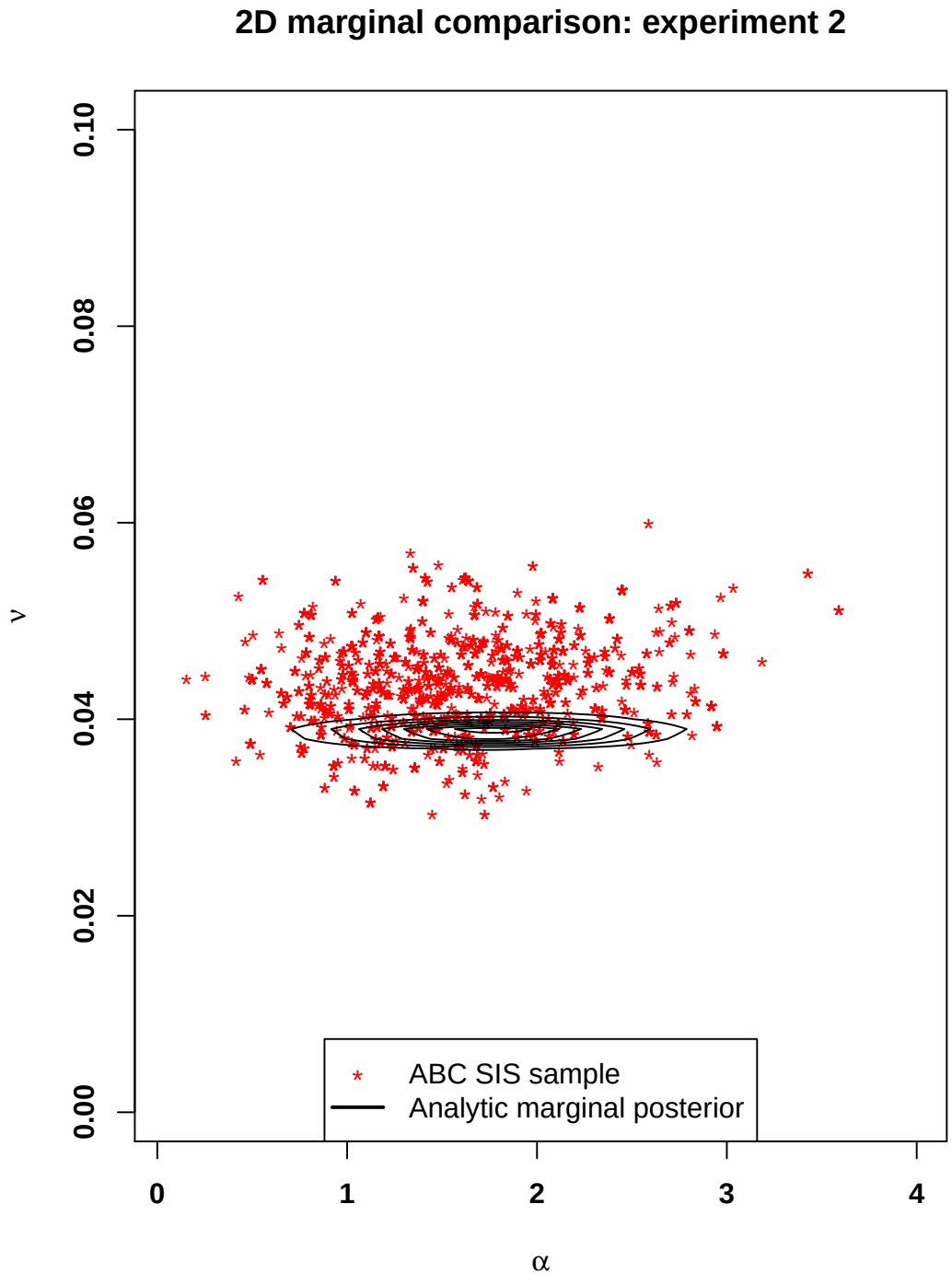

Figure 5.30: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS (red data points) for the mean reversion rate and diffusion parameters. The mean-gradient, sample mean, and MA statistics were used for the mean reversion rate, mean reversion level, and diffusion parameters respectively.

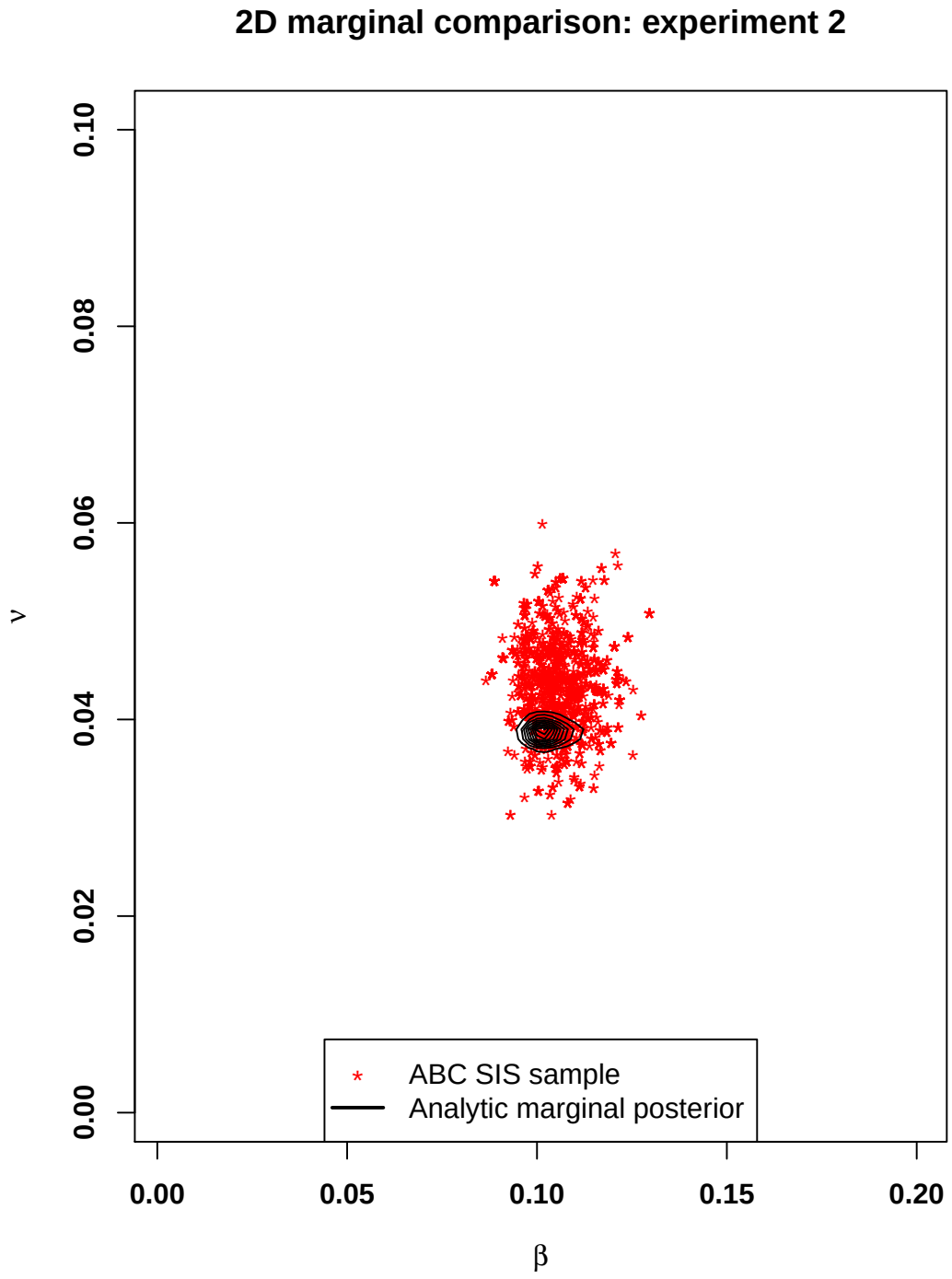

Figure 5.31: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS (red data points) for the mean reversion level and diffusion parameters. The mean-gradient, sample mean, and MA statistics were used for the mean reversion rate, mean reversion level, and diffusion parameters respectively.

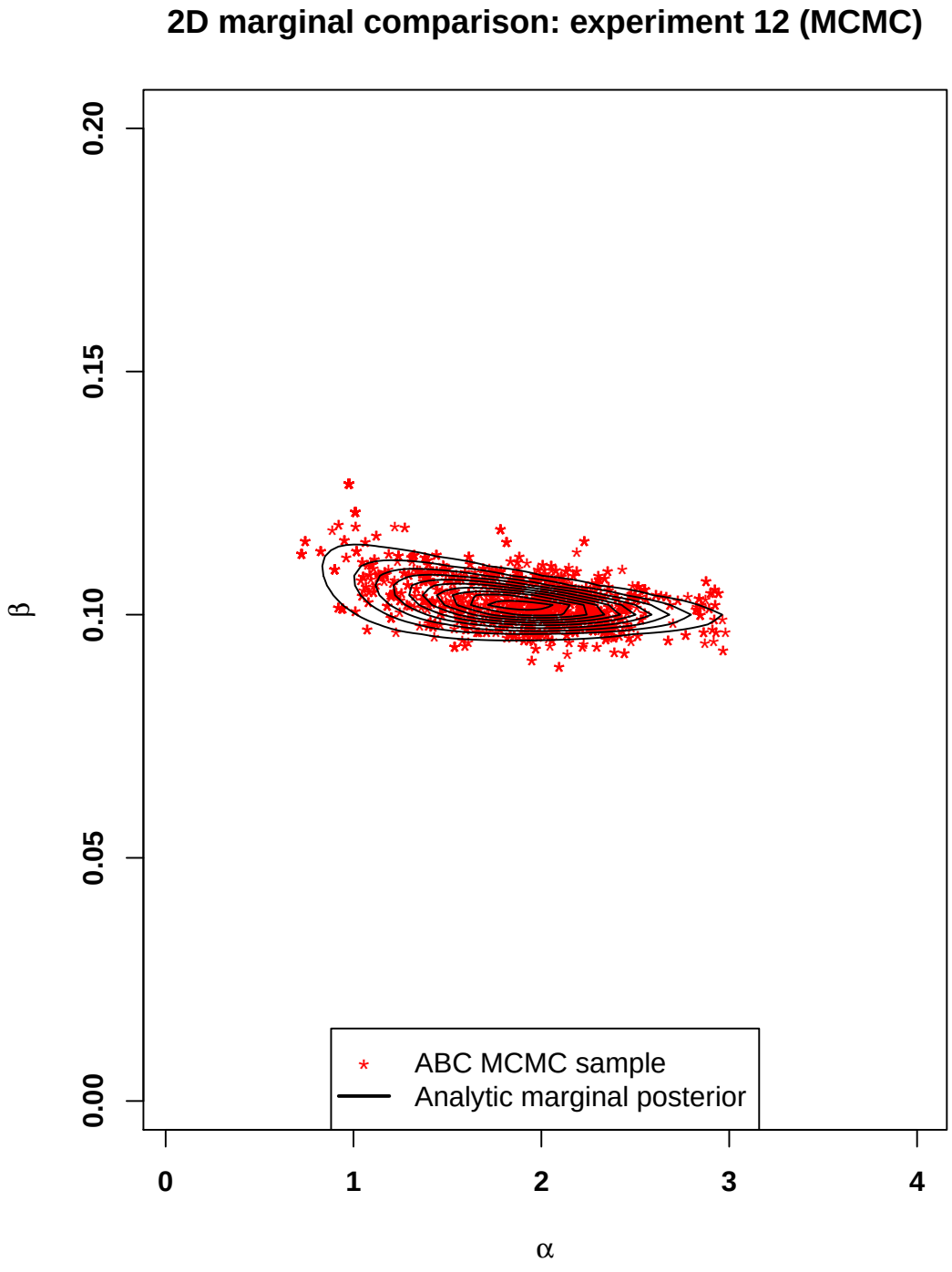

Figure 5.32: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via adapted ABC MCMC (red data points) for the mean reversion rate and mean reversion level parameters. The OLS, semi-automatic (lasso), and MA statistics were used for the mean reversion rate, mean reversion level, and diffusion parameters respectively.

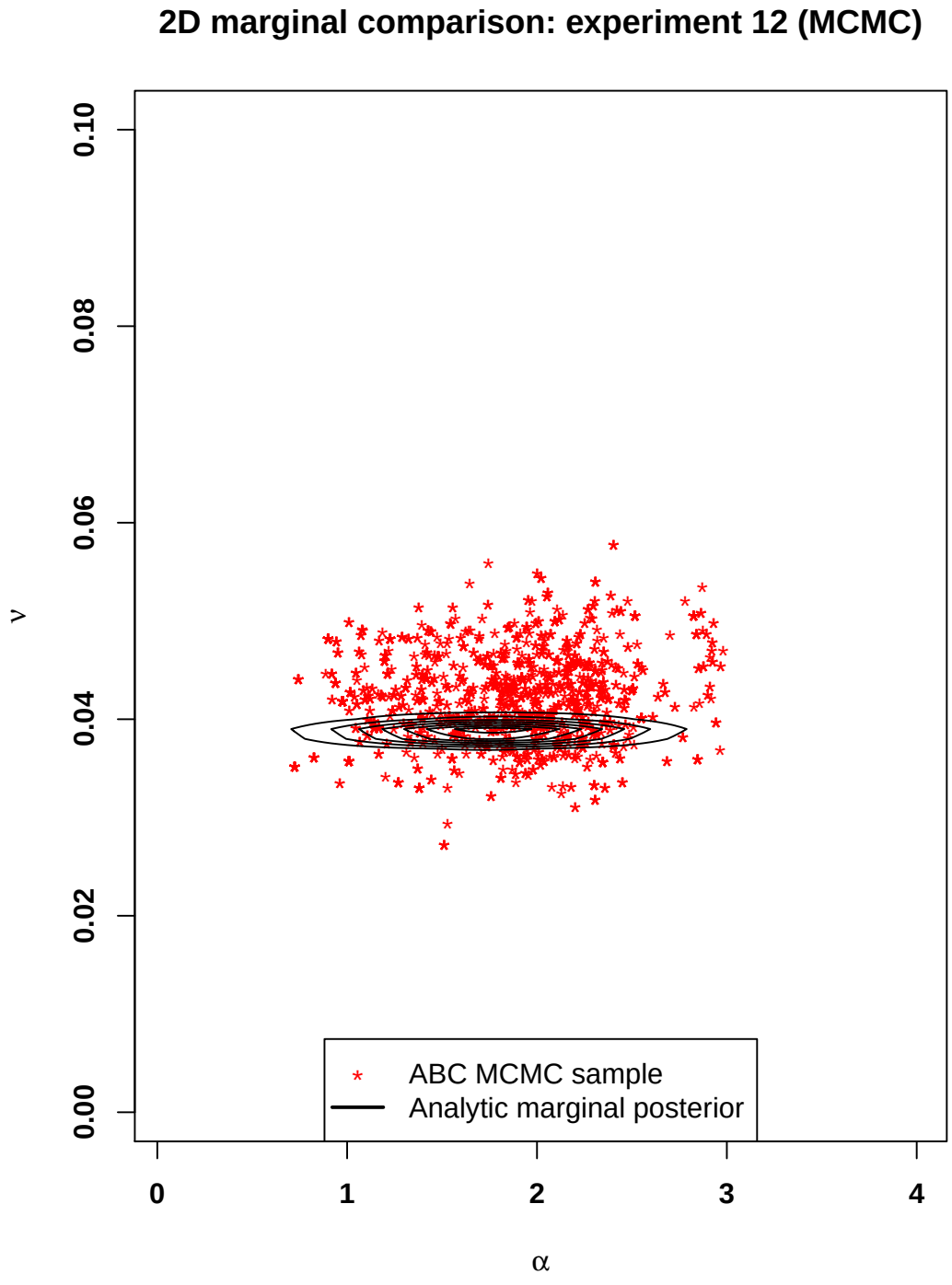

Figure 5.33: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via adapted ABC MCMC (red data points) for the mean reversion rate and diffusion parameters. The OLS, semi-automatic (lasso), and MA statistics were used for the mean reversion rate, mean reversion level, and diffusion parameters respectively.

