Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2009)
71, *Part* 1, *pp.* 49–73

Parameter estimation for partially observed hypoelliptic diffusions

Yvo Pokern and Andrew M. Stuart

University of Warwick, Coventry, UK

and Petter Wiberg

Goldman-Sachs, London, UK

[Received March 2006. Final revision March 2008]

Summary. Hypoelliptic diffusion processes can be used to model a variety of phenomena in applications ranging from molecular dynamics to audio signal analysis. We study parameter estimation for such processes in situations where we observe some components of the solution at discrete times. Since exact likelihoods for the transition densities are typically not known, approximations are used that are expected to work well in the limit of small intersample times Δt and large total observation times $N\,\Delta t$. Hypoellipticity together with partial observation leads to ill conditioning requiring a judicious combination of approximate likelihoods for the various parameters to be estimated. We combine these in a deterministic scan Gibbs sampler alternating between missing data in the unobserved solution components, and parameters. Numerical experiments illustrate asymptotic consistency of the method when applied to simulated data. The paper concludes with an application of the Gibbs sampler to molecular dynamics data.

Keywords: Gibbs sampler; Hypoelliptic diffusion; Numerical methods; Parameter estimation; Partial observation

1. Introduction

In many application areas it is of interest to model some components of a large deterministic system by a low dimensional stochastic model. In some of these applications, insight from the deterministic problem itself forces structure on the form of the stochastic model, and this structure must be reflected in parameter estimation. In this paper, we study the fitting of stochastic differential equations (SDEs) to discrete time series data in situations where the model is a hypoelliptic diffusion process, meaning that the covariance matrix of the noise is degenerate, but the probability densities are smooth, and also where observations are only made of variables that are not directly forced by white noise. Such a structure arises naturally in various applications.

One application is the modelling of macromolecular systems; see Grubmüller and Tavan (1994) and Hummer (2005). In its basic form, molecular dynamics describe the molecule by a large Hamiltonian system of ordinary differential equations. As is commonplace in chemistry and physics, we shall refer to data that are obtained from numerical simulation of such models as molecular dynamics data. If the molecule spends most of its time in a small number of macroscopic configurations then it may be appropriate to model the dynamics within, and in some cases between, these states by a hypoelliptic diffusion. Although this phrasing of the

Address for correspondence: Yvo Pokern, Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK.
E-mail: Y.Pokern@warwick.ac.uk

 1369–7412/09/71049

question is relatively recent, under the name of the 'Kramers problem' it dates back to Kramers (1940) with a brief summary in section 5.3.6a of Gardiner (1985). Another application, audio signal analysis, is referred to in Giannopoulos and Godsill (2001) where a continuous time auto-regressive moving average model is used; see also Godsill and Yang (2006) for more on the type of methodology that is used.

We consider SDE models of the form

$$\begin{aligned} \mathrm{d}x &= \Theta\, A(x)\,\mathrm{d}t + C\,\mathrm{d}B, \\ x(0) &= x_0 \end{aligned} \tag{1}$$

where B is an m-dimensional Wiener process and x a k-dimensional continuous process with $k > m$. $A : \mathbb{R}^k \to \mathbb{R}^l$ is a set of (possibly non-linear) globally Lipschitz force functions. The parameters which we estimate are the last m rows of the drift matrix (the first $k-m$ rows of which are assumed to be known), $\Theta \in \mathbb{R}^{k\times l}$, and the diffusivity matrix C which we assume to be of the form

$$C = \begin{pmatrix} 0 \\ \Gamma \end{pmatrix} \in \mathbb{R}^{k\times m}.$$

where $\Gamma \in \mathbb{R}^{m\times m}$ is a constant non-singular matrix. Thus, we are estimating drift and diffusion parameters only in the co-ordinates which are directly driven by white noise.

It is known that, under suitable hypotheses on A and C, a unique L^2-integrable solution $x(\cdot)$ exists almost surely for all times $t \in \mathbb{R}^+$; see for example theorem 5.2.1 in Øksendal (2000). We also assume that the process that is defined by model (1) is hypoelliptic as defined in Nualart (1991). Intuitively, this corresponds to the noise being spread into all components of the system (1) via the drift.

The structure of C implies that the noise acts directly only on a subset of the variables which we refer to as *rough*. It may then be transmitted, through the coupling in the drift, to the remaining parts of the system which we refer to as *smooth* (we do not mean C^∞ here, but they are at least C^1). To distinguish between rough and smooth variables, we introduce the notation $x(t)^{\mathrm{T}} = (u(t)^{\mathrm{T}}, v(t)^{\mathrm{T}})$ where $u(t) \in \mathbb{R}^{k-m}$ is smooth and $v(t) \in \mathbb{R}^m$ is rough. It is helpful to define projections $\mathcal{P} : \mathbb{R}^k \to \mathbb{R}^{k-m}$ by $\mathcal{P}x = u$ and $\mathcal{Q} : \mathbb{R}^k \to \mathbb{R}^m$ by $\mathcal{Q}x = v$.

We denote the sample path at $N+1$ equally spaced points in time by $\{x_n = x(n\,\Delta t)\}_{n=0}^N$, and we write $x_n^{\mathrm{T}} = (u_n^{\mathrm{T}}, v_n^{\mathrm{T}})$ to separate the rough and smooth components. Also, for any sequence $(z_1, \ldots, z_N)$, $N \in \mathbb{N}$, we write $\Delta z_n = z_{n+1} - z_n$ to denote forward differences. We are mainly interested in cases where only the smooth component u is observed and our focus is on parameter estimation for all of Γ and for entries of those rows of Θ corresponding to the rough path, on the assumption that $\{u_n\}_{n=0}^N$ are samples from a true solution of system (1); such a parameter estimation problem arises naturally in many applications and an example is given in Section 7. We shall describe a deterministic scan Gibbs sampler to approach this problem, sampling alternatingly from the missing path $\{v_n\}_{n=0}^N$, the drift parameters Θ and the covariance $\Gamma\Gamma^{\mathrm{T}}$. It is natural to consider $N\Delta t = T \gg 1$ and $\Delta t \ll 1$.

Given prior distributions for the parameters, $p_0(\Theta, \Gamma\Gamma^{\mathrm{T}})$, the posterior distribution can be constructed as follows:

$$\mathbb{P}(v, \Theta, \Gamma\Gamma^{\mathrm{T}} | u) \propto \mathcal{L}(u, v | \Theta, \Gamma\Gamma^{\mathrm{T}})\; p_0(\Theta, \Gamma\Gamma^{\mathrm{T}}). \tag{2}$$

Here, $\mathcal{L}(u, v|\Theta, \Gamma\Gamma^{\mathrm{T}})$ has been introduced as a measure equal to the probability density $\mathbb{P}(u, v|\Theta, \Gamma\Gamma^{\mathrm{T}})$ up to a constant of proportionality. When u and v are fixed and $\mathcal{L}(u, v|\Theta, \Gamma\Gamma^{\mathrm{T}})$ is thought of as a function of Θ and $\Gamma\Gamma^{\mathrm{T}}$ it is a likelihood.

Similarly, the probability densities $\mathbb{P}(v|\Theta, \Gamma\Gamma^{\mathrm{T}}, u)$, $\mathbb{P}(\Theta|v, \Gamma\Gamma^{\mathrm{T}}, u)$ and $\mathbb{P}(\Gamma\Gamma^{\mathrm{T}}|v, \Theta, u)$ are replaced by corresponding expressions using $\mathcal{L}$ when omitting constants of proportionality that are irrelevant to estimation of the posterior probability. The probability density $\mathbb{P}(u, v|\Theta, \Gamma\Gamma^{\mathrm{T}})$ gives rise to the transition density $\mathbb{P}(u_{n+1}, v_{n+1}|u_n, v_n, \Theta, \Gamma\Gamma^{\mathrm{T}})$, which we shall write as $\mathcal{L}(u_{n+1}, v_{n+1}|u_n, v_n, \Theta, \Gamma\Gamma^{\mathrm{T}})$ when omitting constants of proportionality.

In principle, expression (2) can be used as the basis for Bayesian sampling of $(\Theta, \Gamma\Gamma^{\mathrm{T}})$, viewing v as missing data. However, the exact probability of the path, $\mathbb{P}(u, v|\Theta, \Gamma\Gamma^{\mathrm{T}})$, is typically unavailable. In this paper we shall combine judicious approximations of this density to solve the sampling problem.

The sequence $\{x_n\}_{n=0}^{N}$ that is defined above is generated by a Markov chain. The random map $x_n \mapsto x_{n+1}$ is determined by the integral equation

$$x_{n+1} = x_n + \int_{n\Delta t}^{(n+1)\Delta t} \Theta A\{x(s)\}\,\mathrm{d}s + \int_{n\Delta t}^{(n+1)\Delta t} C\,\mathrm{d}B(s).$$

The Euler–Maruyama approximation of this map gives

$$X_{n+1} \approx X_n + \Delta t\Theta\, A(X_n) + \sqrt{\Delta t}\, R(0, \Theta)\xi_n \tag{3}$$

where $X_n, \xi_n \in \mathbb{R}^k$, ξ_n is an independent and identically distributed (IID) sequence of normally distributed random variables, $\xi_n \sim \mathcal{N}(0, I)$, and

$$R(0, \Theta) = \begin{pmatrix} 0 & 0 \\ 0 & \Gamma \end{pmatrix} \in \mathbb{R}^{k\times k}$$

is not invertible. (Here, as throughout, we use upper-case letters to denote discrete time approximations of the continuous time process.) This approximation corresponds to retaining the terms of order $\mathcal{O}(\Delta t)$ in the drift and of $\mathcal{O}(\sqrt{\Delta t})$ in the noise when performing an Itô–Taylor expansion (see chapter 5 of Kloeden and Platen (1992)). Owing to the non-invertibility of $R(0, \Theta)$, this approximation is unsuitable for many purposes and we extend it by adding the first non-zero noise terms arising in the first $k - m$ rows of the Itô–Taylor expansion for X_{n+1}. This results in the expression

$$X_{n+1} \approx X_n + \Delta t\Theta\, A(X_n) + \sqrt{\Delta t}\, R(\Delta t; \Theta)\xi_n \tag{4}$$

where $X_n \in \mathbb{R}^k, \xi_n \in \mathbb{R}^k$, is distributed as $\mathcal{N}(0, I)$ and $R(\Delta t; \Theta) \in \mathbb{R}^{k\times k}$. Because of the hypoellipticity, $R(\Delta t; \Theta)$ is now invertible, but the 0s in C mean that it is highly ill conditioned (or near degenerate) for $0 < \Delta t \ll 1$. Specific examples for the matrix R will be given later.

Ideally we would like to implement the following deterministic scan Gibbs sampler.

(a) Sample Θ from $\mathbb{P}(\Theta|u, v, \Gamma\Gamma^{\mathrm{T}})$.
(b) Sample $\Gamma\Gamma^{\mathrm{T}}$ from $\mathbb{P}(\Gamma\Gamma^{\mathrm{T}}|u, v, \Theta)$.
(c) Sample v from $\mathbb{P}(v|u, \Theta, \Gamma\Gamma^{\mathrm{T}})$.
(d) Restart from step (a) unless sufficiently equilibrated.