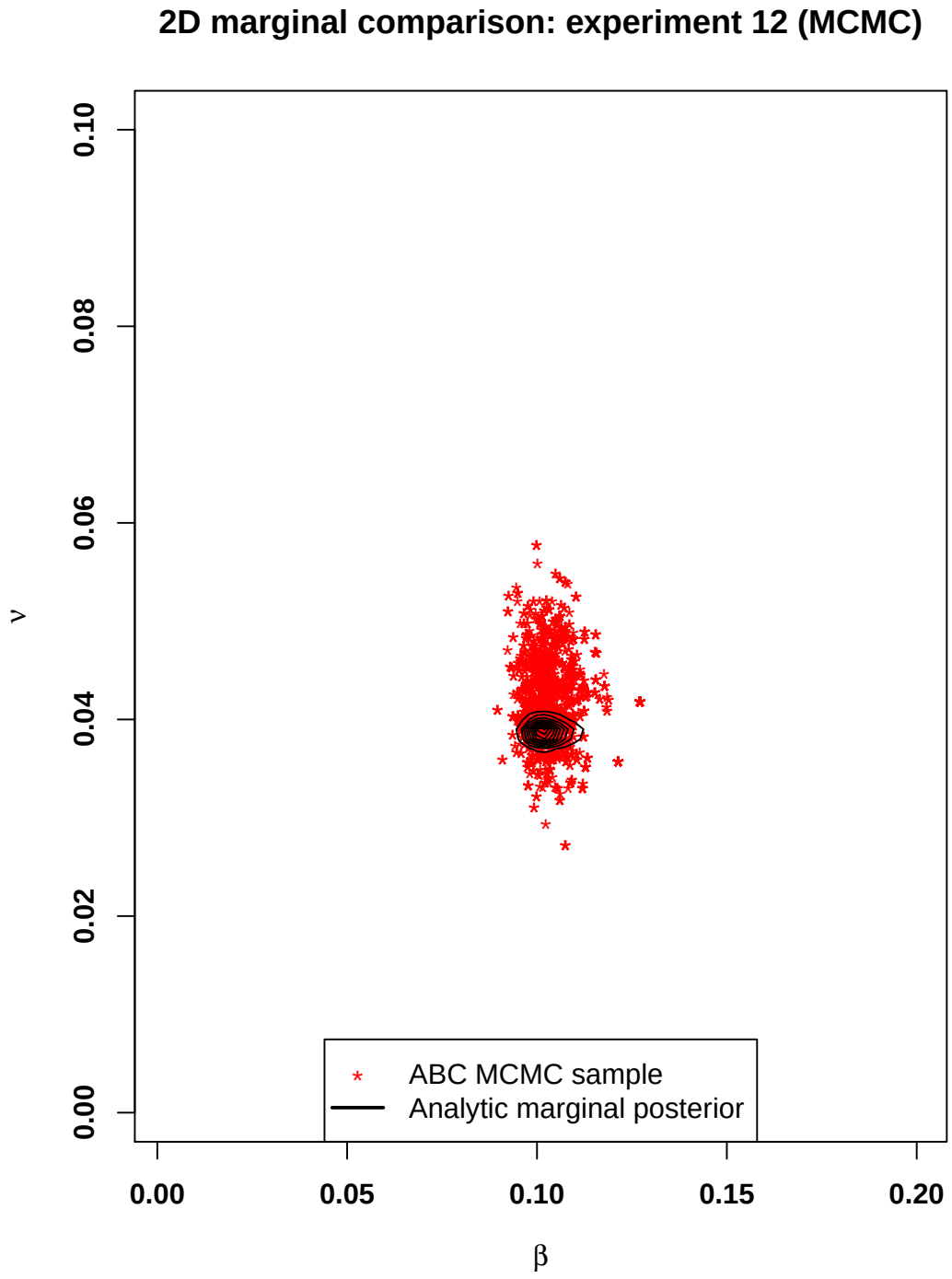

Figure 5.34: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via adapted ABC MCMC (red data points) for the mean reversion level and diffusion parameters. The OLS, semi-automatic (lasso), and MA statistics were used for the mean reversion rate, mean reversion level, and diffusion parameters respectively.

From Figure 5.29 one can see that the Tempered ABC SIS sampler was able to successfully recover the correct location and general shape of the marginal posterior distribution of the mean reversion rate and diffusion parameters. The empirical distribution is slightly more diffuse compared to the analytic posterior, which is to be expected given that our ABC approximation assumes additional Gaussian randomness in the data generating process (see equation 4.2). Similarly, Figure 5.30 illustrates that the Tempered ABC SIS sampler was able to recover the location and shape of the marginal posterior associated with the mean reversion rate and diffusion parameters. As was the case with the estimation of the diffusion parameter in the GBM model covered previously, the MA statistic appears to yield slightly biased samples from the posterior, with most of the data points lying marginally above the location of the mode of the true posterior. Figure 5.31 reinforces these conclusions—the sample points from the ABC approximated posterior are clustered around the true posterior, with a slight positive bias evident in the sampled diffusion parameter values.

Empirical parameter distributions obtained via adapted ABC MCMC sampling, illustrated in Figures 5.32, 5.33, and 5.34, are represent good quality approximations to the analytic posterior. Indeed, the marginal posterior in Figure 5.32 is more concentrated around the location of the analytic contours than the empirical sample obtained via Tempered ABC SIS sampling. The semi-automatic summary statistic derived via lasso regression also performed very well, producing an empricial sample that is highly concentrated around the analytic posterior. The reader is referred to the Appendix 5.8 for the remaining two dimensional marginal posterior plots.

We now present the one dimensional marginal posterior approximations for all 28 experiments. Experiments 1, 2, and 3 yielded the most promising results, with both the Tempered ABC SIS and adapted ABC MCMC samples recovering the

analytic posterior densities. Empirical results for the mean reversion level and diffusion parameters were very promising, suggesting that even in the absence of sufficient summary statistics, and despite the non-zero similarity kernel variance, ε, that was used in the ABC likelihood approximation, both ABC samplers can accurately recover the analytic marginal posterior densities for these parameters. Consonant with the results section relating to the GBM model, we now present the PPD plot for the CIR model. To reiterate: the PPD represents the density of newly simulated data from the model, conditional on the pre-existing model observations. PPD plots can be very useful when making predictions about the distribution of newly simulated data from the model, given the historical data that has already been observed, and explicitly take into account the uncertainty surrounding the 'true' parameter values. In the context of quantitative risk management in finance, it is extremely important to capture different types of risk when making predictions about the future evolution of key state variables; failing to take into account parameter risk when generating Monte Carlo scenarios could lead to inadequate capital being set aside to absorb future financial shocks, the consequences of which could be very serious.

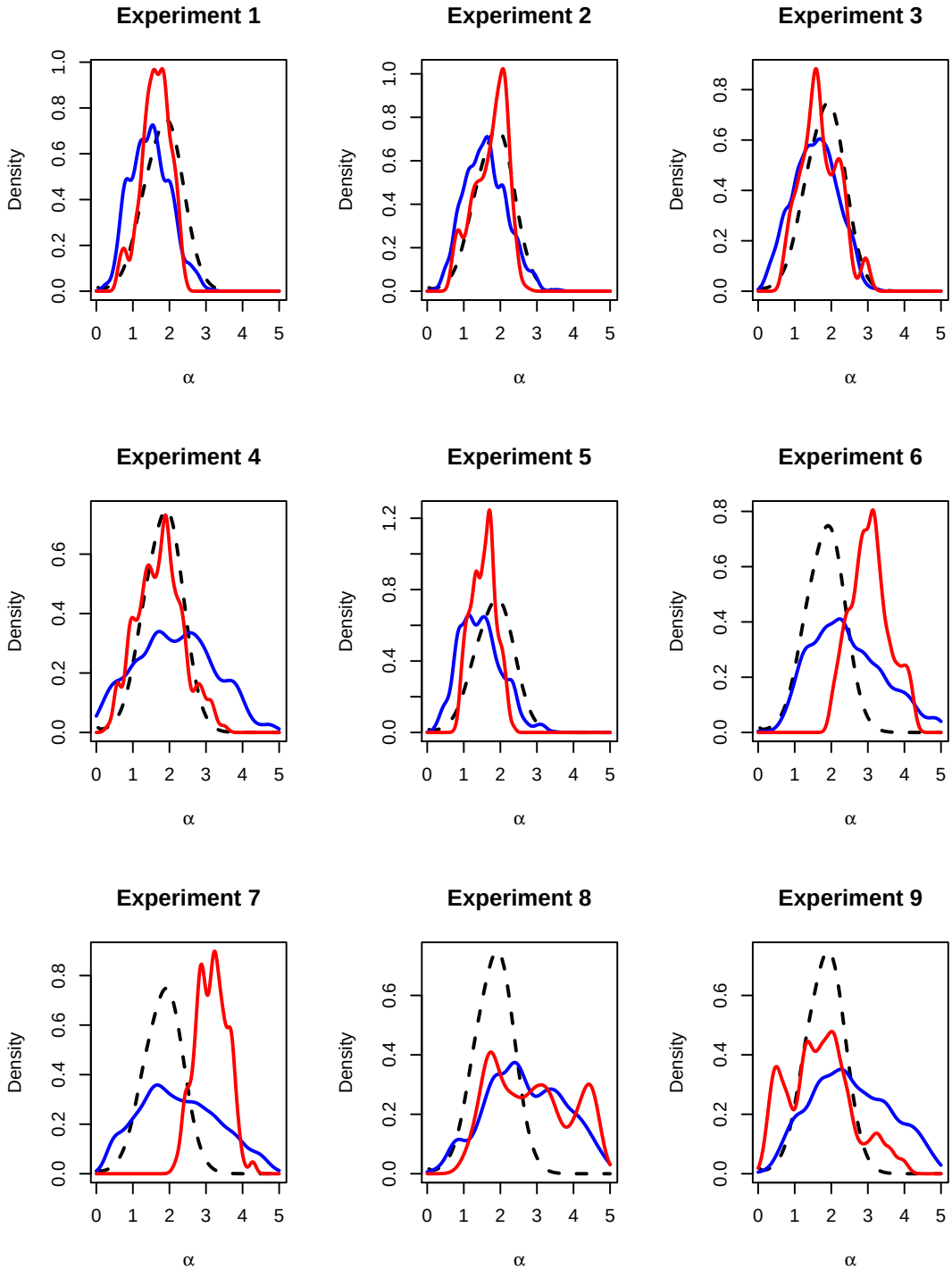

Figure 5.35: Plots of the marginal mean reversion rate posterior densities derived via Tempered ABC SIS (blue) and ABC MCMC (red), and the analytic marginal posterior (black dotted line).

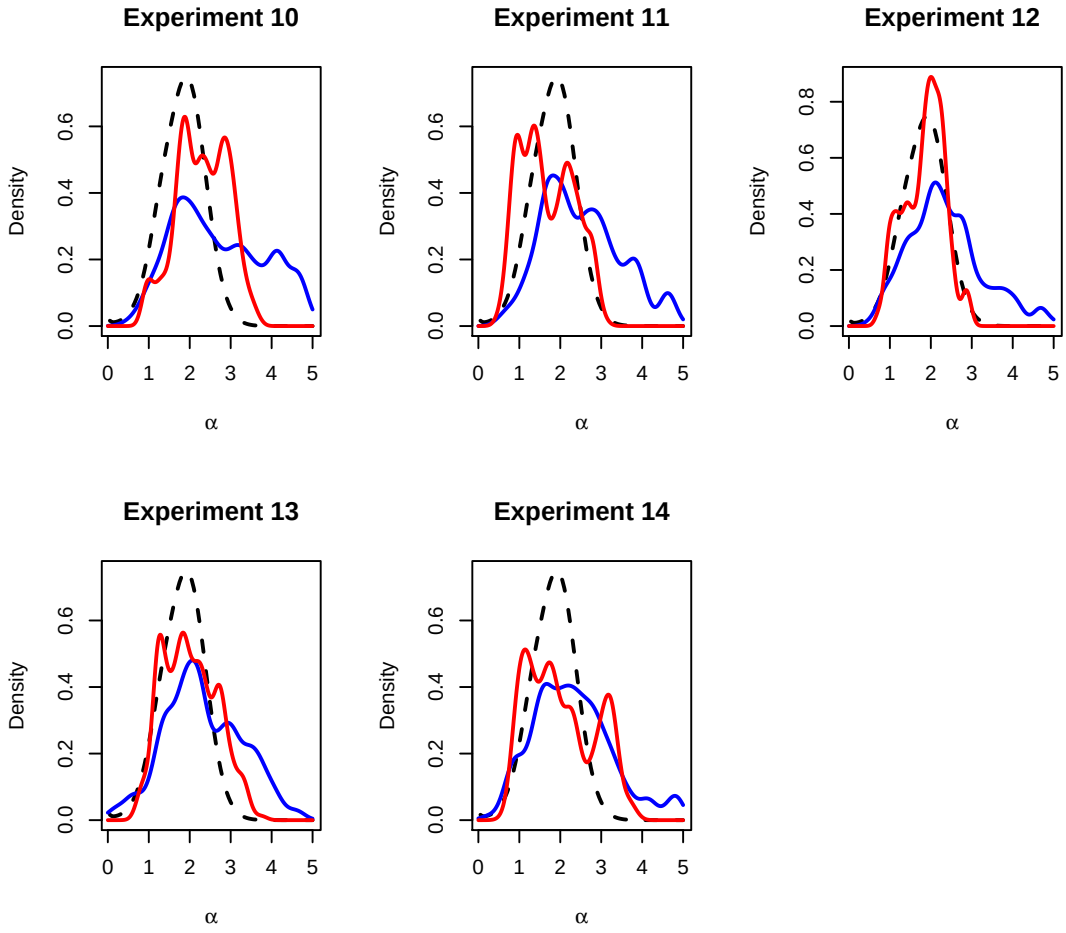

Figure 5.36: Plots of the marginal mean reversion level posterior densities derived via Tempered ABC SIS (blue) and ABC MCMC (red), and the analytic marginal posterior (black dotted line).

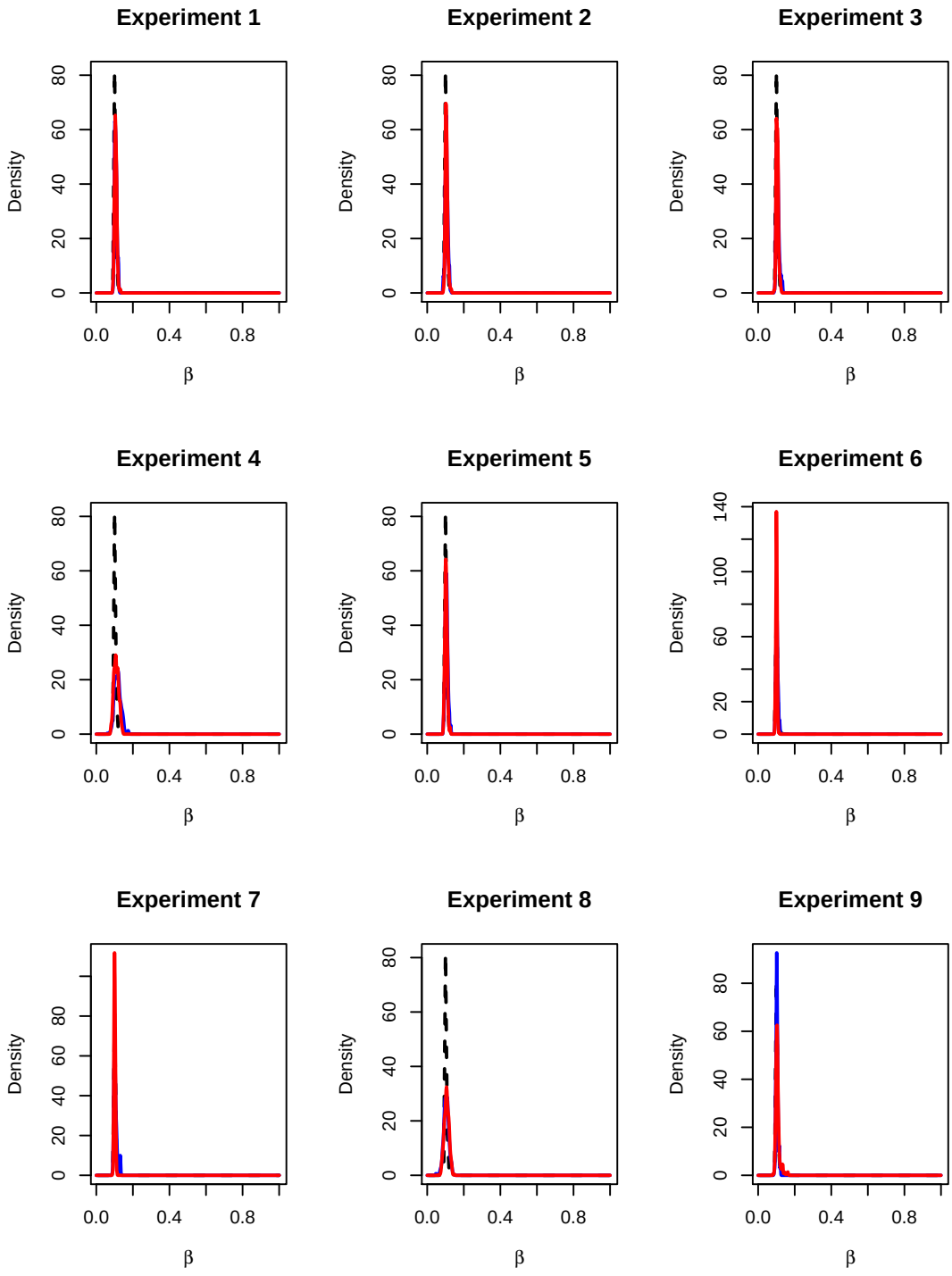

Figure 5.37: Plots of the marginal diffusion posterior densities derived via Tempered ABC SIS (blue) and ABC MCMC (red), and the analytic marginal posterior (black dotted line).

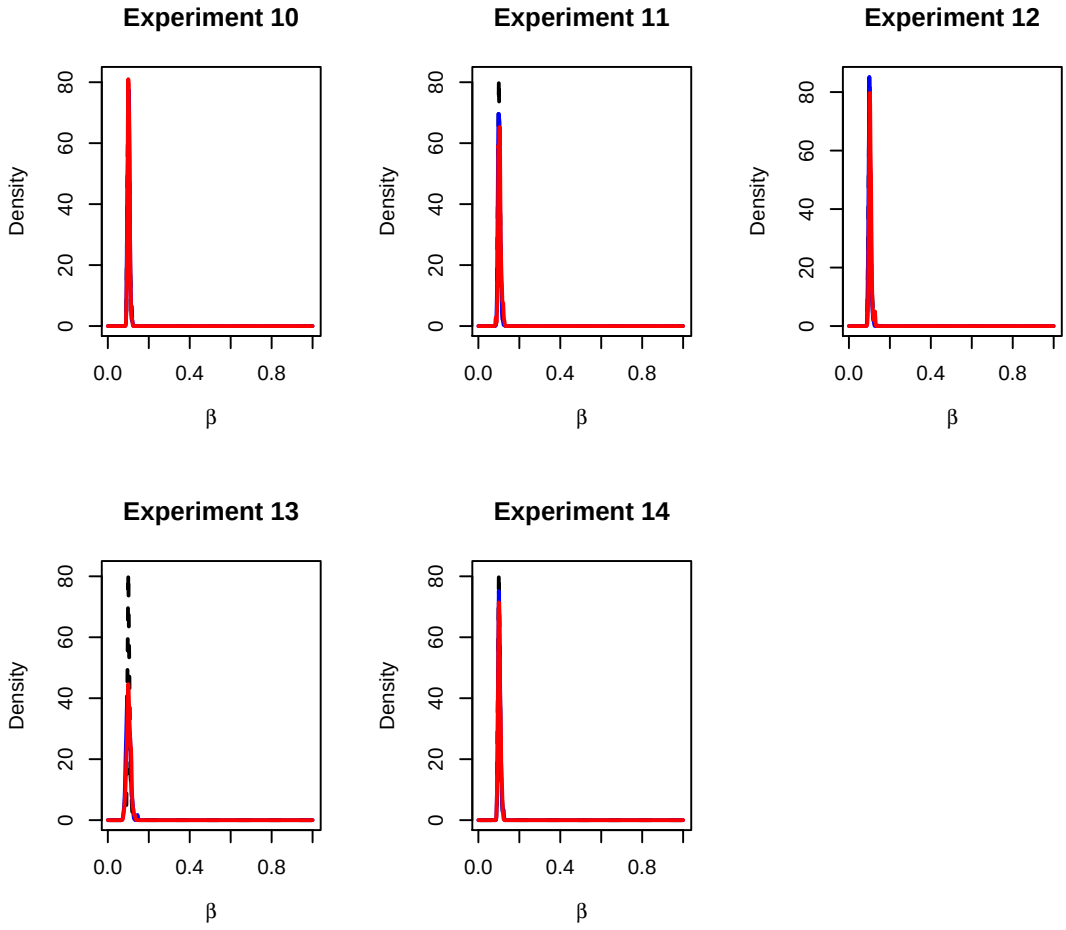

Figure 5.38: Plots of the marginal densities of the mean reversion rate parameter, derived via Tempered ABC SIS (blue) and ABC MCMC (red) and the analytic marginal posterior (black dotted line).

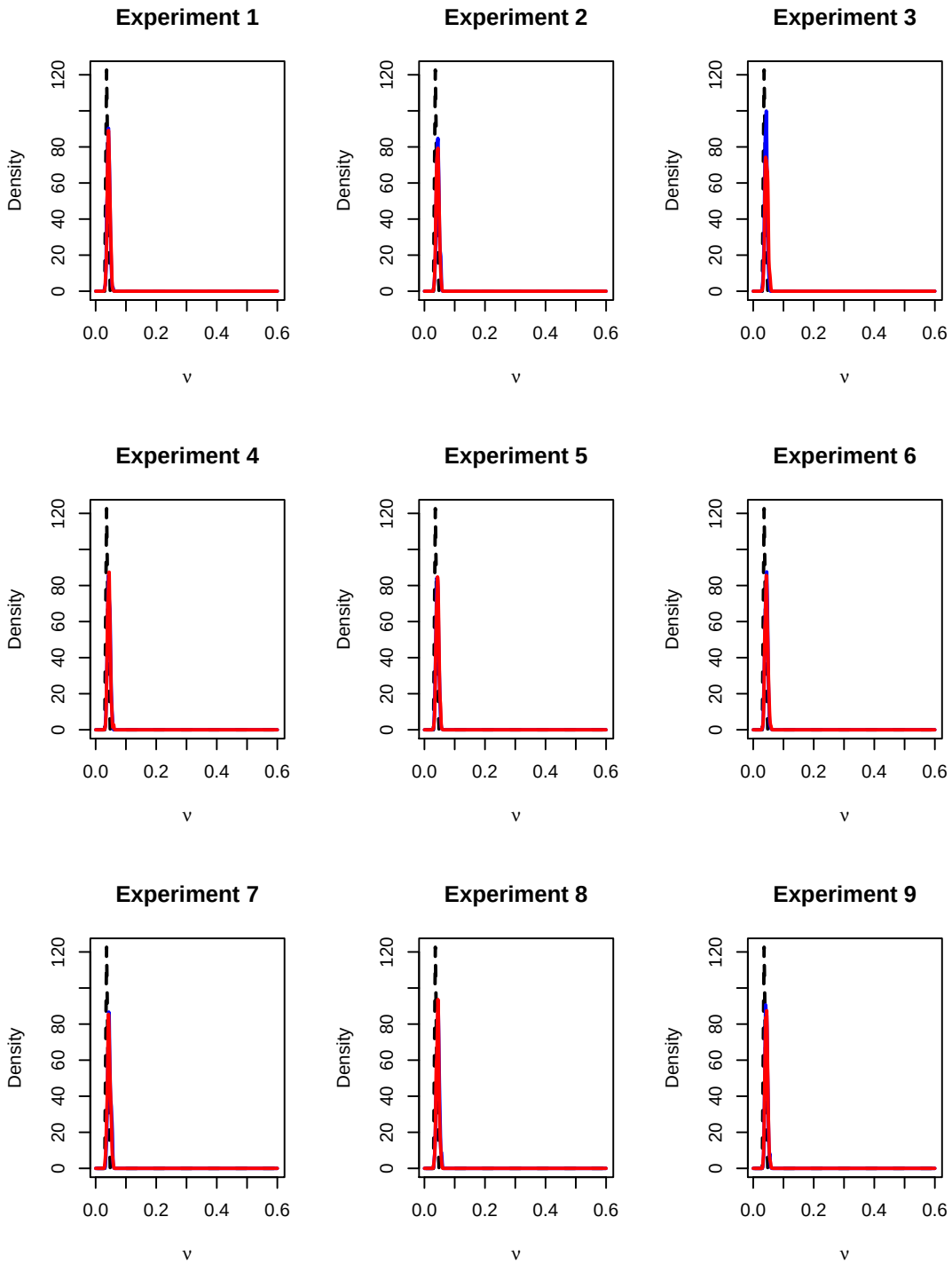

Figure 5.39: Plots of the marginal densities of the mean reversion level parameter, derived via Tempered ABC SIS (blue) and ABC MCMC (red) and the analytic marginal posterior (black dotted line).