In practice, however, approximations to the densities $\mathbb{P}$ will be needed. We refer to expressions of the form (4) as auxiliary models and we shall use them to approximate the exact density on path space, $\mathbb{P}(u, v|\Theta, \Gamma\Gamma^{\mathrm{T}})$, of the path u, v for parameter values Θ and $\Gamma\Gamma^{\mathrm{T}}$. The resulting approximations are denoted $\mathbb{P}_{\mathrm{E}}(U, V|\Theta, \Gamma\Gamma^{\mathrm{T}})$ for the Euler–Maruyama approximation found from expression (3) and $\mathbb{P}_{\mathrm{IT}}(U, V|\Theta, \Gamma\Gamma^{\mathrm{T}})$ for the Itô–Taylor approximation found from expression (4). We again use $\mathcal{L}_{\mathrm{E}}$ and $\mathcal{L}_{\mathrm{IT}}$ in the same way as for the exact distribution $\mathbb{P}$ above when omitting constants of proportionality.

The questions that we address in this paper are as follows.

(a) How does the ill conditioning of the Markov chain $\{x_n\}_{n=0}^N$ affect parameter estimation for $\Gamma\Gamma^{\mathrm{T}}$ and for the last m rows of Θ in the regime $\Delta t \ll 1, N\,\Delta t = T \gg 1$?
(b) In many applications, it is natural that only the smooth data $\{u_n\}_{n=0}^N$ are observed, and not the rough data $\{v_n\}_{n=0}^N$. What effect does the absence of observations of the rough data have on the estimation for $\Delta t \ll 1$ and $N\,\Delta t = T \gg 1$?
(c) The exact likelihood is usually not available; what approximations of the likelihood should be used, in view of the ill conditioning?
(d) How should the answers to these questions be combined to produce an effective Gibbs loop to sample the distribution of parameters $\Theta, \Gamma\Gamma^{\mathrm{T}}$ and the missing data $\{v_n\}_{n=0}^N$?

To tackle these issues, we use a combination of analysis and numerical simulation, based on three model problems which are conceived to highlight issues that are central to the questions above. We shall use analysis to explain why some seemingly reasonable methods fail, and simulation will be used both to extend the validity of the analysis and to illustrate good behaviour of the new method that we introduce.

For the numerical simulations, we shall use either exact discrete time samples of system (1) in simple Gaussian cases, or trajectories that are obtained by Euler–Maruyama simulation of the SDE on a temporal grid with a spacing that is considerably finer than the observation time interval Δt.

In Section 2 we shall introduce our three model problems and in Section 3 we study the performance of $\mathcal{L}_{\mathrm{E}}$ to estimate the diffusion coefficient. Observing and analysing its failure in the case with partial observation leads to the improved statistical model yielding $\mathcal{L}_{\mathrm{IT}}$ which eliminates these problems; we introduce this in Section 4. In Section 5 we show that $\mathcal{L}_{\mathrm{IT}}$ is inappropriate for drift estimation, but that $\mathcal{L}_{\mathrm{E}}$ is effective in this context. In Section 6, the individual estimators will be combined into a Gibbs sampler to solve the overall estimation problem with asymptotically consistent performance being demonstrated numerically. Section 7 contains an application to molecular dynamics and Section 8 provides concluding discussion.

We introduce one item of notation to simplify the presentation. Given an invertible matrix $R \in \mathbb{R}^{n\times n}$ we introduce a new norm using the Euclidean norm on $\mathbb{R}^n$ by setting $\|x\|_R = \|R^{-1}x\|_2$ for vectors $x \in \mathbb{R}^n$.

1.1. Two classical estimators

From previous work on hypoelliptic diffusions, we note a classical estimator for the covariance matrix and for the drift matrix in the linear fully observed case which will be useful for reference later in the paper.

Firstly, it is straightforward to estimate the covariance matrix $\Gamma\Gamma^{\mathrm{T}}$ from the quadratic variation: noting that

$$\frac{1}{T}\sum_{n=0}^{N-1}(v_{n+1}-v_n)(v_{n+1}-v_n)^{\mathrm{T}} \to \Gamma\Gamma^{\mathrm{T}} \qquad \text{as } N\to\infty, \tag{5}$$

with $T = N\,\Delta t$ fixed; see Durrett (1996).

The Girsanov formula gives rise to a maximum likelihood estimator for the lower rows of Θ, and in the linear case, where A is just the identity, the maximum likelihood estimate for the whole of Θ is given by

$$\hat{\Theta} = \left(\int_0^T \mathrm{d}x x^{\mathrm{T}}\right)\left(\int_0^T x x^{\mathrm{T}}\,\mathrm{d}t\right)^{-1}. \tag{6}$$

For the hypoelliptic case, this was proved to be consistent as $T \to \infty$ in Breton and Musiela (1985).

2. Model problems

To study the performance of parameter estimators, we have selected a sequence of three model problems ranging from simple linear stochastic growth through a linear oscillator subject to noise and damping to a non-linear oscillator of similar form. All these problems are second order hypoelliptic and they have a physical background, so we use q (position) and p (momentum) to denote smooth and rough components in the model problems instead of u and v which we used in the general case. Their general form is given as the second-order Langevin equation

$$\begin{aligned} \mathrm{d}q &= p\,\mathrm{d}t, \\ \mathrm{d}p &= \{-\gamma p + f(q; D)\}\,\mathrm{d}t + \sigma\,\mathrm{d}B \end{aligned} \tag{7}$$

where f is some (possibly non-linear) force function parameterized by D and the variables q and p are scalar. The parameters γ, D and σ are to be estimated.

2.1. Model problem I: stochastic growth

Here, $x = (q, p)^{\mathrm{T}}$ satisfies

$$\begin{aligned} \mathrm{d}q &= p\,\mathrm{d}t, \\ \mathrm{d}p &= \sigma\,\mathrm{d}B. \end{aligned} \tag{8}$$

The process has one parameter, the diffusion parameter σ, that describes the size of the fluctuations. In the setting of model (1) we have

$$A(x) = x,$$
$$\Theta = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$
$$C = \begin{pmatrix} 0 \\ \sigma \end{pmatrix}$$

and $u = q$ and $v = p$. The process is Gaussian with mean and covariance

$$\mu(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} q_0 \\ r_0 \end{pmatrix},$$
$$\Sigma(t) = \sigma^2 \begin{pmatrix} t^3/3 & t^2/2 \\ t^2/2 & t \end{pmatrix}.$$

The exact discrete samples may be written as

$$\begin{aligned} q_{n+1} &= q_n + p_n\,\Delta t + \sigma \frac{(\Delta t)^{3/2}}{\sqrt{12}} \zeta_n^{(1)} + \sigma \frac{(\Delta t)^{3/2}}{2} \zeta_n^{(2)}, \\ p_{n+1} &= p_n + \sigma \sqrt{\Delta t}\, \zeta_n^{(2)}, \end{aligned} \tag{9}$$

with

$$\zeta_0 \sim \mathcal{N}\left\{0, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right\}$$

and $\{\zeta_n\}_{n=0}^N$ being IID; individual components of ζ_n are referred to as $\zeta_n^{(1)}$ and $\zeta_n^{(2)}$ respectively. The matrix R from approximation (4) is given here as

$$R = \sigma \begin{pmatrix} \frac{1}{\sqrt{12}}\Delta t & \frac{1}{2}\Delta t \\ 0 & 1 \end{pmatrix}.$$

In the case of this model problem, the auxiliary model (4) is actually exact.

2.2. Model problem II: harmonic oscillator

As our second model problem we consider a damped harmonic oscillator that is driven by a white noise forcing where $x = (q, p)^{\mathrm{T}}$:

$$\begin{aligned} \mathrm{d}q &= p\,\mathrm{d}t, \\ \mathrm{d}p &= -Dq\,\mathrm{d}t - \gamma p\,\mathrm{d}t + \sigma\,\mathrm{d}B. \end{aligned} \tag{10}$$

This model is obtained from the general SDE (1) for the choice

$$\begin{aligned} A(x) &= x, \\ \Theta &= \begin{pmatrix} 0 & 1 \\ -D & -\gamma \end{pmatrix}, \\ C &= \begin{pmatrix} 0 \\ \sigma \end{pmatrix} \end{aligned}$$

and $u = q$ and $v = p$. The process is Gaussian and the mean and covariance of the solution can be explicitly calculated. The matrix R is the same as in model problem I.

2.3. Model problem III: oscillator with trigonometric potential

In the third model problem, $x = (q, p)^{\mathrm{T}}$ describes the dynamics of a particle moving in a potential which is a superposition of trigonometric functions and in contact with a heat bath obeying the fluctuation–dissipation relation; see Lasota and Mackey (1994). This potential is sometimes used in molecular dynamics in connection with the dynamics of dihedral angles—see Section 7. The model is

$$\begin{aligned} \mathrm{d}q &= p\,\mathrm{d}t, \\ \mathrm{d}p &= \left\{ -\gamma p - \sum_{j=1}^{c} D_j \sin(q)\cos^{j-1}(q) \right\} \mathrm{d}t + \sigma\,\mathrm{d}B. \end{aligned} \tag{11}$$

This equation has parameters γ, $D_i, i = 1, \ldots, c$, and σ. It can be obtained from the general SDE (1) for the choice

$$\begin{aligned} A\left\{ \begin{pmatrix} q \\ p \end{pmatrix} \right\} &= \begin{pmatrix} \sin(q) \\ \sin(q)\cos(q) \\ \vdots \\ \sin(q)\cos^{c-1}(q) \\ p \end{pmatrix}, \\ \Theta &= \begin{pmatrix} 0 & \ldots & 0 & 1 \\ -D_1 & \ldots & -D_c & -\gamma \end{pmatrix}, \\ C &= \begin{pmatrix} 0 \\ \sigma \end{pmatrix} \end{aligned}$$

and $u = q$ and $v = p$. No explicit closed form expression for the solution of the SDE is known in this case; the process is not Gaussian. The matrix R in the statistical model (4) is the same as that obtained in model problem I.

3. Euler auxiliary model

As discussed in Section 1, we need to find appropriate approximations for $\mathbb{P}$ in steps (a)–(c) of the desired Gibbs loop. The purpose of this section is to show that use of $\mathbb{P}_{\mathrm{E}}$ in step (c), to sample the missing component of the path, leads to incorrect estimation of the diffusion coefficient. The root cause is the numerical differentiation for the missing path which is implied by the Euler approximation.

3.1. Auxiliary model

If the force function $A(\cdot)$ is non-linear, closed form expressions for the transition density are in general unavailable. To overcome this obstacle, we can use a discrete time auxiliary model. The Euler model (3) is commonly used and we apply it to a simple linear model problem to highlight its deficiencies in the case of partially observed data from hypoelliptic diffusions.

The Euler–Maruyama approximation of the SDE (1) is

$$X_{n+1} = X_n + \Delta t \Theta\, A(X_n) + \sqrt{\Delta t} C \xi_n \tag{12}$$

where $\xi_n \sim \mathcal{N}(0, I)$ is an IID sequence of m-dimensional vectors with standard normal distribution. This corresponds to approximation (4) with $R(\Delta t; \Theta)$ replaced by $R(0; \Theta)$ from approximation (3). Thus we obtain

$$\begin{aligned} U_{n+1} &= U_n + \Delta t \mathcal{P}\Theta\, A(X_n), \\ V_{n+1} &= V_n + \Delta t \mathcal{Q}\Theta\, A(X_n) + \sqrt{\Delta t}\Gamma\xi_n \end{aligned} \tag{13}$$

where now each element of the IID sequence ξ_n is distributed as $\mathcal{N}(0, I)$ in $\mathbb{R}^m$. This model gives rise to the density

$$\mathcal{L}_{\mathrm{ND}}(U, V|\Theta, \Gamma\Gamma^{\mathrm{T}}) = \prod_{n=0}^{N-1} \frac{\exp\{-\frac{1}{2}\|\Delta V_n - \Delta t \mathcal{Q}\Theta\, A(X_n)\|_\Gamma^2\}}{\sqrt{(2\pi|\Gamma\Gamma^{\mathrm{T}}|)}} \delta\left\{\frac{U_{n+1} - U_n}{\Delta t} - \mathcal{P}\Theta\, A(X_n)\right\}. \tag{14}$$

The Dirac mass insists that the data are compatible with the auxiliary model (12), i.e. the V-path must be given by numerical differentiation of the U-path in the case of expression (7), and similar formulae in the general case. To estimate parameters we shall use the expression

$$\mathcal{L}_{\mathrm{E}}(U, V|\Theta, \Gamma\Gamma^{\mathrm{T}}) = \prod_{n=0}^{N-1} \frac{\exp\{-\frac{1}{2}\|\Delta V_n - \Delta t \mathcal{Q}\Theta\, A(X_n)\|_\Gamma^2\}}{\sqrt{(2\pi|\Gamma\Gamma^{\mathrm{T}}|)}}. \tag{15}$$

In the case when the Euler model is used to estimate missing components we assume that $\{U_n\}$ and $\{V_n\}$ are related so that the data are compatible with the auxiliary model—i.e. numerical differentiation is used to find $\{V_n\}$ from $\{U_n\}$.

3.2. Model problem I

The Euler auxiliary model for this model problem is

$$\begin{aligned} Q_{n+1} &= Q_n + P_n\, \Delta t, \\ P_{n+1} &= P_n + \sigma\sqrt{\Delta t}\xi_n. \end{aligned} \tag{16}$$

Here, $\{\xi_n\}$ is an IID $\mathcal{N}(0, 1)$ sequence. The root cause of the phenomena that we discuss in this paper is manifest in comparing expressions (9) and (16). The difference is that the $\mathcal{O}\{(\Delta t)^{3/2}\}$ white noise contributions in the exact time series (9) do not appear in the equation for Q_n.

We shall see that this plays havoc with parameter estimation, even though the Euler method is pathwise convergent.

We assume that observations of the smooth component only, Q_n, are available. In this case the Euler method for estimation (16) gives the formula

$$P_n = \frac{Q_{n+1} - Q_n}{\Delta t} \tag{17}$$

for the missing data. In the following numerical experiment we generate exact data from expression (9) by using the parameter value $\sigma = 1$. We substitute P_n given by equation (17) into equation (15) and find the maximum likelihood estimator for σ in the case of partial observation. In the case of complete observation we use the exact value for $\{P_n\}$, from expression (9), and again use a maximum likelihood estimator for σ from equation (15).

Using $N = 100$ time steps for a final time of $T = 10$ with $\sigma = 1$ the histograms for the estimated diffusion coefficient that are presented in Figs 1(b) and 1(e) are obtained. Figs 1(a)–1(c) contain histograms that were obtained in the case of complete observation where good agreement between the true σ and the estimates is observed. Figs 1(d)–1(f) contain the histograms that were obtained for partial observation by using equation (17). The observed mean value of $\mathbb{E}(\hat{\sigma}) = 0.806$ indicates that the method yields biased estimates. Increasing the final time to $T = 100$ (see Figs 1(a) and 1(d)) or increasing the resolution to $\Delta t = 0.01$ (see Figs 1(c) and 1(f)) does not remove this bias.

Thus we see that, in the case of partial observation, $\hat{\sigma}$ contains $\mathcal{O}(1)$ errors which do not diminish with decreasing Δt and/or increasing $T = N\,\Delta t$.

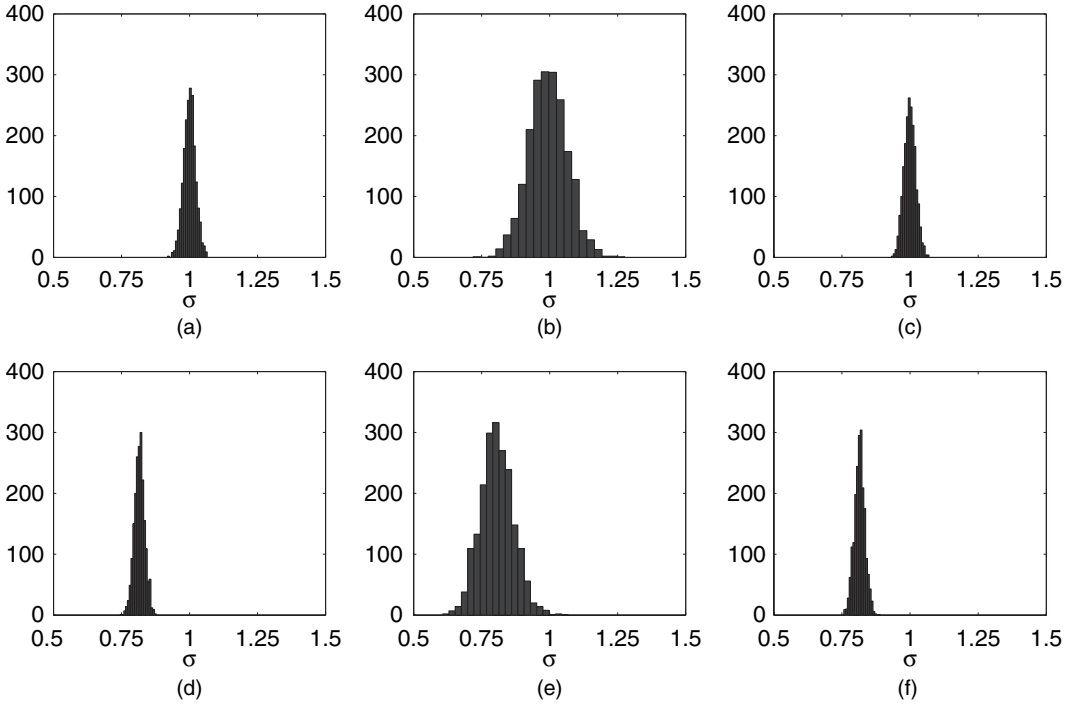

Fig. 1. Maximum likelihood estimates of σ by using the Euler model for model problem I: (a) complete observation, $T = 100$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 1.0001$); (b) complete observation, $T = 10$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 0.9942$); (c) complete observation, $T = 10$, $\Delta t = 0.01$ (■, $\langle\sigma\rangle = 0.99921$); (d) partial observation, $T = 100$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 0.81607$); (e) partial observation, $T = 10$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 0.80636$); (f) partial observation, $T = 10$, $\Delta t = 0.01$ (■, $\langle\sigma\rangle = 0.81491$)

3.3. Analysis of why the missing data method fails

Model problem I can be used to illustrate why this method fails. We first argue that the method works without hidden data. Interpreting equation (15) as a log-likelihood function with respect to σ, we obtain the following expression in the case of stochastic growth:

$$\log\{\mathcal{L}_{\mathrm{E}}(\sigma|Q,P)\} = -2N\log(\sigma) - \frac{1}{\sigma^2\,\Delta t}\sum_{n=0}^{N-1}(\Delta P_n)^2$$

where Δ is the forward difference operator. The maximum of the log-likelihood function gives the maximum likelihood estimate,

$$\hat{\sigma}^2 = \frac{1}{N\,\Delta t}\sum_{n=0}^{N-1}(\Delta P_n)^2. \tag{18}$$

In the case of complete data, expression (9) gives

$$\hat{\sigma}^2 = \frac{\sigma^2}{N}\sum_{n=0}^{N-1}(\zeta_n^{(2)})^2. \tag{19}$$

By the law of large numbers, $\hat{\sigma}^2 \to \sigma^2$ almost surely as $N \to \infty$. This shows that the method works when the complete data are observed.

Let us consider what happens when P is hidden. In this case, P_n is estimated by

$$\hat{P}_n = \frac{Q_{n+1} - Q_n}{\Delta t}.$$

But since q_n is generated by expression (9) we find that

$$\hat{P}_n = \frac{P_{n+1} + P_n}{2} + \sigma\frac{\sqrt{\Delta t}}{\sqrt{12}}\zeta_n^{(1)}$$

and

$$\begin{aligned}\Delta\hat{P}_n &= \frac{\Delta P_{n+1}}{2} + \frac{\Delta P_n}{2} + \sigma\frac{\sqrt{\Delta t}}{\sqrt{12}}(\zeta_{n+1}^{(1)} - \zeta_n^{(1)}) \\ &= \frac{\sigma\sqrt{\Delta t}}{2}\left(\zeta_{n+1}^{(2)} + \zeta_n^{(2)} + \frac{1}{\sqrt{3}}\zeta_{n+1}^{(1)} - \frac{1}{\sqrt{3}}\zeta_n^{(1)}\right).\end{aligned}$$

When $\Delta\hat{P}_n$ is inserted in equation (18) it follows that

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sigma^2}{4N}\sum_{n=0}^{N-1}\left(\zeta_{n+1}^{(2)} + \zeta_n^{(2)} + \frac{\zeta_{n+1}^{(1)} - \zeta_n^{(1)}}{\sqrt{3}}\right)^2 \\ &= \frac{\sigma^2}{4N}\left\{\sum_{n=0}^{N-1}\left(\zeta_{n+1}^{(2)} + \frac{\zeta_{n+1}^{(1)}}{\sqrt{3}}\right)^2 + \sum_{n=0}^{N-1}\left(\zeta_n^{(2)} - \frac{\zeta_n^{(1)}}{\sqrt{3}}\right)^2 + 2\sum_{n=0}^{N-1}\left(\zeta_n^{(2)} - \frac{\zeta_n^{(1)}}{\sqrt{3}}\right)\left(\zeta_{n+1}^{(2)} + \frac{\zeta_{n+1}^{(1)}}{\sqrt{3}}\right)\right\}.\end{aligned}$$

The random variables $\{\zeta_n\}_{n=0}^N$ are IID with $\zeta_0 \sim N(0, I)$. So, by the law of large numbers, $\hat{\sigma}^2 \to \frac{2}{3}\sigma^2$ almost surely as $N \to \infty$. Furthermore, the limits hold in either of the cases where $N\,\Delta t = T$ or Δt are fixed as $N \to \infty$. This means that, independently of what limit is considered, a seemingly reasonable estimation scheme based on Euler approximation results in $\mathcal{O}(1)$ errors in the diffusion coefficient. There is similarity here with work of Gaines and Lyons (1997) showing that adaptive methods for SDEs get the quadratic variation wrong if the adaptive strategy is not chosen carefully.