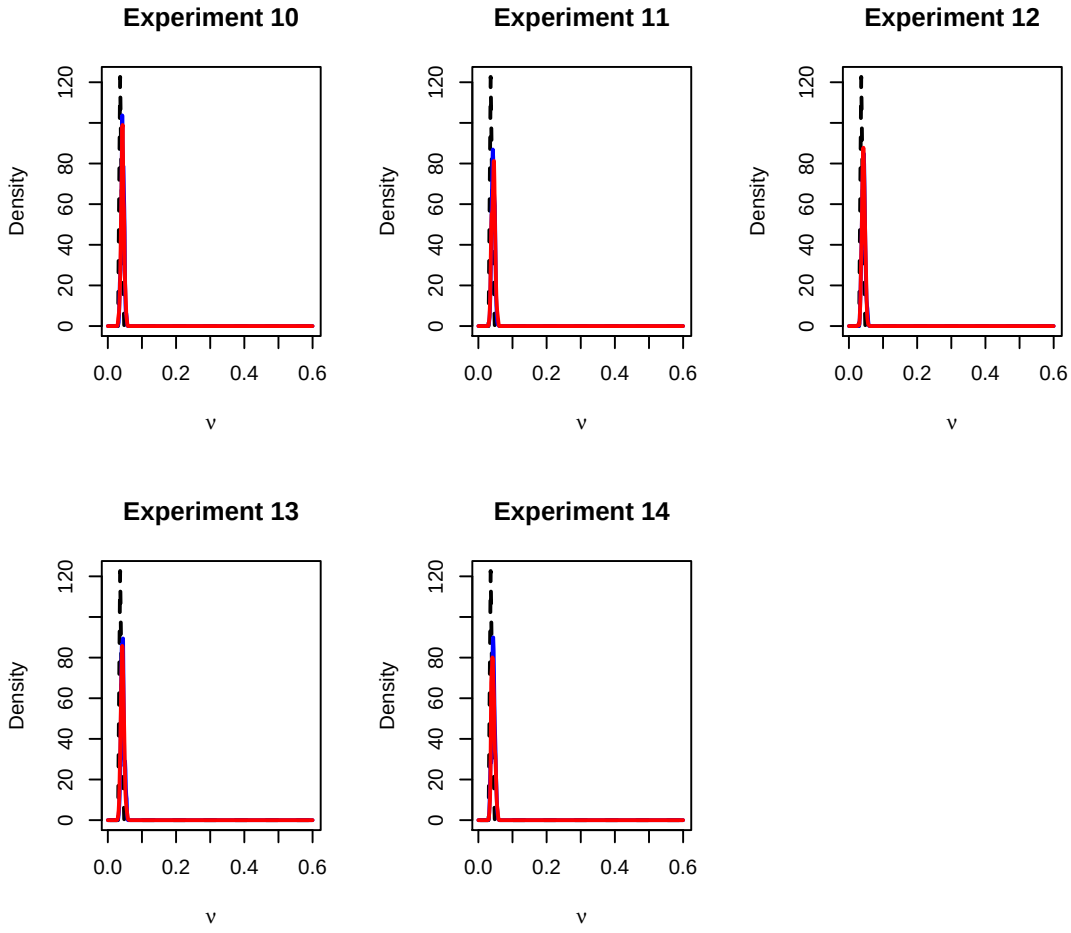

Figure 5.40: Plots of the marginal densities of the diffusion parameter, derived via Tempered ABC SIS (blue) and ABC MCMC (red) and the analytic marginal posterior (black dotted line).

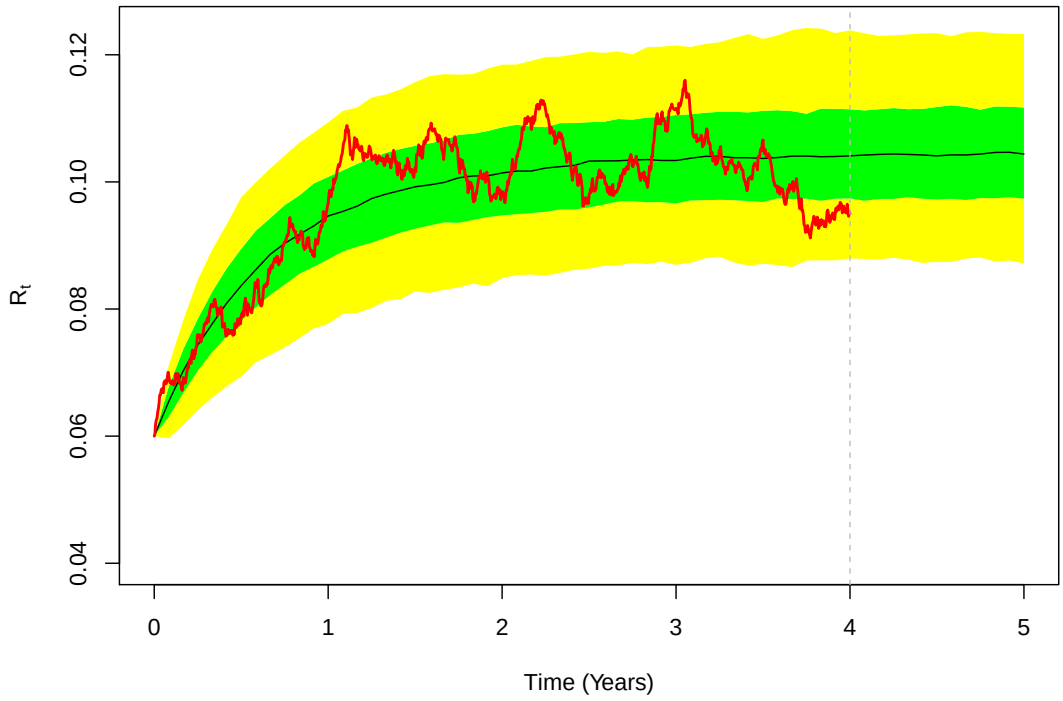

Figure 5.41: Posterior predictive density of five years' worth of new data from the CIR model, conditional on the observations used for inference. The black line represents the median of the PPD, the green area represents the inter-quartile range, and the yellow regions represent the ranges between the 5th and 25th, and the 75th and 95th percentiles. The red line represents the observations that were used to generate our samples from the posterior distribution of parameters via Temepered ABC SIS and ABC MCMC.

5.6 Discussion & summary

In this chapter we provided an overview of the models chosen to test the MC ABC samplers presented in the previous chapter, and explained the rationale behind the choice of these models in particular, namely that they are very well known and can be solved analytically which makes assessing the effectiveness of the samplers easier. We then presented some background theory on summary statistics, and reiterated

the need to utilise summary statistics in order to reduce the dimensionality of the data and maintain the efficiency of the samplers. The various methods of constructing summary statistics were presented, before outlining the experiment set up and discussing the practical issues relating to the experiments, e.g. choosing sampler parameters such as the similarity kernel variance. Finally, the results of the various numerical experiments were presented, which indicated that both MC ABC samplers were able to reproduce reasonably good approximations to the true posterior with certain combinations of summary statistics, for both models considered in this chapter. The semi-automatic summary statistics were able to capture information about the drift parameters of the GBM model, but were not useful when it came to the estimation of the parameter in the diffusion coefficient of (5.2). A potential avenue for further investigation might be to try constructing semi-automatic summary statistics by regressing against different functions of the observed data, e.g. using $f(D) = (D, D^2, \ldots, D^k)$ as explanatory variables in the regression. This approach seemed to work relatively well when we applied it during the CIR model estimation exercise to capture information about the mean reversion rate and mean reversion level. One potential problem associated with this approach is that data sets associated with financial models are often large, which makes the regression stage difficult due to there being a large number of explanatory variables being fitted, e.g. in order to produce reasonably stable regression coefficients it is generally a good idea to ensure that the number of parameters and corresponding data sets used in the regression be a multiple (of perhaps ten or fifteen) of the number of predictor variables (which is just the size of the time series of observations in this case), which means that the design matrix of the regression may be very large, potentially leading to computational issues. With this consideration in mind, it may be neccessary to use the lasso if regressing

against a very large number of predictor variables. The moving average based statistic seemed to perform well in both Tempered ABC SIS and the adapted ABC MCMC samplers, with the relatively sharp peak of the marginal posterior for the diffusion parameter roughly coinciding with the peak of the true marginal posterior. As noted previously, there does seem to be a bias in the empirical densities produced with this statistic, observed in the parameter estimates of the GBM model and the CIR model; however, it is by far the most informative choice of data summary for the diffusion parameters estimated in this thesis. The other ad hoc summary statistics that were tested for each model had mixed results—the mean reversion level parameter in the CIR model was easily identified using a variety of summary statistics, but the statistics associated with the mean reversion rate of the CIR model produced results of varying quality. Specifically, the choice of summary statistics used for the estimation of other parameters seemed to affect the performance of the summary statistics associated with the mean reversion rate. A natural extension of this work would be to test the ABC sampling techniques against more challenging models; both the GBM and CIR models are univariate, and formulating summary statistics based on an intuitive interpretation of the model parameters was a luxury that would not be available to the statistician if larger models were to be estimated using the ABC sampling techniques developed in this thesis. It is clear that the applicability of ABC based samplers to larger models, with larger numbers of parameters, will depend crucially on the ability to construct informative statistics from the observed data—a task the merits investigation on its own.

5.7 Appendix A

In this appendix we explicate the steps involved in constructing the EM based non-sufficient summary statistics associated with the CIR model. The steps followed are exactly the same as those taken in constructing the EM statistics associated with the GBM model (5.2). First, we discretise (5.3) to obtain the EM approximation

$$R_{k+1} - R_k = \alpha(\beta - R_k)\Delta t + \nu R_k^{1/2}\Delta B_k,$$

which implies that the transition density of the approximation is Gaussian

$$f_{EM}(R_{k+1}|R_k, \theta) \sim \mathcal{N}(\mu(R_k;\theta), \sigma^2(R_k;\theta)),$$

where

$$\mu(R_k;\theta) = \bar{\alpha\beta} + R_k(1-\bar{\alpha}), \quad \sigma^2(R_k;\theta) = \nu^2\Delta t R_k,$$

and $\bar{\alpha\beta} = \alpha\beta\Delta t$, $\bar{\alpha} = \alpha\Delta t$. Substituting this expression into the formula for the likelihood of the data, which is given by (5.8), we obtain the following expression for the likelihood function associated with the EM approximated model

$$f_{EM}(R|\theta) = (2\pi\Delta t)^{-N/2}\left(\prod_{k=1}^{N} R_{k-1}^{-1/2}\right)\times \left[\nu^{-N}\exp\left\{\frac{-1}{2\nu^2\Delta t}\sum_{k=1}^{N}\frac{(R_k - \bar{\alpha\beta} - R_{k-1}(1-\bar{\alpha}))^2}{R_{k-1}}\right\}\right], \tag{5.10}$$

where $R = \{R_0, R_1, \ldots, R_N\}$ represents the vector of $N+1$ observations. By utilising the factorisation criterion introduced earlier (see (9)), one observes that expression (5.10) above can be factorised into the product of two functions: one

dependent solely on the data (no parameter dependence), and one that depends on the parameters and various functionals of the data, i.e.

$$f_{EM}(R|\theta) = g(S(R), \theta) \cdot h(R),$$

where

$$g(S(R), \theta) = \nu^{-N} \exp\left\{ \frac{-1}{2\sigma^2 \Delta t} \sum_{k=1}^{N} \frac{(R_k - \bar{\alpha}\bar{\beta} - R_{k-1}(1-\bar{\alpha}))^2}{R_{k-1}} \right\}$$
$$h(R) = (2\pi\Delta t)^{-N/2} \prod_{k=1}^{N} R_{k-1}^{-1/2}.$$

By rearranging the function g, we obtain the following four EM statistics for the CIR model

$$T_1(R) = \sum_{k=1}^{N} R_k^2 R_{k-1}^{-1}, \quad T_2(R) = \sum_{k=1}^{N} R_k R_{k-1}^{-1} \tag{5.11}$$

$$T_3(R) = \sum_{k=1}^{N} R_{k-1}^{-1}, \qquad T_4(R) = \sum_{k=1}^{N} R_k. \tag{5.12}$$

5.8 Appendix B

We present the remaining results from the CIR parameter estimation exercises below.

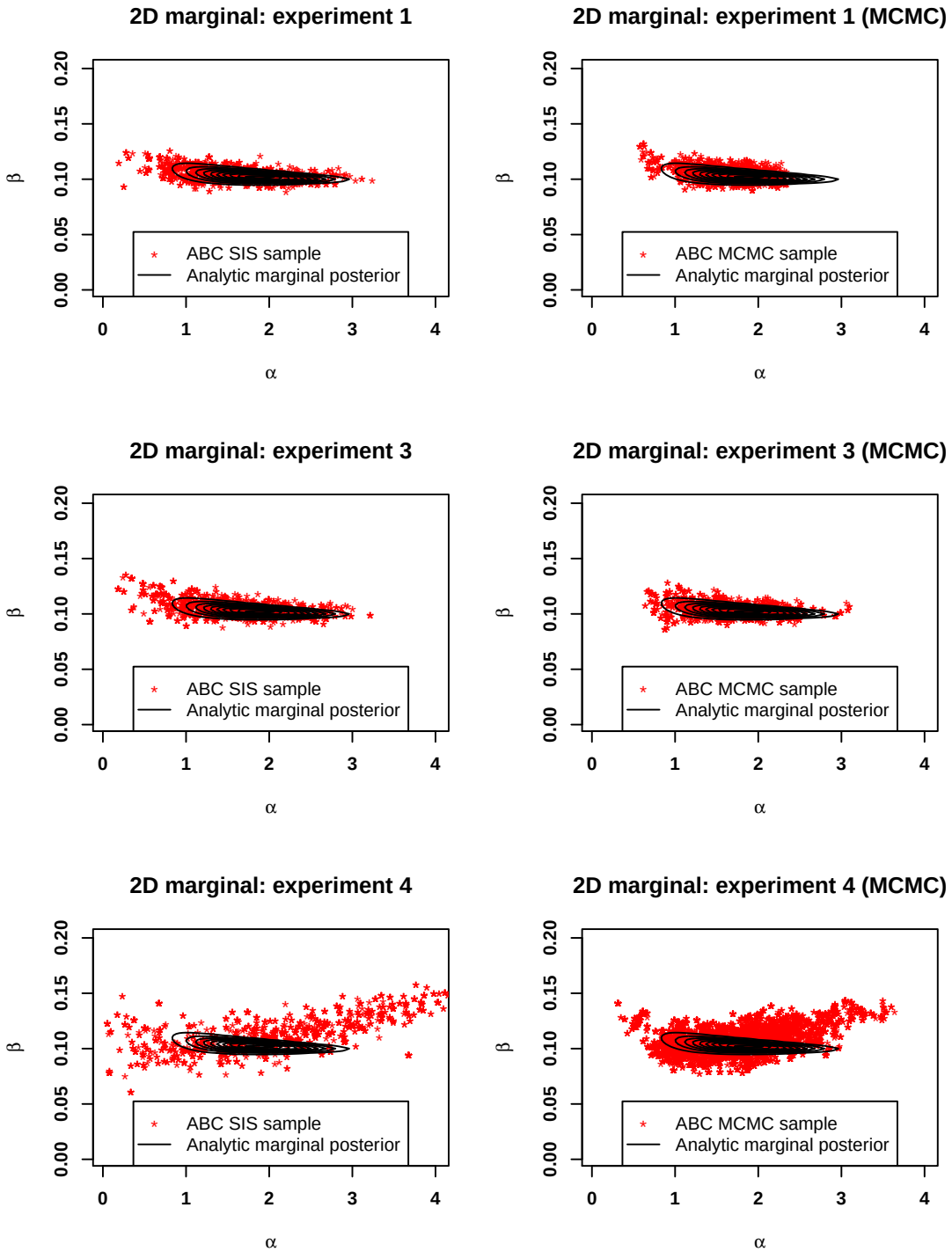

Figure 5.42: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and mean reversion level parameters.