4. Improved auxiliary model

The Euler auxiliary model fails to propagate noise to the smooth component of the solution and thus leads to estimating missing paths v with incorrect quadratic variation. A new auxiliary model is thus proposed which propagates the noise by using what amounts to an Itô–Taylor expansion, retaining the leading order component of the noise in each row of the equation. The model is used to set up an estimator for the missing path by using a Langevin sampler from path space which is then simplified to a direct sampler in the Gaussian case. Numerical experiments indicate that the method yields the correct quadratic variation for the simulated missing path.

The model is motivated by using our common framework for the model problems I–III, namely expression (7). The improved auxiliary model is based on the observation that in the second row of an Itô–Taylor expansion of expression (7) the drift terms are of size $\mathcal{O}(\Delta t)$ whereas the random forcing term is 'typically' (in root mean square) of size $\mathcal{O}(\sqrt{\Delta t})$. Thus, neglecting the contribution of the drift term in the second row on the first row leads to the following approximation of expression (7):

$$\begin{pmatrix} Q_{n+1} \\ P_{n+1} \end{pmatrix} = \begin{pmatrix} Q_n \\ P_n \end{pmatrix} + \Delta t \begin{pmatrix} P_n \\ f(Q_n) - \gamma P_n \end{pmatrix} + \sigma \begin{pmatrix} \int_{n\Delta t}^{(n+1)\Delta t} \{B(s) - B(n\,\Delta t)\}\,\mathrm{d}s \\ B\{(n+1)\,\Delta t\} - B(n\,\Delta t) \end{pmatrix}.$$

The random vector on the right-hand side is Gaussian and can be expressed as a linear combination of two independent normally distributed Gaussian random variables. Computation of the variances and the correlation is straightforward, leading to the following statistical model:

$$\begin{pmatrix} Q_{n+1} \\ P_{n+1} \end{pmatrix} = \begin{pmatrix} Q_n \\ P_n \end{pmatrix} + \Delta t \begin{pmatrix} P_n \\ f(Q_n) - \gamma P_n \end{pmatrix} + \sigma \sqrt{\Delta t} R \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}. \tag{20}$$

Here, ξ_1 and ξ_2 are independent normally distributed Gaussian random variables and R is given as

$$R = \begin{pmatrix} \Delta t/\sqrt{12} & \Delta t/2 \\ 0 & 1 \end{pmatrix}.$$

This is a specific instance of approximation (4). It should be noted that this model is in agreement with the Itô–Taylor approximation up to error terms of order $\mathcal{O}(\Delta t^2)$ in the first row and $\mathcal{O}(\Delta t^{3/2})$ in the second row and that higher order hypoelliptic processes can be approximated by using a similarly truncated Itô–Taylor expansion. The key important idea is to propagate noise into all components of the system, to leading order.

If complete observations are available, this model performs satisfactorily for estimation of σ. This can be verified analytically for model problem I in the same fashion as in Section 3.3. Numerically, this can be seen from Figs 2(a)–2(c) (referring to complete observation) for model problem I and from Figs 3(a)–3(c) for model problem II. In both cases the true value is given by $\sigma = 1$. See Section 4.2 for a full discussion of these numerical experiments.

If only partial observations are available, however, a means of reconstructing the hidden component of the path must be procured. A standard procedure would be the use of the Kalman filter or smoother (Kalman, 1960; Catlin, 1989), which could then be combined with the expectation–maximization algorithm (Dempster *et al.*, 1977; Meng and van Dyk, 1997) to estimate parameters. In this paper, however, we employ a Bayesian approach sampling directly from the posterior distribution for the rough component p without factorizing the sampling into forward and backward sweeps.

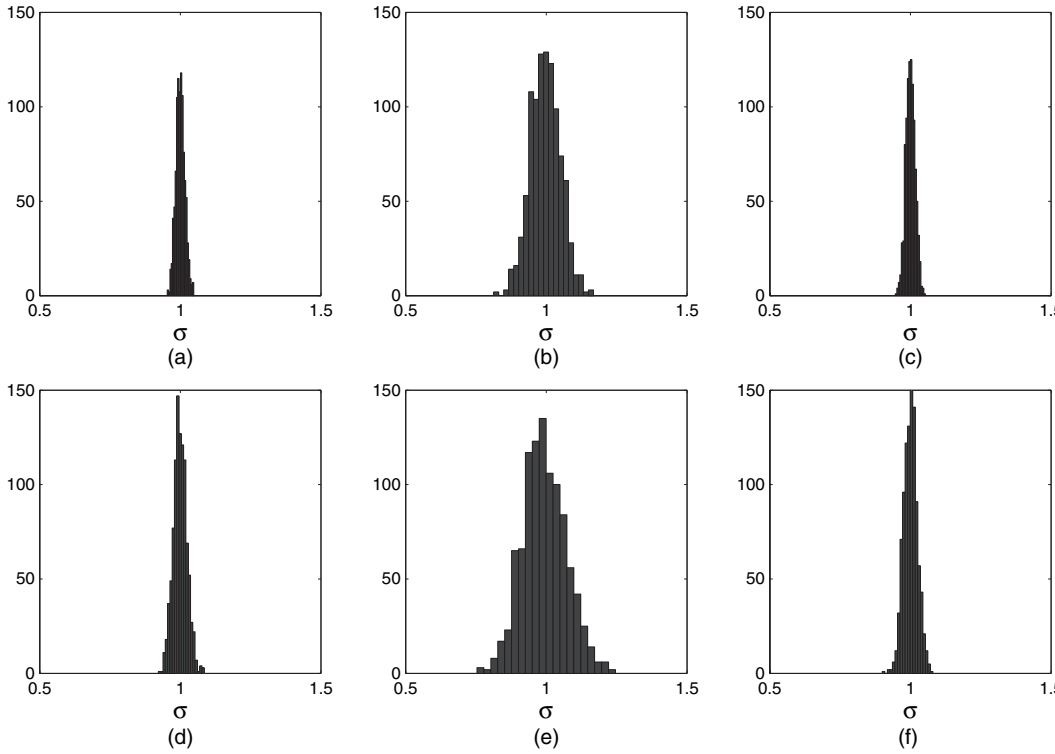

Fig. 2. Estimates of σ by using the $\mathcal{L}_{\mathrm{IT}}$ model for model problem I: (a) maximum likelihood estimates, complete observation, $T = 100$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 1.0002$, standard deviation 0.016077); (b) maximum likelihood estimates, complete observation, $T = 10$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 0.99637$, standard deviation 0.05272); (c) maximum likelihood estimates, complete observation, $T = 10$, $\Delta t = 0.01$ (■, $\langle\sigma\rangle = 1.0002$, standard deviation 0.016538); (d) mean Gibbs estimates, partial observation, $T = 100$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 0.99932$, standard deviation 0.02416); (e) mean Gibbs estimates, partial observation, $T = 10$, $\Delta t = 0.1$ (■, $\langle\sigma\rangle = 0.99333$, standard deviation 0.07741); (f) mean Gibbs estimates, partial observation, $T = 10$, $\Delta t = 0.01$ (■, $\langle\sigma\rangle = 1.0002$, standard deviation 0.024443)

4.1. Path sampling

The logarithm of the density on path space for the missing data induced by the auxiliary model (4) can be written as

$$\log\{\mathcal{L}_{\mathrm{IT}}(p|q,\Theta,\Gamma\Gamma^{\mathrm{T}})\} = -\frac{1}{2}\sum_{l=0}^{N}\|\Delta X_l - \Theta\, A(X_l)\Delta t\|_R^2 + \text{constant}. \tag{21}$$

We shall apply this in the case (20) which is a specific instance of model (4).

One way to sample from the density on path space, $\mathcal{L}_{\mathrm{IT}}(P)$, for rough paths $\{P_i\}_{i=0}^N$ is via the Langevin equation (see section 6.5.2 in Robert and Casella (1999)) and, in general, we expect this to be effective in view of the high dimensionality of P. Other Markov chain Monte Carlo approaches may also be used.

However, when the joint distribution of $\{P_i\}_{i=1}^N$ is Gaussian it is possible to generate independent samples as follows: note first that in the Gaussian case, when $\mathcal{L}_{\mathrm{IT}}$ in equation (21) is quadratic in P, the derivative of $\log(\mathcal{L}_{\mathrm{IT}})$ with respect to the rough path P can be computed explicitly, which was carried out in Pokern (2007). For our oscillator framework, the derivative can be expressed by using a tridiagonal, negative definite matrix P_{mat} with highest order stencil $-1\ -4\ -1$ acting on the P-vector plus a possibly non-linear contribution $\mathcal{Q}(Q)$ acting on the Q-vector only:

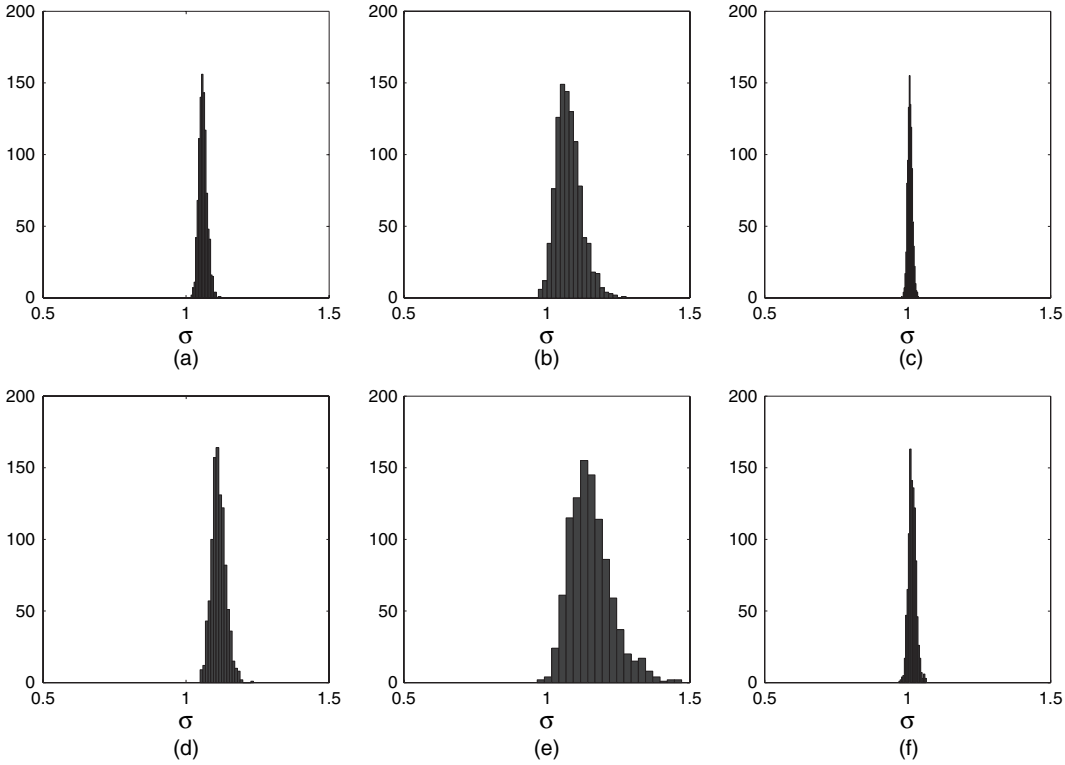

Fig. 3. Estimates of σ by using the $\mathcal{L}_{\mathrm{IT}}$ model for model problem II: (a) maximum likelihood estimates, complete observation, $T = 100$, $\Delta t = 0.02$ (■, $\langle\sigma\rangle = 1.0591$, standard deviation 0.014348); (b) maximum likelihood estimates, complete observation, $T = 10$, $\Delta t = 0.02$ (■, $\langle\sigma\rangle = 1.0768$, standard deviation 0.04359); (c) maximum likelihood estimates, complete observation, $T = 10$, $\Delta t = 0.002$ (■, $\langle\sigma\rangle = 1.0085$, standard deviation 0.0082836); (d) mean Gibbs estimates, partial observation, $T = 100$, $\Delta t = 0.02$ (■, $\langle\sigma\rangle = 1.114$, standard deviation 0.024739); (e) mean Gibbs estimates, partial observation, $T = 10$, $\Delta t = 0.02$ (■, $\langle\sigma\rangle = 1.1538$, standard deviation 0.073624); (f) mean Gibbs estimates, partial observation, $T = 10$, $\Delta t = 0.0002$ (■, $\langle\sigma\rangle = 1.0163$, standard deviation 0.013044)

$$\nabla_p \log\{\mathcal{L}_{\mathrm{IT}}(Q, P)\} = P_{\mathrm{mat}} P + \mathcal{Q}(Q).$$

Then, the suggested direct sampler for P-paths is simply

$$P = -P_{\mathrm{mat}}^{-1}\, \mathcal{Q}(Q) + U^{-1}\xi. \tag{22}$$

Here $U^{\mathrm{T}}U = -P_{\mathrm{mat}}$ is a Cholesky factorization and ξ is a dimension N vector of IID normally distributed random numbers.

4.2. Estimating diffusion coefficient and missing path

The approximation $\mathcal{L}_{\mathrm{IT}}(P, Q|\sigma, \Theta)$ can be used to estimate both the missing path p and the diffusion coefficient σ for our model problems I–III.

To estimate σ, the derivative of the logarithm of $\mathcal{L}_{\mathrm{IT}}$

$$\log\{\mathcal{L}_{\mathrm{IT}}(\sigma|P, Q, \Theta)\} = \log\{\mathcal{L}_{\mathrm{IT}}(P, Q|\sigma, \Theta)\} + \log\{p_0(\Theta, \sigma)\} + \text{constant}$$

(where priors $p_0(\Theta, \sigma)$ are assumed to be given and constants in σ have been omitted) with respect to σ is computed:

$$\frac{\partial}{\partial\sigma}\log(\mathcal{L}_{\mathrm{IT}}) = -\frac{2N}{\sigma} + \frac{1}{\sigma^3}Z + \frac{\partial}{\partial\sigma}\log\{p_0(\Theta,\sigma)\}.$$

Here, we have used the abbreviation

$$Z := \sum_{n=0}^{N-1} \left\| \begin{pmatrix} Q_{n+1} \\ P_{n+1} \end{pmatrix} - \begin{pmatrix} Q_n \\ P_n \end{pmatrix} - \Delta t \begin{pmatrix} P_n \\ f(Q_n) - \gamma P_n \end{pmatrix} \right\|_R^2.$$

In this case no prior distribution was felt necessary as, when $N \to \infty$, its importance would diminish rapidly. Thus we set $p_0 \equiv 1$.

We use a Langevin-type sampler for this distribution. To avoid the singularity at $\sigma = 0$ we use the transformation $\zeta(\sigma) = \sigma^4$. Using the Itô formula, this yields the following Langevin equation which we use to sample ζ and hence σ:

$$d\zeta = \{(12 - 8N)\sqrt{\zeta} + 4Z\}\,\mathrm{d}s + 4\sqrt{2}\zeta^{3/4}\,\mathrm{d}W. \tag{23}$$

A simple explicit Euler–Maruyama discretization in s is used to simulate paths for this SDE. The time step Δs needs to be tuned with N to ensure convergence of the explicit integrator. Since this is a one-dimensional problem, conservatively small time steps and long integration times can be afforded. With such a choice of time step Δs the theoretically possible transient behaviour (see Roberts and Tweedie (1997)) was not observed and we expect accurate samples from the posterior in σ.

This Langevin-type sampler (23) can then be alternated in a systematic scan Gibbs sampler (as described on page 130 of Liu (2001)) using N_{Gibbs} iterations with the direct sampler for the paths, equation (22). This yields estimates of the missing path and the diffusion coefficient which is estimated by averaging over the latter half of the N_{Gibbs} samples. We illustrate this with an example using model problem I with the parameters $\sigma = 1$, $T \in \{10, 100\}$, $\Delta t \in \{0.1, 0.01\}$ and $N_{\mathrm{Gibbs}} = 50$. The sample paths that were used for the fitting are generated by using a subsampled Euler–Maruyama method with temporal grid $\Delta t/k$ where $k = 30$. The resulting histogram of mean posterior estimators is given in Fig. 2 where Figs 2(a)–2(c) correspond to the behaviour when complete observations are available and Figs 2(d)–2(f) correspond to only the smooth component being observed and missing data being sampled according to equation (22). For model problem II we use the parameters $\sigma = 1, D = 4, \gamma = 0.5$, $T \in \{10, 100\}, \Delta t \in \{0.02, 0.002\}$ and $N_{\mathrm{Gibbs}} = 50$. The sample paths that are used for the fitting are generated as for model problem I and the experimental results are given in Fig. 3.

It appears from Figs 2 and 3 that the estimator for this joint problem performs well for model problems I and II for Δt sufficiently small and T sufficiently large. A more careful investigation of the convergence properties is postponed to Section 6 when drift estimation will be incorporated in the procedure.

5. Drift estimation

5.1. Overview

With the approximations $\mathcal{L}_{\mathrm{E}}$ and $\mathcal{L}_{\mathrm{IT}}$ in place, the question arises which of these should be used to estimate the drift parameters. Using model problem II we numerically observe that an $\mathcal{L}_{\mathrm{E}}$-based maximum likelihood estimator performs well. In contrast, ill conditioning due to hypoellipticity leads to error amplification and affects the performance of the $\mathcal{L}_{\mathrm{IT}}$-based maximum likelihood estimator.

5.2. Drift parameters from $\mathcal{L}_{\mathrm{E}}$

To simplify analysis, we illustrate the estimator by using model problems II, expression (10), and III, expression (11). For the latter, the Euler auxiliary model is

$$\begin{aligned} Q_{n+1} &= Q_n + \Delta t P_n, \\ P_{n+1} &= P_n - \Delta t \sum_{i=1}^{c} D_i \, f_i(Q_n) - \Delta t \gamma P_n + \sqrt{\Delta t}\sigma \xi_n, \end{aligned} \tag{24}$$

where we abbreviated the trigonometric expressions using $f_j(q) = \sin(q)\cos^{j-1}(q)$. The functional $\mathcal{L}_{\mathrm{E}}$ in this case is given by

$$\mathcal{L}_{\mathrm{E}}(\gamma, D|Q, P, \sigma) \propto \exp\left[- \sum_{n=0}^{N-1} \frac{\left\{ \Delta P_n + \Delta t \sum_{i=1}^{c} D_i \, f_i(Q_n) + \Delta t \gamma P_n \right\}^2}{2\,\Delta t \sigma^2} \right], \tag{25}$$

Clearly, this posterior is Gaussian with distribution

$$\hat{\Theta} \sim \mathcal{N}(M_{\mathrm{E}}^{-1} b_{\mathrm{E}}, M_{\mathrm{E}}^{-1}), \tag{26}$$

where the matrix M_{E} and the vector b_{E} can be read off from expression (25).

5.3. Drift parameters from $\mathcal{L}_{\mathrm{IT}}$

As the approximate model based on $\mathcal{L}_{\mathrm{IT}}$ is observed to resolve the difficulty with estimating σ for hidden p-paths, it is interesting to see whether it can also be used to estimate the drift parameters.

The logarithm of the density on path space up to an additive constant is given by equation (21). To illustrate the problems arising from the use of $\mathcal{L}_{\mathrm{IT}}$ we use model problem II, so that equation (21) becomes

$$\log\{\mathcal{L}_{\mathrm{IT}}(\Theta|Q, P, \sigma)\} = \frac{1}{2\,\Delta t} \sum_{n=0}^{N-1} \|(\Delta X_n - \Delta t \Theta \, A(X_n))\|_R^2 + \text{constant} \tag{27}$$

where

$$R = \sigma \begin{pmatrix} \Delta t/\sqrt{12} & \Delta t/2 \\ 0 & 1 \end{pmatrix},$$

irrelevant constants have been omitted and we have

$$\begin{aligned} A\left\{ \begin{pmatrix} Q_n \\ P_n \end{pmatrix} \right\} &= \begin{pmatrix} Q_n \\ P_n \end{pmatrix}, \\ \Theta &= \begin{pmatrix} 0 & 1 \\ -D & -\gamma \end{pmatrix}. \end{aligned}$$

To obtain a maximum likelihood estimator from this, we take the derivative with respect to the parameters D and γ and equate to 0. This yields the following linear system:

$$\begin{pmatrix} \sum_n Q_n^2 \, \Delta t & \sum_n P_n Q_n \, \Delta t \\ \sum_n P_n Q_n \, \Delta t & \sum_n P_n^2 \, \Delta t \end{pmatrix} \begin{pmatrix} \hat{D} \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} -\sum_n Q_n \, \Delta P_n \\ -\sum_n P_n \, \Delta P_n \end{pmatrix} + \begin{pmatrix} \sum_n \frac{3}{2} Q_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \\ \sum_n \frac{3}{2} P_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \end{pmatrix}. \tag{28}$$

Comparing this linear system with the mean of the successful estimator (26) we note the presence of an additional term on the right-hand side. This term leads to the failure of the above estimator. Thus, $\mathcal{L}_{\mathrm{IT}}$ is not an appropriate approximation for use in step (a) of the Gibbs sampler.