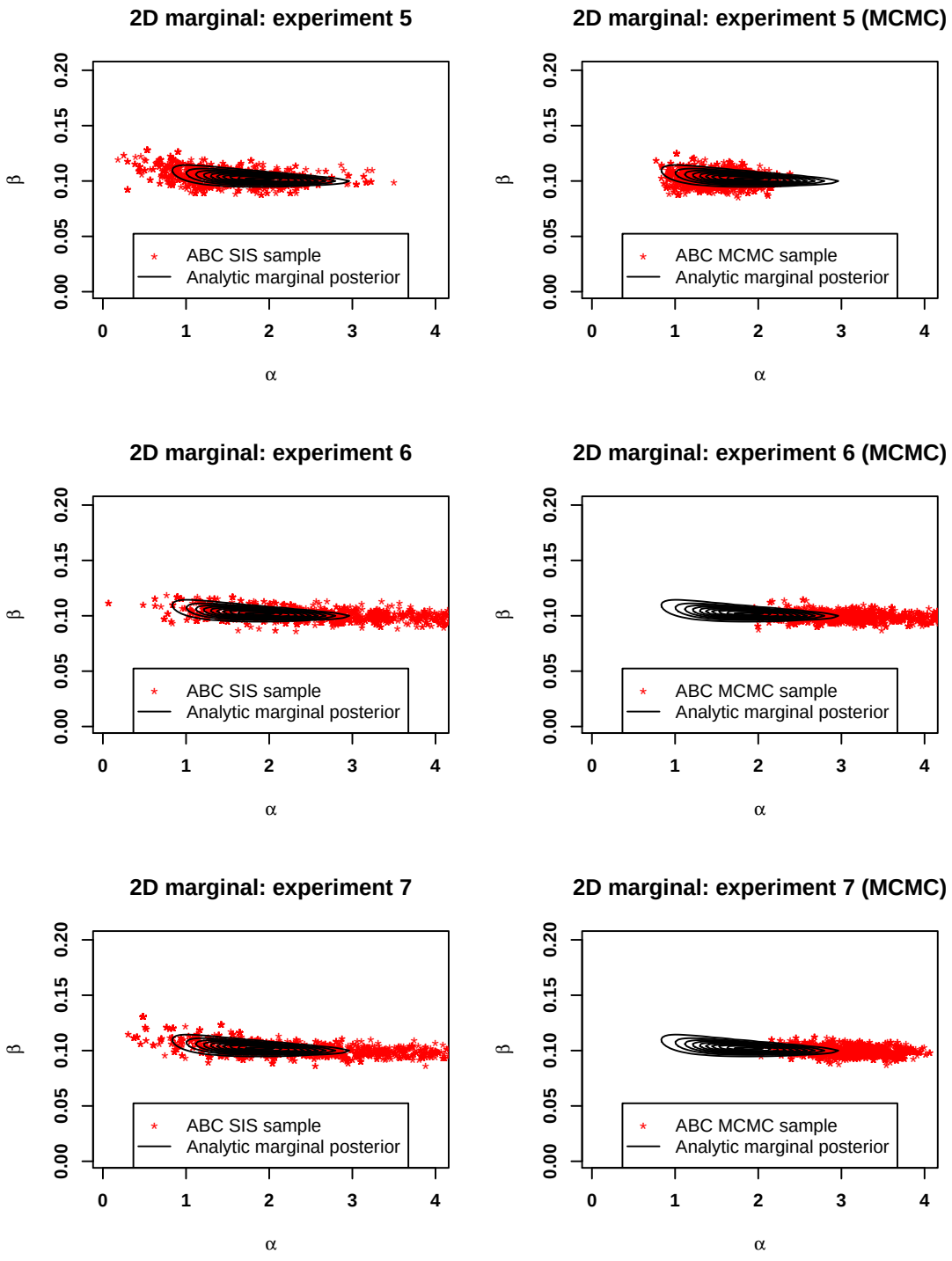

Figure 5.43: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and mean reversion level parameters.

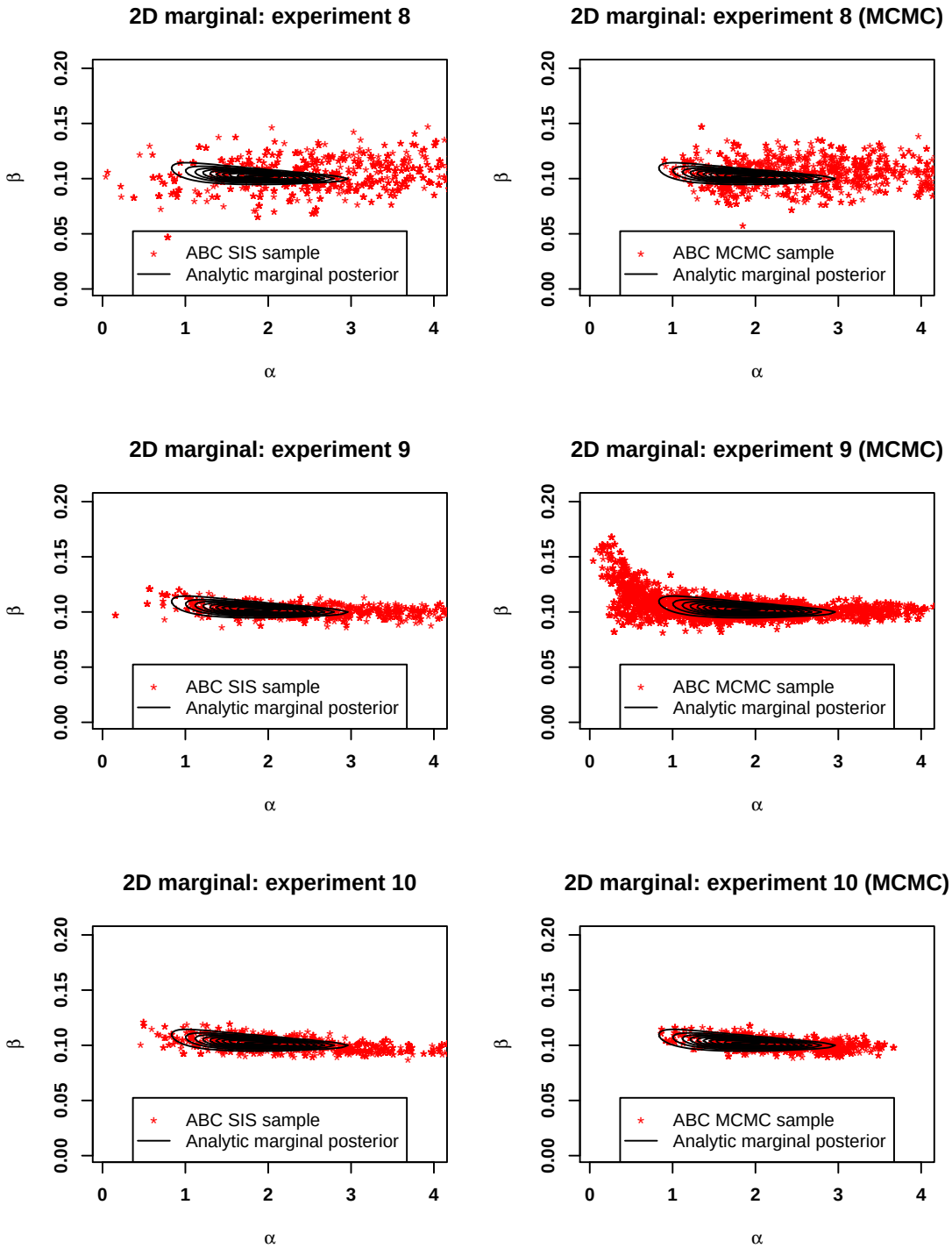

Figure 5.44: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and mean reversion level parameters.

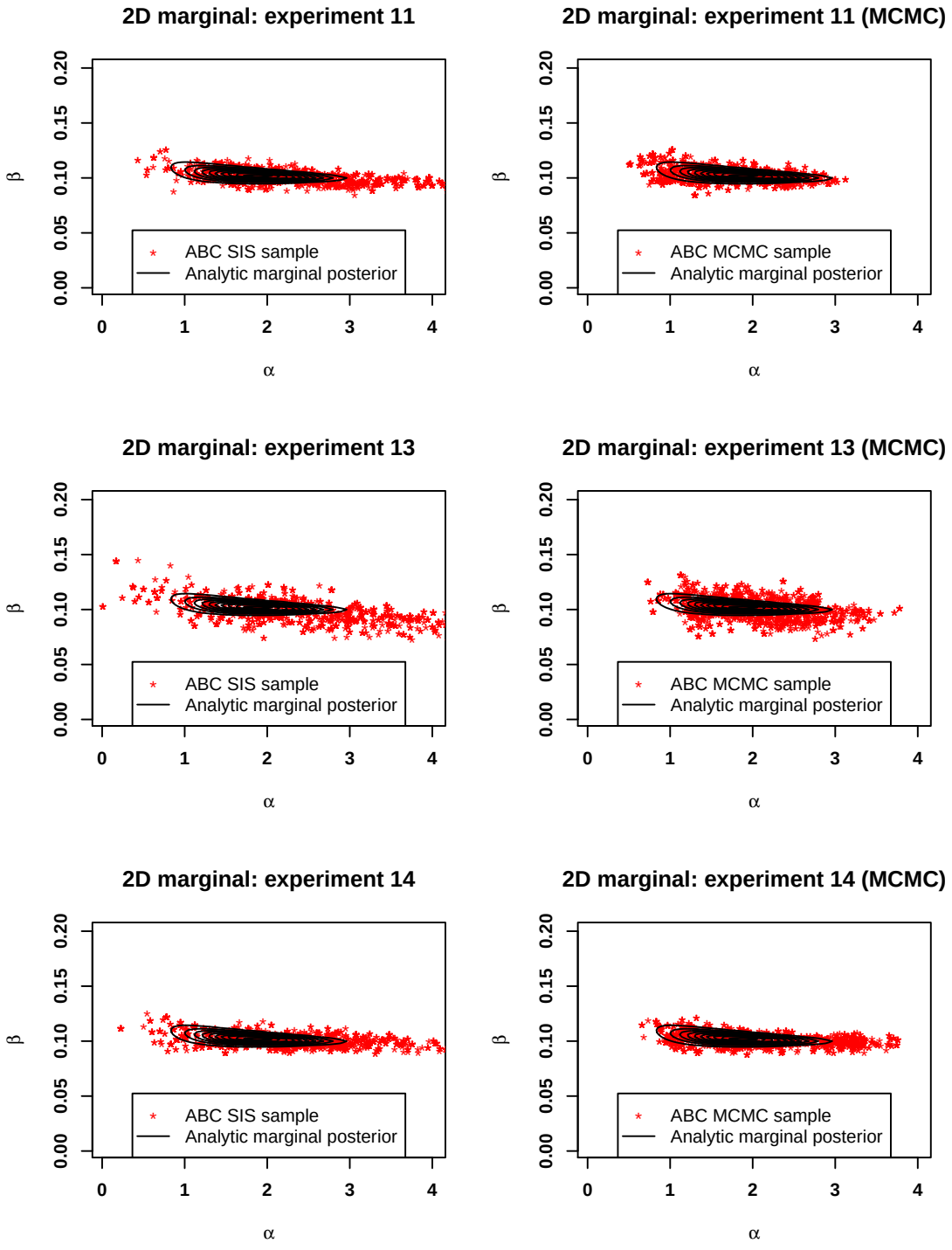

Figure 5.45: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and mean reversion level parameters.

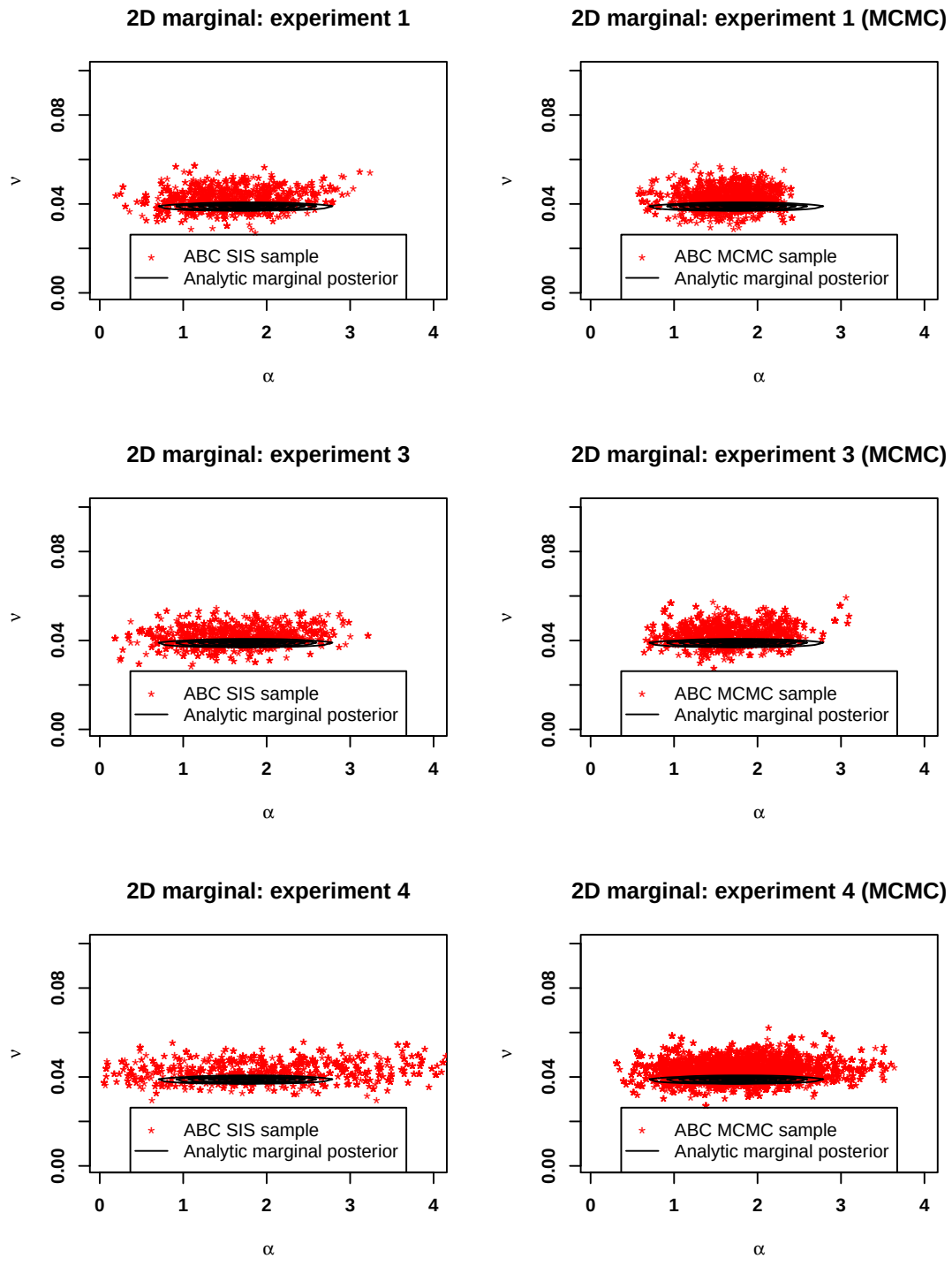

Figure 5.46: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and diffusion parameters.

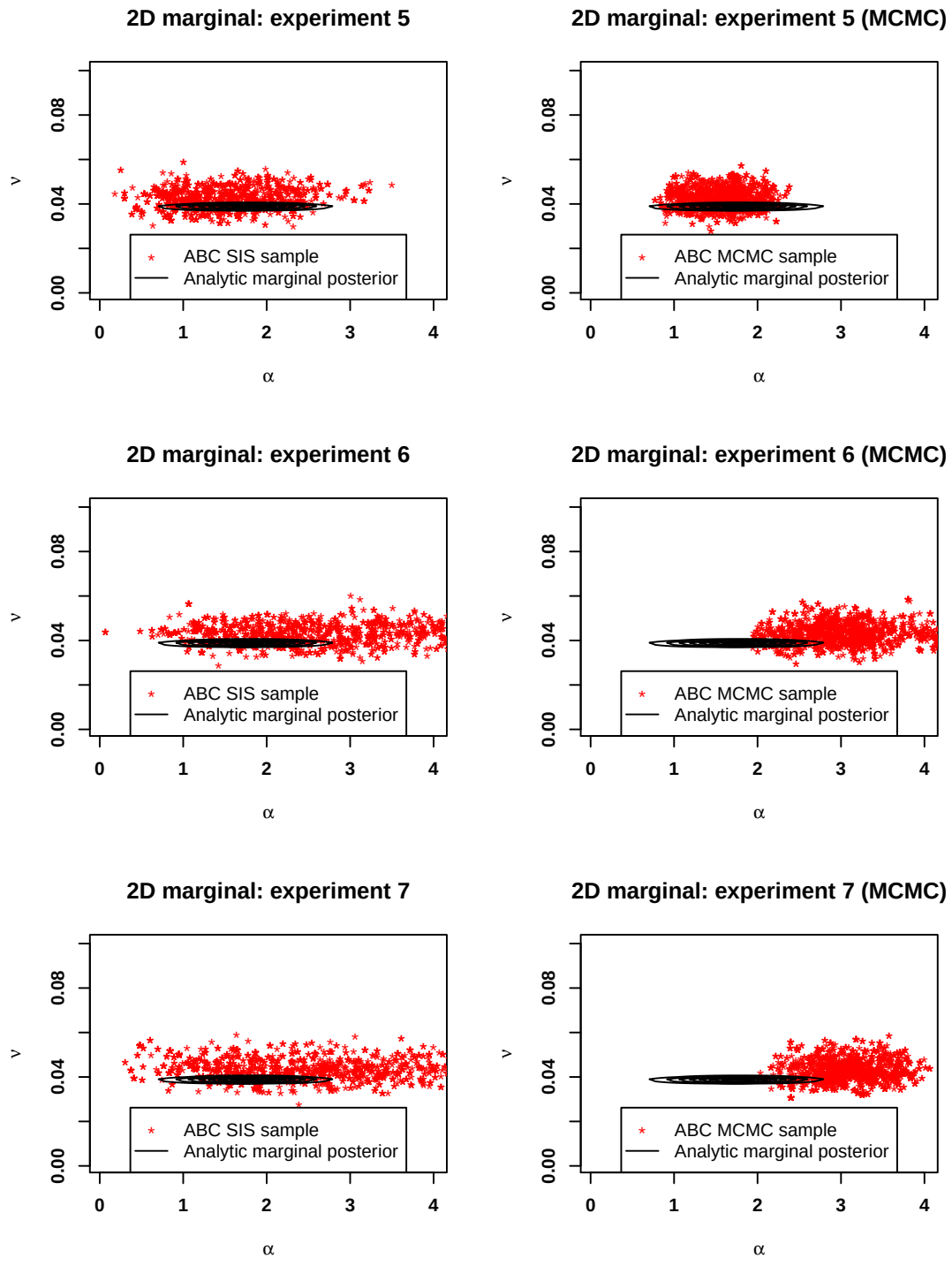

Figure 5.47: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and diffusion parameters.

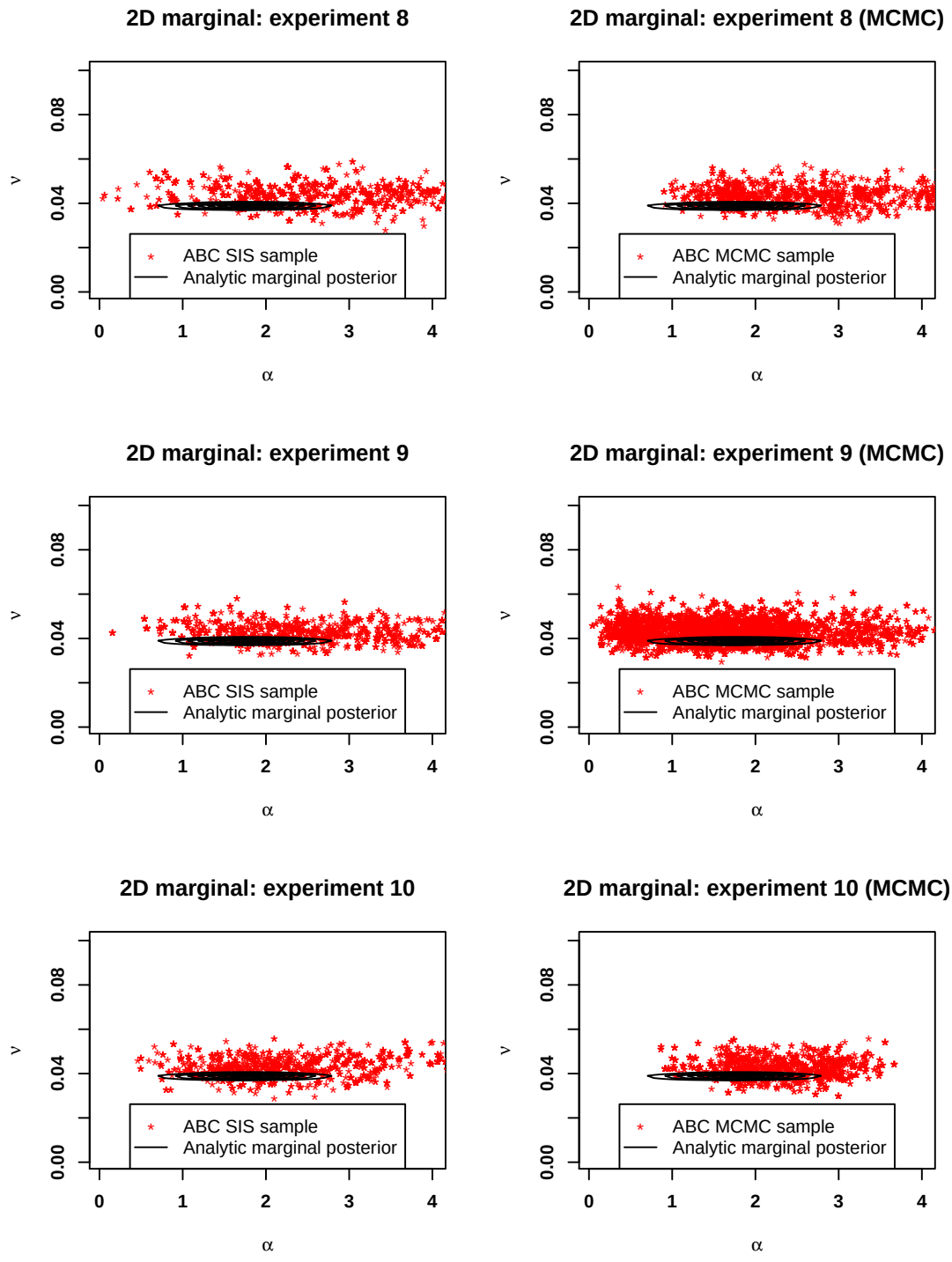

Figure 5.48: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and diffusion parameters.

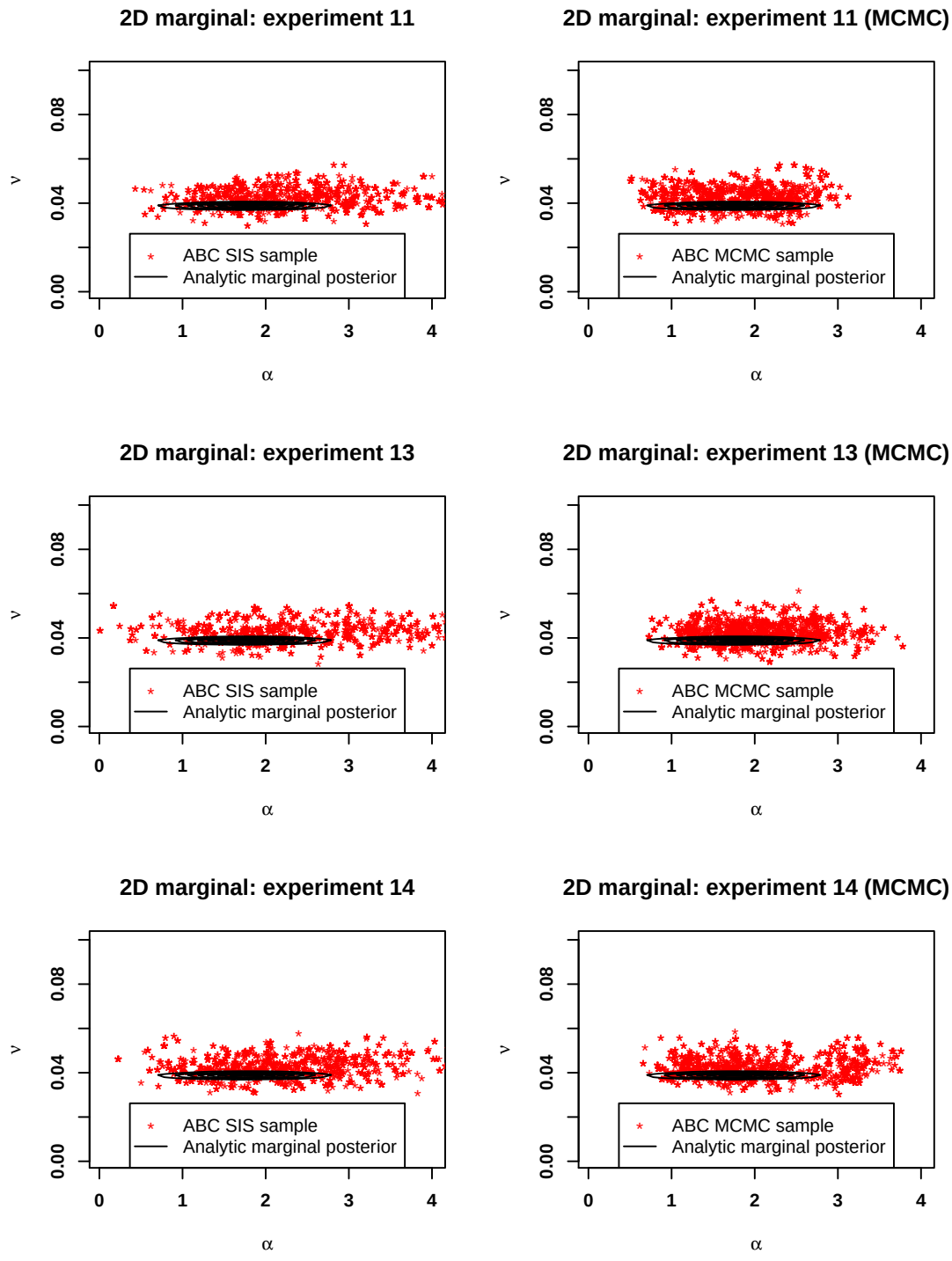

Figure 5.49: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion rate and diffusion parameters.