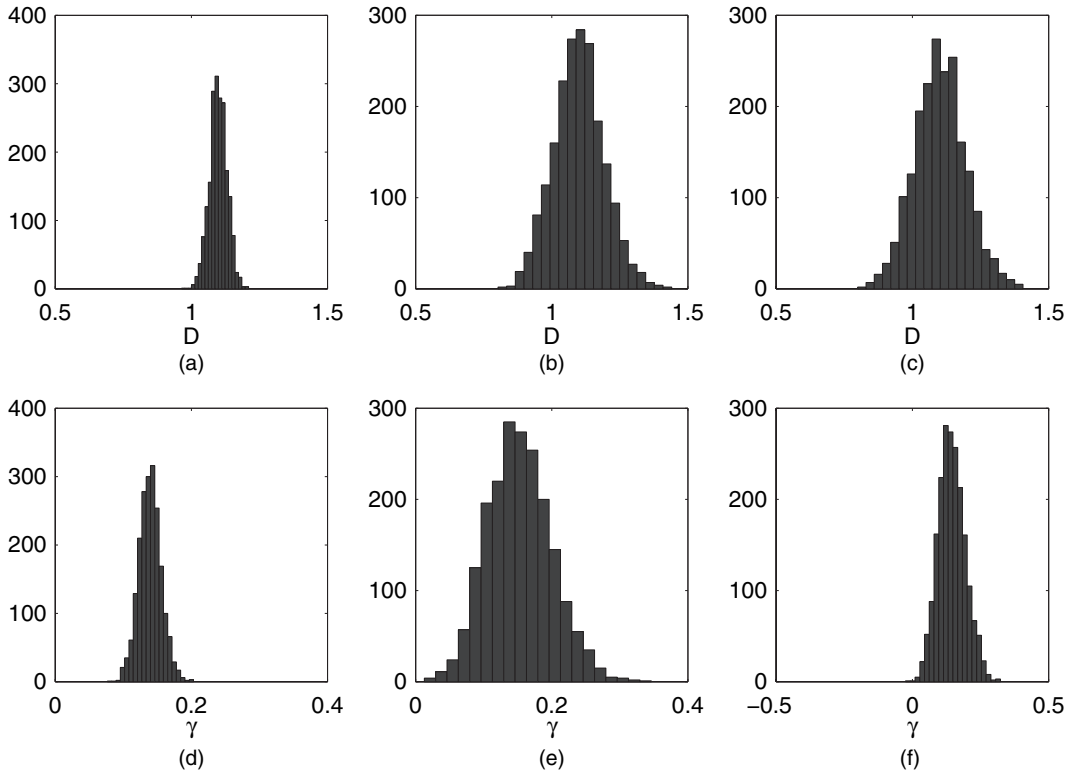

Fig. 4. Maximum likelihood drift estimates for model problem II, by using $\mathcal{L}_{\mathrm{IT}}$: (a) $T = 1000$, $\Delta t = 0.01$ (■, $\langle D \rangle = -1.0989$); (b) $T = 100$, $\Delta t = 0.01$ (■, $\langle D \rangle = -1.0993$); (c) $T = 100$, $\Delta t = 0.001$ (■, $\langle D \rangle = -1.1015$); (d) $T = 1000$, $\Delta t = 0.01$ (■, $\langle \gamma \rangle = -0.13925$); (e) $T = 100$, $\Delta t = 0.01$ (■, $\langle \gamma \rangle = -0.15268$); (f) $T = 100$, $\Delta t = 0.001$ (■, $\langle \gamma \rangle = -0.14457$)

5.4. Numerical check: drift

There are two factors influencing convergence: T and Δt. To illustrate their influence, consider the following series of numerical tests. All the tests share the parameters $D = 4, \gamma = 0.5, \sigma = 0.5$ and $k = 30$. Data for the tests are again generated by using an Euler–Maruyama method on a finer temporal grid with resolution $\Delta t/k$. Figs 4(a)–4(c) contain histograms for the maximum likelihood estimate for the drift parameter D whereas Figs 4(d)–4(f) contain histograms for the drift parameter γ in any case using the full sample path for maximum likelihood inference, i.e. formula (28). It is clear from the experiments summarized in Fig. 4 that both D and γ are grossly underestimated by $\hat{D}$ and $\hat{\gamma}$ from equation (28). This problem does not resolve for smaller Δt (see Figs 4(c) and 4(f)); it does not disappear for longer intervals of observation, either, as can be inferred from Figs 4(a) and 4(d).

5.5. Why the $\mathcal{L}_{\mathrm{IT}}$ model fails for the drift parameters

The key is to compare equation (28) with the mean in distribution (26). This reveals that the last term in equation (28) is an error term which we now study.

Using the second-order Itô–Taylor approximation

$$X_{n+1} = X_n + \Delta t A X_n + \begin{pmatrix} 1 & 0 \\ -\gamma & 1 \end{pmatrix} R \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \frac{1}{2} \Delta t^2 A^2 X_n + \mathcal{O}(\Delta t^{5/2})$$

we can compute the second term on the right-hand side of equation (28):

$$\begin{pmatrix} \sum_n \frac{3}{2} Q_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \\ \sum_n \frac{3}{2} P_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \end{pmatrix} = \begin{pmatrix} -\frac{3}{4}\gamma \sum_n Q_n P_n \, \Delta t - \frac{3}{4} D \sum_n Q_n^2 \, \Delta t \\ -\frac{3}{4} D \sum_n Q_n P_n \, \Delta t - \frac{3}{4}\gamma \sum_n P_n^2 \, \Delta t \end{pmatrix} + I_s + \mathcal{O}(\Delta t). \tag{29}$$

Here, D and γ refer to the exact drift parameters that are used to *generate* the sample path, whereas $\hat{D}$ and $\hat{\gamma}$ in equations (28) and (29) are the drift parameters that are estimated by using the improved auxiliary model. The term I_s on the right-hand side contains stochastic integrals whose expected value is 0.

As the mean error terms can be written in terms of the matrix elements themselves, equation (29) can be substituted in equation (28) to obtain

$$\mathbb{E}(\hat{D}) = \tfrac{1}{4} D + \mathcal{O}(\Delta t), \tag{30}$$

$$\mathbb{E}(\hat{\gamma}) = \tfrac{1}{4}\gamma + \mathcal{O}(\Delta t). \tag{31}$$

This seems to be corroborated by the numerical tests.

5.6. Conclusion for drift estimation

We observed numerically but do not show here that $\mathcal{L}_{\mathrm{E}}$ associated with an Euler model for the SDE (1) yields asymptotically consistent Langevin and maximum likelihood estimators for model problem II.

Although it is aesthetically desirable to base the estimation of all parameters as well as the missing data on the same approximation $\mathcal{L}_{\mathrm{IT}}$ of the true density (up to multiplicative constants) $\mathcal{L}$ and, although this approximation was found to work well for the estimation of missing data and the diffusion coefficient, it does not work for the drift parameters.

It is possible to trace this failure to the fact that only the second row of Θ is estimated where $\mathcal{O}(\Delta t)$ errors in the first row become amplified to $\mathcal{O}(1)$ errors in the second row. Estimating all entries of Θ, although being outside the specification of the problem under consideration, also yields $\mathcal{O}(1)$ errors if $\mathcal{L}_{\mathrm{IT}}$ is used and so does not remedy the problem. This problem is not shared by the discretized version of the diffusion-independent estimator (6), but this is not a maximum likelihood estimator for $\mathcal{L}_{\mathrm{IT}}$.

In summary, for the purposes of fitting our model problems to observed data we employ the Euler auxiliary model (25) for the drift parameters.

6. The Gibbs loop

In this section, we combine the insights that were obtained in previous sections to formulate an effective algorithm to fit hypoelliptic diffusions to partial observations of data at discrete times. We apply a deterministic scan Gibbs sampler alternating between missing data (the rough component of the path, v), drift parameters and diffusion parameters.

We combine the approximations that were developed and motivated in previous sections in the following Gibbs sampler.

(a) Sample Θ from $\mathbb{P}_{\mathrm{E}}(\Theta|U, V, \sigma)$.
(b) Sample σ from $\mathbb{P}_{\mathrm{IT}}(\sigma|U, V, \Theta)$.
(c) Sample v from $\mathbb{P}_{\mathrm{IT}}(V|U, \Theta, \sigma)$.
(d) Restart from step (a) unless sufficiently equilibrated.

Our numerical results will show that this judicious combination of approximations results in an effective algorithm. Theoretical justification remains an interesting open problem.

When applied to model problem III the detailed algorithm (algorithm 1) reads as follows.

Given observations Q_i, $i = 1, \ldots, N$, the initial P-path is obtained by using numerical differentiation:

$$P_i^{(0)} = \frac{\Delta Q_i}{\Delta t}. \tag{32}$$

The initial drift parameter estimate is just set to 0: $\{D_j^{(0)}\}_{j=1}^c = 0$; $\gamma^{(0)} = 0$. Then start the Gibbs loop.

For $k = 1, \ldots, N_{\text{Gibbs}}$:

(a) estimate the drift parameters $\gamma^{(k)}$ and $\{D_j^{(k)}\}_{j=1}^c$ by using sampling based on $\mathcal{L}_{\text{E}}$ given $\{P_i^{(k-1)}\}_{i=0}^N$ via distribution (26);
(b) estimate the diffusivity $\sigma^{(k)}$ by using the Langevin sampler (23) based on $\mathcal{L}_{\text{IT}}$ given $\{P_i^{(k-1)}\}_{i=0}^N$ and $\gamma^{(k)}$, $\{D_j^{(k)}\}_{j=1}^c$;
(c) obtain an independent sample of the P-path, $\{P_i^{(k)}\}_{i=0}^N$ by using equation (22) derived from $\mathcal{L}_{\text{IT}}$ given parameters $\gamma^{(k)}$, $\{D_j^{(k)}\}_{j=1}^c$ and $\sigma^{(k)}$.

We test this algorithm numerically where sample paths of expression (11) are generated by using a subsampled Euler–Maruyama approximation of the SDE. The data are generated by using a time step that is smaller than the observation time step by a factor of either $k = 30$ or $k = 60$. Comparing the results for these two and other non-reported cases, they are found not to depend on the rate of subsampling, k, if this is chosen sufficiently large. The parameters that were used for these simulations are $D_0 = 1, D_1 = -8, D_2 = 8, \gamma = 0.5, \sigma = 0.7, T = 500, \Delta t \in \{1/2, \ldots, 1/128\}$ and $N_{\text{Gibbs}} = 50$. The trigonometric potential resulting from this choice of drift parameters is depicted in Fig. 5(a) and a typical sample path for q is given in Fig. 5(b). It should be noted that all sample paths are started at $(q, p) = (1, 1)$.

The performance of the Gibbs sampler for the sample q-path that is given in Fig. 5 is shown in Fig. 6 where 100 Gibbs steps sampling from the posterior distribution of drift and diffusion parameters are shown for the set-up that is shown above except that here $N_{\text{Gibbs}} = 100$ and $\Delta t = 0.01$. Mean posterior estimators are computed averaging over the latter half of N_{Gibbs} iterations as before. This sampling is repeated up to 64 000 times and we label the repeated sampling average of these mean posterior estimators as $\langle \hat{D}_i \rangle$ and $\langle \hat{\gamma} \rangle$. We then compute their deviation from the true values, $\Delta D_i = \langle \hat{D}_i \rangle - D_i$, and plot ΔD_i and $\Delta\gamma$ against Δt in a doubly logarithmic plot given in Fig. 7.