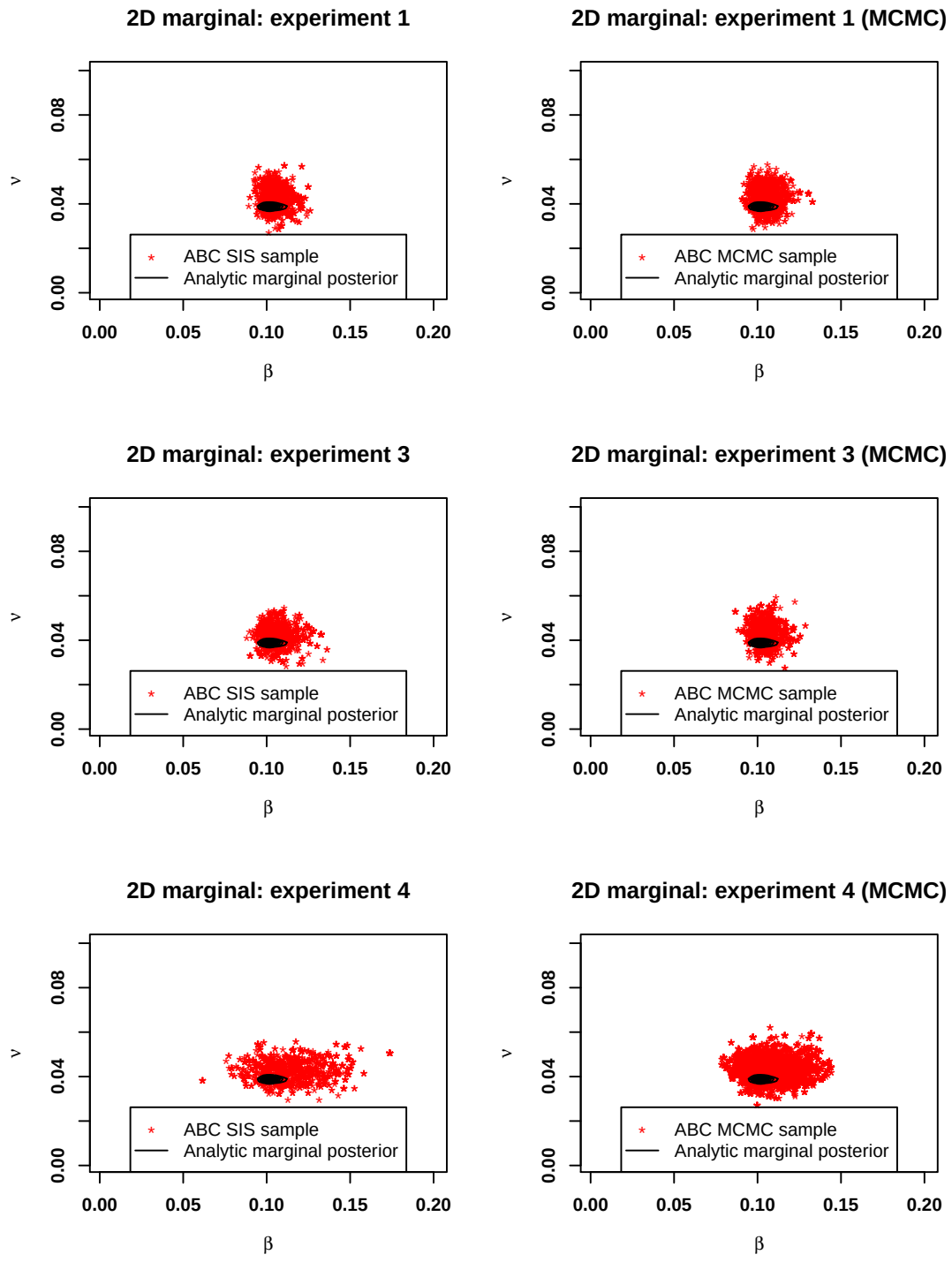

Figure 5.50: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion level and diffusion parameters.

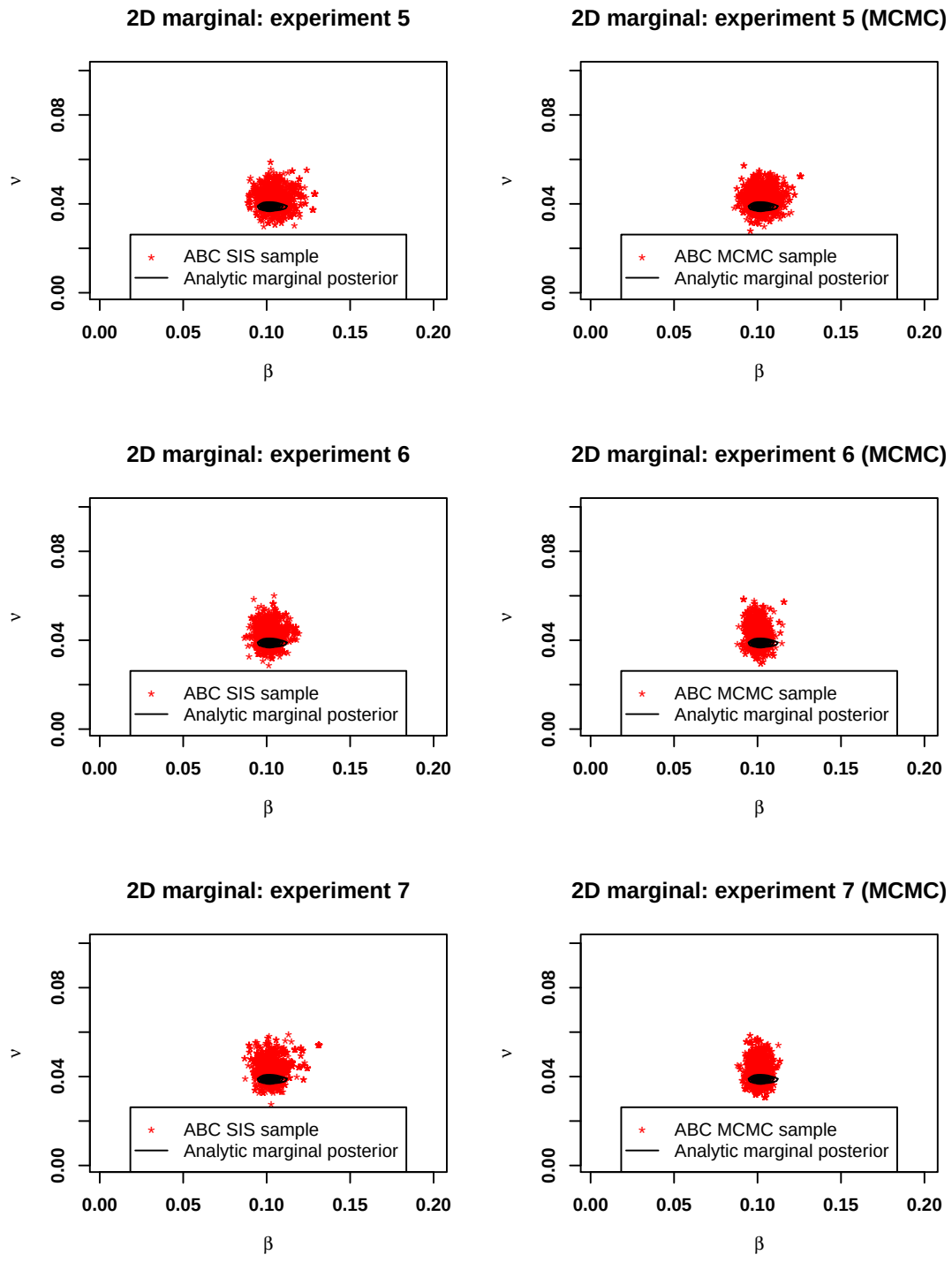

Figure 5.51: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion level and diffusion parameters.

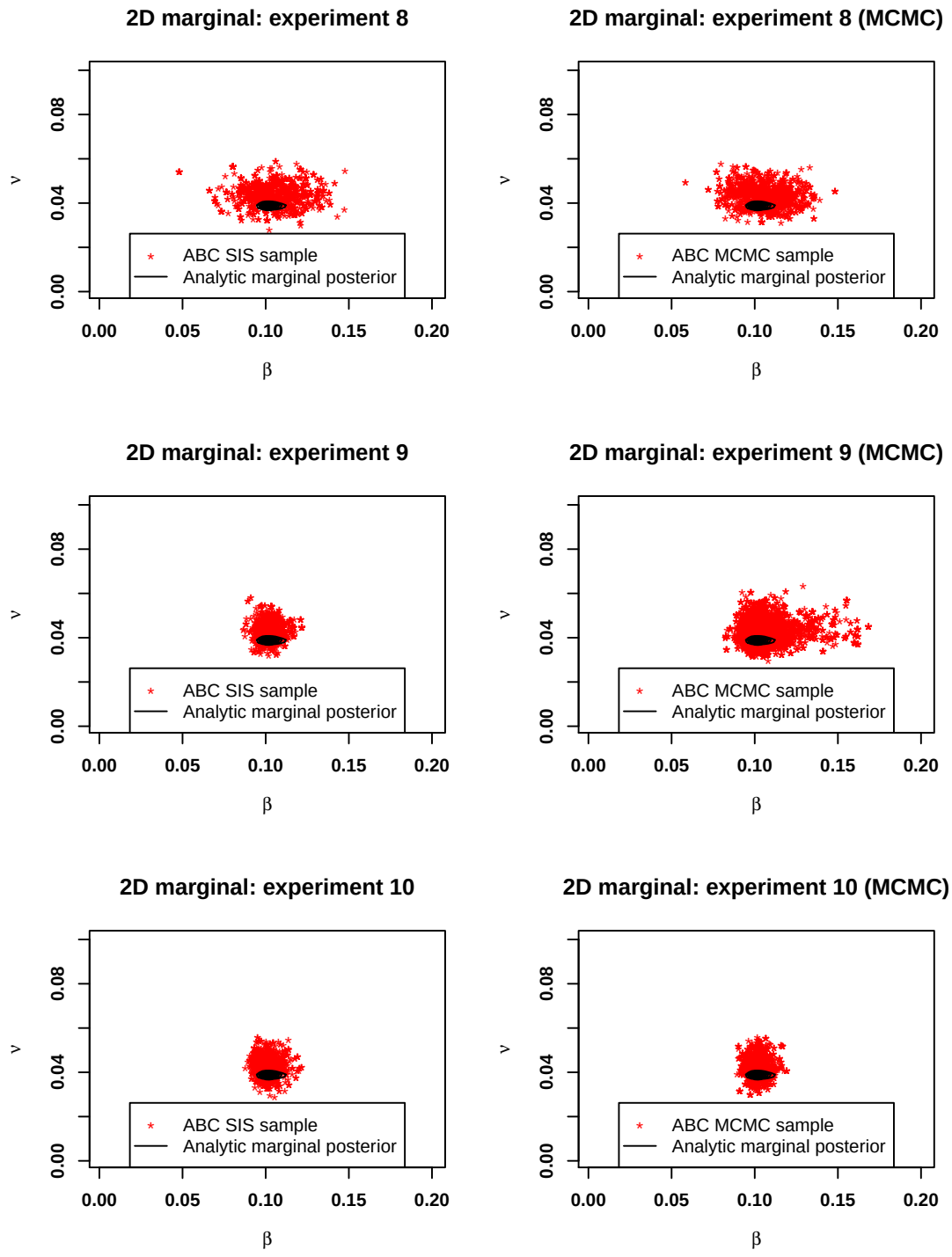

Figure 5.52: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion level and diffusion parameters.

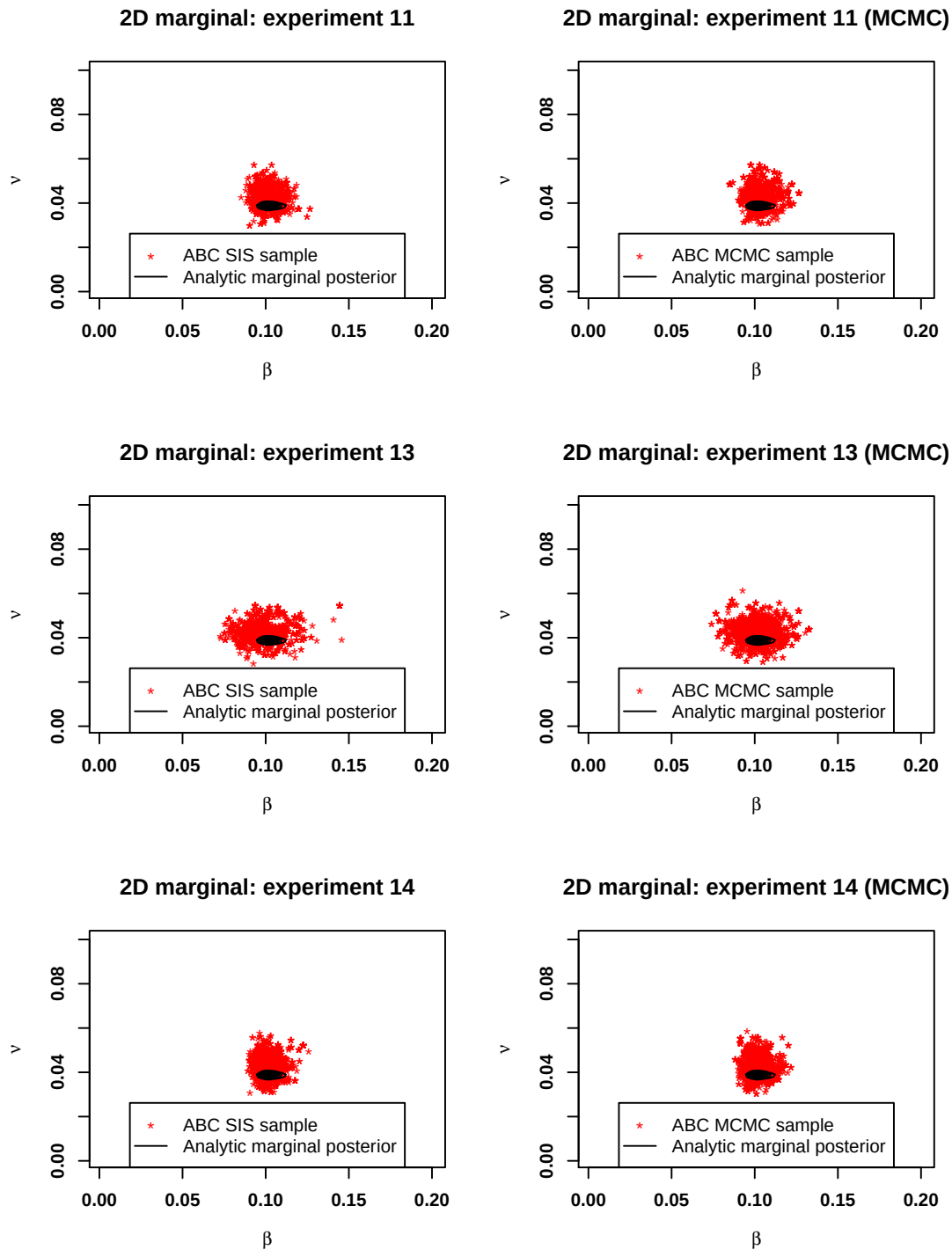

Figure 5.53: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points) for the mean reversion level and diffusion parameters.

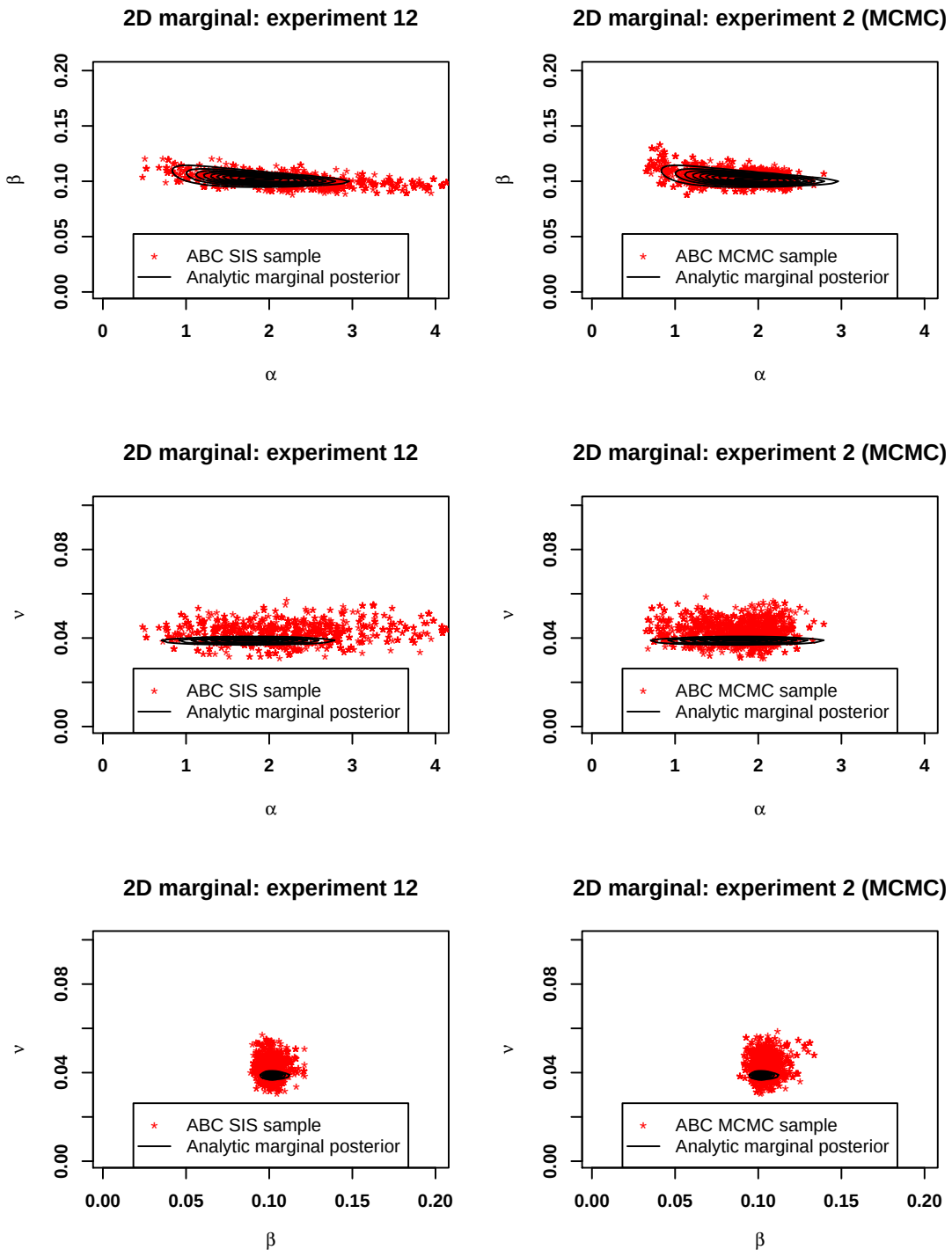

Figure 5.54: Contour plot of the analytic marginal posterior (solid black line) overlaid with the empirical samples from the posterior derived via Tempered ABC SIS and adapted ABC MCMC (red data points).

Chapter 6

Conclusion

6.1 Review

In the course of this thesis we have investigated problems concerning both the design of numerical simulation schemes and parameter estimation methods for SDEs. In Chapter 2 we studied the class of SDEs with linear mean reverting drift and CEV diffusion coefficients (see (2.1)), introducing a drift-implicit numerical approximation scheme that we used to prove the strong convergence of the numerical scheme to the true solution and to provide a lower bound on the rate of strong convergence. Establishing the strong convergence properties of numerical approximations to SDEs is not only a challenging mathematical task, but very important from a practical perspective, e.g. in situations where one has to price path dependent options by Monte Carlo methods, it is the strong convergence properties of the numerical approximation schemes that are relevant when considering the accuracy of the approximated price.

In Chapter 3 we switched focus to the problem of parameter estimation in the context of SDEs. We focussed our analysis on a highly non-linear SDE that has been suggested as a model for the instantaneous, nominal rate of interest, which

we referred to as the Ait-Sahalia (AS) short rate model. We presented a new method of estimating the parameters of the AS SDE, the development of which was based on the insight that when simulating SDEs numerically, drift-implicit discretisation schemes often retain some important qualitative properies possessed by the analytic solution of the SDE. We tested the new method of parameter estimation against a standard estimation technique and found that both methods of estimation were unsatisfactory. Following this conclusion, we presented some analysis to further examine why the parameter estimation techniques performed so poorly, and discovered parameter identifiability problems with the model itself; in particular, there were a large number of parameter values that could credibly have given rise to the data. This conclusion motivated a discussion regarding the effectiveness of standard, likelihood based inference, and whether a Bayesian approach to parameter estimation might represent a more appropriate approach.

Following the conclusions of Chapter 3, we began Chapter 4 by introducing Bayesian inference and discussing the pros and cons of the Bayesian approach relative to the likelihood based approach to parameter estimation. We then discussed approximate Bayesian computation (ABC) and its usefulness when it comes to estimating complicated stochastic models that are intractible. After presenting a summary of the standard ABC sampling algorithms in the literature, we proposed some new ABC based samplers that were capable of deriving samples from an approximation to the model posterior without any knowledge of the model likelihood.

In Chapter 5 we applied our newly developed ABC samplers to two test cases that are representative of the sort of models commonly encountered in financial applications. We spent time discussing the need for summary statistics that are capable of condensing the information contained within observed data into a smaller dimensional object, and proposed various methods of constructing such statistics

when the data are assumed to come from SDEs. We then applied our newly developed ABC sampling algorithms (Tempered ABC SIS and adapted ABC MCMC) to two widely used models in mathematical finance, geometric Brownian motion and the square root process, and presented the resulting approximate posterior distributions alongside the analytic solutions in order to assess the quality of the respective approximations. Both of the newly developed ABC samplers were able to produce good approximations to the model posterior, for both geometric Brownian motion and the square root process. This was the first time (to our knowledge) that ABC methods have been used to estimate the parameters of SDEs, and the promising results obtained suggest that, with further work, this approach to model estimation could be extremely useful in practice.

6.2 Avenues for further research

While investigating the efficacy of the ABC samplers presented in Chapter 4, it became clear that the applicability of ABC to the problem of estimating the parameters of SDEs was critically dependent on being able to construct low dimensional summary statistics from the data. Without summary statistics, Tempered ABC SIS produced highly degenerate samples from the approximate posterior, and the adapted ABC MCMC sampler produced Markov chains that exhibited very poor mixing. The estimation results obtained for the GBM model using sufficient summary statistics clearly demonstrate that both ABC samplers presented in this thesis are capable of producing very high quality approximations to the true model posterior, but when less informative summary statistics were utilised in the ABC samplers the results were, unsurprisingly, of lower quality. The results relating to the square root process are consistent with this observation. In the examples considered in Chapter 5 we were able to construct summary statistics for each

model parameter by considering the role each model parameter plays in determining the dynamics of the modelled process. For example, the mean reversion level parameter of the square root process can be interpreted as the long run level to which the process tends to drift over time, which suggests that the sample mean of the observations should contain information about this parameter. While this approach is possible for some simpler models, it is generally not possible to attach a physical interpretation to the parameters associated with more complicated models (e.g. multidimensional SDEs), and therefore this method of summary statistic construction may not be available in general. For this method of parameter estimation to be applicable to larger, more complex models (for which a likelihood free method of parameter estimation would be most valuable), more general methods of constructing summary statistics need to be developed. As was pointed out in Chapter 5, Fearnhead and Prangle (2012) have developed a *semi automatic* procedure for constructing summary statistics, which we applied with mixed success in the parameter estimation case studies in Chapter 5. While this method of statistic construction worked well for some parameters, it failed to produce informative statistics for others, especially those found in the diffusion coefficient of the SDEs. To our knowledge there are no other generic procedures for constructing summary statistics in the literature, and therefore this challenge represents an interesting topic that merits further investigation. If a generic procedure for summary statistic construction can be developed, ABC parameter estimation could become one of the most flexible and powerful methods of calibrating models to data.

Bibliography

Ait-Sähalia, Y. (1996). Testing continuous-time models of the spot interest rate. *Review of Financial Studies 9*(2), 385–426.

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

Beaumont, M. (2010). Approximate bayesian computation in evolution and ecology. *Annual Review of Ecol., Evol., and Systematics 41*(1), 379–406.

Beaumont, M., J. Cornuet, J. Marin, and C. Robert (2009). Adaptive approximate bayesian computation. *Biometrika 96*(4), 983–990.

Bernardo, J. and A. Smith (2000). *Bayesian Theory*. John Wiley Sons.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy 81*(3), 637–654.

Bortot, P., S. Coles, and S. Sisson (2007). Inference for stereological extremes. *J. Am. Stat. Assoc. 102*, 84–92.

Brigo, D. and F. Mercurio (2006). *Interest Rate Models - Theory and Practice*. Springer.

Calderhead, B. and M. Girolami (2003). Estimating bayes factors via thermodunamic integration and population mcmc. *Computational Statistics and Data Analysis 53*(12), 4028–4045.

Chan, K., G. Karolyi, F. Longstaff, and A. Sanders (1992). An empirical comparison of alternative models of the short-term interest rate. *The Journal of Finance 47*(3), 1209–1227.

Chopin, N. (2004). Central limit theorem for sequential monte carlo methods and its application to bayesian inference. *Ann. Stat. 32*(6), 2385–2411.

Cox, J., J. Ingersoll, and S. Ross (1985). A theory of the term structure of interest rates. *Econometrica 53*, 385–408.

Del Moral, P., A. Doucet, and A. Jasra (2006). Sequential monte carlo samplers. *J. R. Statist. Soc. B 68*(3), 411–436.

Dereich, S., A. Neuenkirch, and L. Szpruch (2012). An euler-type method for the strong approximation of the cox-ingersoll-ross process. *Proc. R. Soc. A 468*(2140), 1105–1115.

Fearnhead, P. and D. Prangle (2012). Constructing summary statistics for approximate bayesian computation: Semi-automatic approximate bayesian computation. *J. R. Statist. Soc. B 74*(3), 419–474.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. of the R. Soc. of London, Series A* (222), 309–368.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis*. Chapman Hall/CRC Press.

Gelman, A. and X. Meng (1999). Simulating normalising constants: From importance sampling to bridge sampling to path sampling. *Statistical Science 13*(2), 163–185.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman Hall/CRC Press.

Glasserman, P. (2010). *Monte Carlo Methods in Financial Engineering*. Springer.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Higham, D., X. Mao, and A. Stuart (2002). Strong convergence of euler-type methods for nonlinear stochastic differential equations. *SIAM J. Numer. Anal. 40*(3), 1041–1063.

Hull, J. (2009). *Options, Futures and Other Derivatives*. Prentice Hall.

Hurn, A., J. Jeisman, and K. Lindsay (2007). Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *Journal of Financial Econometrics 5*(3), 390–455.

Iacus, S. (2008). *Simulation and Inference for Stochastic Differential Equations*. Springer.

Johannes, M. and N. Polson (2010). *Handbook of Financial Econometrics*, Chapter : MCMC Methods for Continuous-Time Financial Econometrics. North-Holland.

Joyce, P. and P. Marjoram (2008). Approximately sufficient statistics and bayesian computation. *Stat. App. in Gen. and Mol. Biol. 7*(1).

Mao, X. (2008). *Stochastic Differential Equations and Applications*. Horwood.

Mao, X., A. Truman, and C. Yuan (2006). Euler-maruyama approximations in mean-reverting stochastic volatility model under regime-switching. *International Journal of Stochastic Analysis 2006*.

Mao, X. and C. Yuan (2006). *Stochastic Differential Equations with Markovian Switching*. Imperial College Press.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavare (2003). Markov chain monte carlo without likelihoods. *Proc. Nat. Acad. Sci. USA 100*(26), 15324–15328.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machine. *J. Chem. Phys. 21*, 1087–1091.

Meyn, S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag.

Müller-Gronbach, T. (2002). The optimal uniform approximation of systems of stochastic differential equations. *Ann. App. Prob. 12*(2), 664–690.

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

Nelder, J. and R. Mead (1965). A simplex method for function minimisation. *Computer Journal* (7), 308–313.

Neuenkirch, A. and L. Szpruch (2014). First order strong approximations of scalar sdes with values in a domain. *Numerische Mathematik*, 1–34.

Nocedal, J. and S. Wright (2006). *Numerical Optimisation* (2nd Ed. ed.). Springer.

Norris, J. (1997). *Markov Chains*. Cambridge University Press.

Nowman, K. (1997). Gaussian estimation of single-factor continuous time models of the term structure of interest rates. *The Journal of Finance 52*(4), 1695–1706.

Øksendal, B. (2007). *Stochastic Differential Equations, An Introduction with Applications*, Chapter 7. Springer.

Pritchard, J., M. Seielstad, A. Perez-Leznaun, and M. Feldman (1999). Population growth of human y chromosomes: A study of y chromosome microsatellites. *Mol. Biol. Evol. 16*(12), 1791–1798.

Ratmann, O., O. Jorgensen, T. Hinkley, M. Stumpf, S. Richardson, and C. Wiuf (2007). Using likelihood-free inference to compare evolutionary dynamics of the protein networks of h. pylori and p. falciparum. *PLoS Comput. Biol. 3*, 2266–2278.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat. 12*, 1151–1172.

Shreve, S. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer.

Sisson, S., Y. Fan, and M. Tanaka (2007). Sequential monte carlo without likelihoods. *Proc. Nat. Acad. Sci. USA 104*, 1760–1765.

Szpruch, L., X. Mao, D. Higham, and J. Pan (2011). Strongly nonlinear ait-sähalia-type interest rate model and its numerical approximation. *BIT Numerical Mathematics 51*(2), 405–425.

Tiribshani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Stat. Soc.*, 267–288.

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. Stumpf (2009). Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface 6*, 187–202.

Wade, A. (2010). Markov chains in discrete time. *SMSTC lecture notes (probability stream)*.

Wilkinson, R. (2013). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology 12*(2), 129–141.

Zeidler, E. (1989). *Nonlinear Monotone Operators*. Springer.