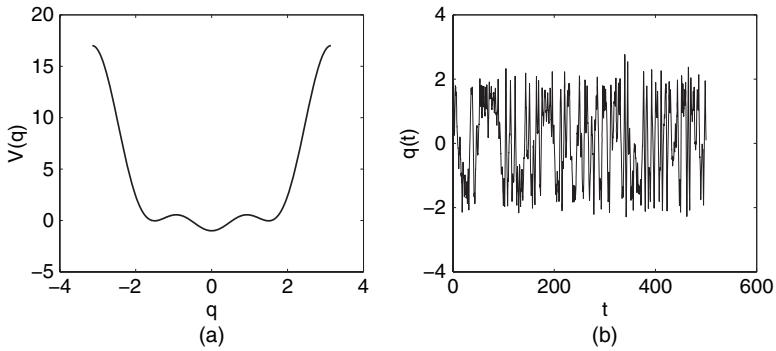

Fig. 5. Typical sample path for model problem III, $T = 500$: (a) trigonometric potential; (b) typical q-path

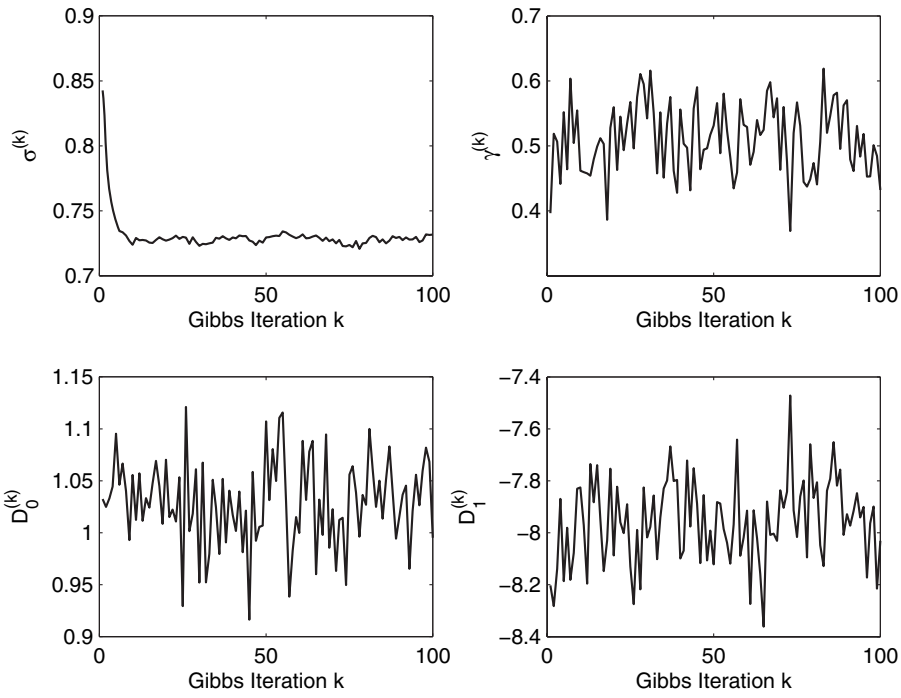

Fig. 6. Model problem III: burn-in of the Gibbs sampler

We seek to fit a straight line to the ΔD_i in a doubly logarithmic plot to ascertain the order of convergence. Since a standard least squares fit proves inadequate, we employ the following procedure.

Given averaged numerically observed parameter estimates y_i and their numerically observed Monte Carlo standard deviations α_i obtained at time steps Δt_i we fit b and c in the linear regression

$$\alpha_i \xi_i = y_i - b - c\,\Delta t_i. \tag{33}$$

Assuming that the errors ξ_i are normally distributed (which is empirically found to be so) a maximum likelihood fit for the parameters b and c can be performed and yields the asymptotic (for $\Delta t \to 0$) drift parameter values that are reported in Fig. 7. Note that this fit constrains the slope of the fitted line in the doubly logarithmic plot to 1. This is to minimize the number of parameters fitted and to improve the accuracy of the extrapolated value b which is the predicted value for y at $\Delta t = 0$. It can be observed in Fig. 7 that this leads to good agreement with the observed average parameter values y_i, and this corroborates the estimator's bias being of order $\mathcal{O}(\Delta t)$.

Comparing the results for the two final times tested, $T = 50$ and $T = 500$, we find that the deviation of the asymptotic drift parameter (b in equation (33)) from the true parameter value is consistent with it being $\mathcal{O}(1/T)$. This error is attributed to all sample paths having been started at $(q, p) = (1, 1)$ rather than from a point that was sampled from the equilibrium measure.

For the diffusion parameter σ, results analogous to those in Fig. 7, using the same parameter values, are shown in Fig. 8 (although Fig. 8 displays results for $k = 30$ only). Asymptotic consistency can be observed from Fig. 8 with a naive least squares fit yielding a slope of $\mathcal{O}(\Delta t^{0.93})$. This is consistent with an $\mathcal{O}(\Delta t)$ error in the estimated diffusion parameter.

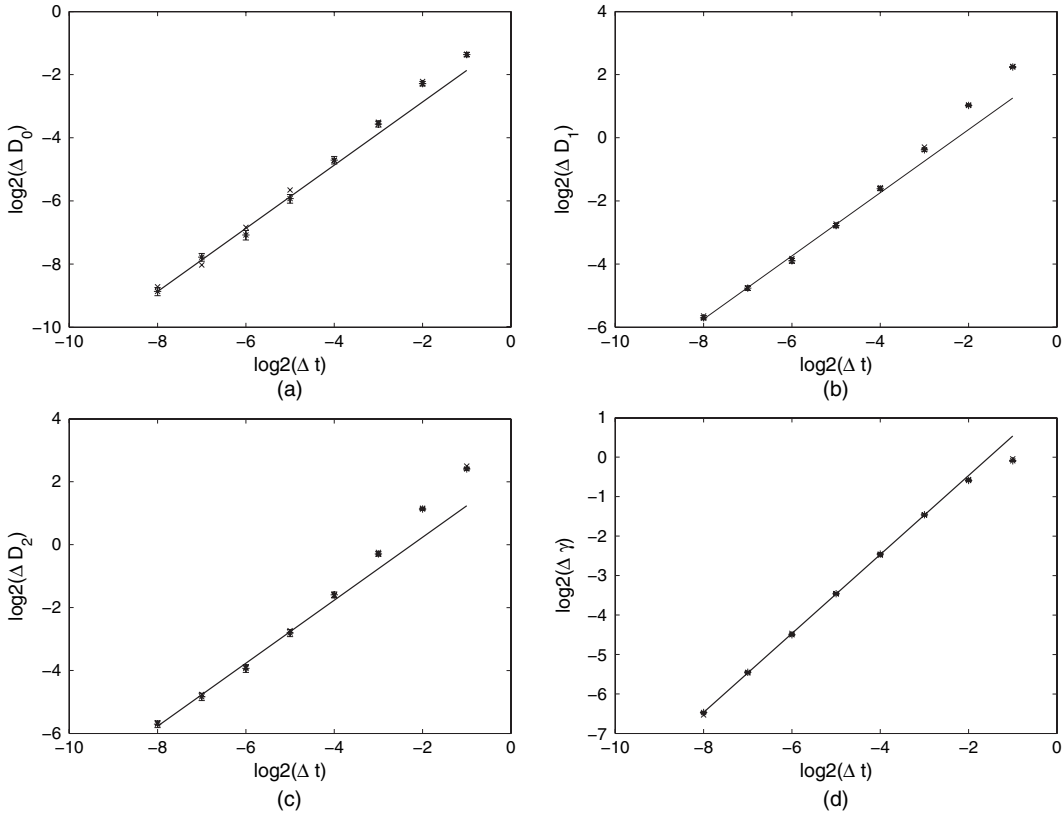

Fig. 7. Model problem III, $T = 500$—averaged mean posterior deviations of the drift parameters ($*$, mean value $k = 30$ with standard deviation; $\times$, mean value $k = 60$): (a) drift parameter D_1 (———, maximum likelihood fit $b = -1.0018$); (b) drift parameter D_2 (———, maximum likelihood fit $b = 8.0101$); (c) drift parameter D_3 (———, maximum likelihood fit $b = -8.0055$); (d) drift parameter γ (———, maximum likelihood fit $b = -0.50457$)

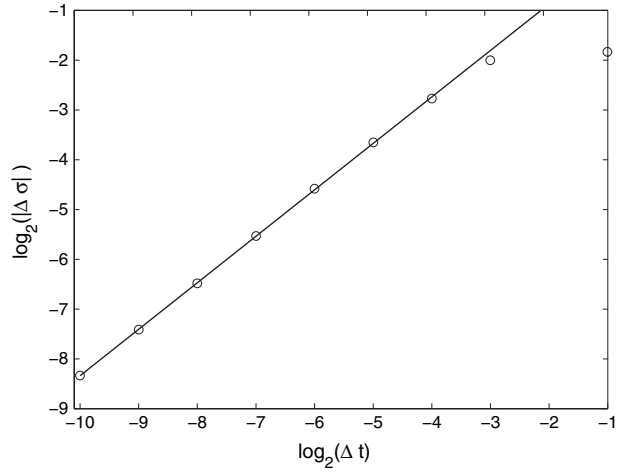

Fig. 8. Model problem III, $T = 500$—averaged mean posterior deviations for σ ($\circ$, $\langle\sigma\rangle$; ———, least squares fit, slope 0.93333)

From these considerations it is apparent that the numerical experiments' outcome is consistent with an $\mathcal{O}(\Delta t) + \mathcal{O}(1/T)$ bias, so algorithm 1 is numerically observed to be an asymptotically unbiased estimator of the drift and diffusion parameters in the cases that were studied.

7. Application to molecular conformational dynamics

As an application of fitting hypoelliptic diffusions by using partial observations we consider data arising from molecular dynamics simulations of a butane molecule by using a simple heat bath approximation.

By considering the origin of the data we demonstrate that it is natural to fit a hypoelliptic diffusion process which yields convergent results for diminishing intersample intervals Δt. Also, stabilization of the fitted force function $f(q) = \Sigma_{j=1}^{c} D_j \, f_j(q)$ as the number of terms to be included, c, increases, is observed. Thus algorithm 1 is shown to be effective on molecular dynamics data. It is also clear, though, that the resulting fit has only limited predictive abilities as it fails to fit the invariant measure of the data at all well. However, this is a *modelling* issue which is not central to this paper.

7.1. Molecular dynamics

The data that are used for this fitting example are generated by using a molecular dynamics simulation for a single molecule of butane. To avoid explicit computations for solvent molecules, several *ad hoc* approximate algorithms have been developed in molecular dynamics. One of the more sweeping approximations that is nonetheless fairly popular, at least as long as electrostatic effects of the solvent can be neglected or treated otherwise, is Langevin dynamics. Here, the time evolution of the Cartesian co-ordinates of the four extended atoms of butane (Fig. 9) is simulated by using a damped driven Hamiltonian system; details of the force field that was used can be found in Brooks (1983).

From a chemical point of view interest is focused on the dihedral angle ω, which is the angle between the two planes in $\mathbb{R}^3$ that is formed by atoms 1, 2 and 3, and atoms 2, 3 and 4; see the sketch in Fig. 9. Conformational change is manifest in this angle, and the Cartesian co-ordinates themselves are of little direct chemical interest. Hence it is natural to try to describe the stochastic dynamics of the dihedral angle in a self-contained fashion.

One molecular dynamics run is produced by using a time step of $\Delta t = 0.1$ fs (throughout this section, we use the time unit femtosecond; 1 fs $= 10^{-15}$ s) and a Verlet variant (see page 435 in Schlick (2000)) covering a total time of $T = 4 \times 10^{-9}$ s (4 ns). A section of the path of the dihedral angle as a function of time can be seen in Fig. 10(a); the corresponding histogram for the whole of the path is depicted in Fig. 10(b).

It should be stressed that the Itô process governing the behaviour of the dihedral angle ω is *not* of the form (11); in particular, it will have a non-constant diffusivity σ. So, fitting to these data

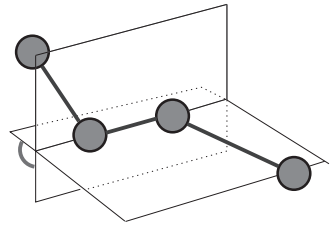

Fig. 9. Sketch of the dihedral angle

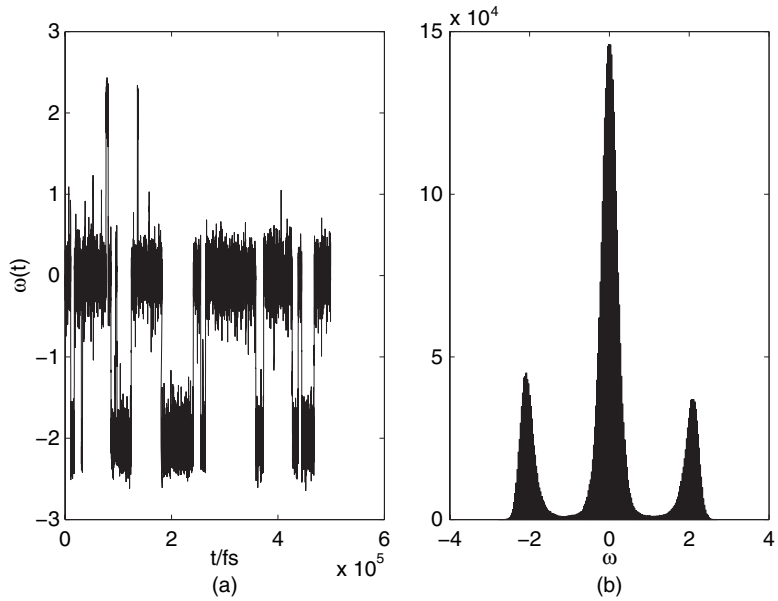

Fig. 10. Molecular dynamics sample path for butane: (a) first 500 ps of the sample path; (b) histogram of the whole sample path (■, $N = 4 \times 10^6$)

tests the robustness of the fitting algorithm in a way that the experiments in previous sections did not.

7.2. Fitting

We aim to fit the process from model problem III, equation (11), to a subsampled trajectory of $\omega(t_i)$ (viewed as the smooth component q) obtained from the molecular dynamics simulation that was described previously. Subsampling is performed because we have a profusion of data and because the hypoelliptic diffusion is expected to be a good fit only at some timescales.

The simulation that was used to obtain the dihedral angle data is such that $\omega(t)$ will be a C^1-function of time assuming a suitable interpretation of the periodicity in ω, so it is natural to fit a hypoelliptic process of damped driven Hamiltonian form.

The physical time units in seconds are minuscule and do not lead to estimated SDE parameters of order 1. It transpires that, to obtain parameter values of order 1, rescaling time so that the final time becomes $T = 80000$ is a good choice. This rescaling is useful in comparing convergence properties with what was observed in Section 6. To assess consistency, the molecular dynamics data are subsampled, at time steps $\Delta t \in \{1\text{ fs}, 2\text{ fs}, 3\text{ fs}\ldots\}$ in physical time units, corresponding to $\{0.02k\}_{k\in\mathbb{N}}$ in the rescaled time units. Algorithm 1 is then run for $N_{\text{Gibbs}} = 40$ outer iterations on each path, using a potential *ansatz*

$$V(\omega;\Theta) = \sum_{k=1}^{c} \Theta_k \cos^k(\omega)$$

which corresponds to the force functions in expression (11) setting $D_k = k\Theta_k$ and $f = V'$; the values $c \in \{3, 5, 7\}$ are used in what follows. These periodic *ansatz* functions are a natural choice for dihedral angle potentials; in fact, the dihedral angle potential that was given in Brooks (1983) is of this form. The drift parameter estimates obtained under subsampling at time step Δt can be seen from Fig. 11 in the case $c = 5$. In Fig. 11, the sampling time step Δt is the abscissa

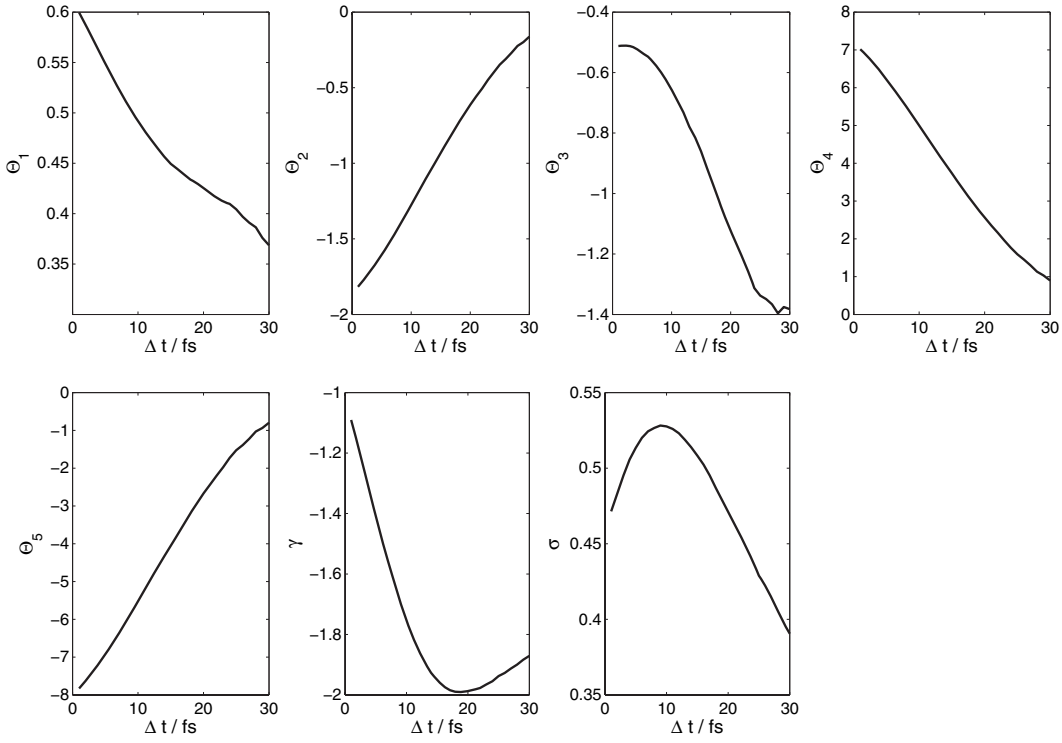

Fig. 11. Convergence for fitted molecular dynamics path with subsampling: mean Gibbs estimates of the drift and diffusion parameters as a function of subsampling interval Δt

and the drift and diffusion parameter estimates ($\Theta_1, \ldots, \Theta_5$, γ and σ) that are obtained from fitting to the sample path subsampled at time step Δt are shown as the ordinate. Fig. 11 shows the behaviour of the drift and diffusion parameter estimates averaged over $N_{\text{Gibbs}} = 100$ Monte Carlo samples $\theta_1, \ldots, \theta_5$ and γ for various values of the subsampling rate. The behaviour as $k \to 0$ indicates that the fitted parameter values converge to a well-defined limit; σ in particular varies relatively little over a large range of subsampling rates. This suggests that the algorithm proposed can fit model problem III to molecular dynamics data. The fact that different (especially drift) parameter values are obtained at different subsampling rates indicates limitations in the fit to model problem III and this will be addressed in the next subsection.

7.3. Limitations

The desirable convergence properties of the algorithm in Δt and T should not be confused with inference about whether fitting this kind of model to this kind of molecular dynamics data gives a good or a bad fit; it merely indicates that, using the algorithm that is suggested in this paper, it is possible to perform such fitting.

To show limitations of the model in this particular application we focus on the implied invariant density of the fitted SDEs, since this object is of interest in computational chemistry. Thus, we consider the push forward of the posterior measure for the parameters D_i, γ and σ onto the set of probability densities on the real line. We can then consider the mean and variance of these densities at any point in $\mathbb{R}$. To do this, we convert the posterior drift parameter samples $\{D_j^{(m)}\}_{j=1}^c$ that are obtained at step m using input data subsampled at rate $k = 1$ to an

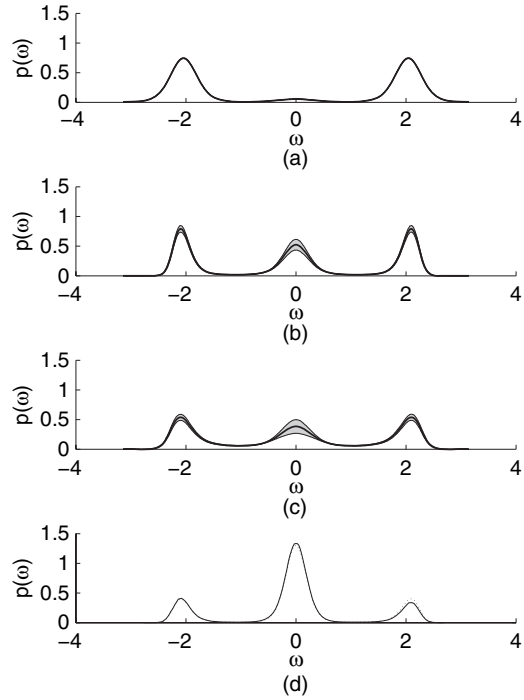

Fig. 12. Probability density functions from fitted potentials for various orders of trigonometric potential (■, posterior variance; ———, empirical density; · · · · · · ·, analytical density): (a) $c = 3$; (b) $c = 5$; (c) $c = 7$; (d) empirical probability density function, butane

invariant density $\rho^{(m)}$ which is specified by its values on an equidistant grid on the interval $[-\pi, \pi]$. These densities for $m \in \{1, \ldots, 1000\}$ are then averaged and their standard deviation is computed pointwise on the grid. This results in Fig. 12. There, we display results for three orders of trigonometric potential c to be fitted. These are contrasted with the empirically observed invariant density and the density arising from the classic canonical thermodynamic ensemble which is proportional to $\exp\{-V(\omega)/kT\}$ which are given in Fig. 12(d). For the force field that was used in the molecular dynamics simulation, it is known that the latter two agree in the limit $T \to \infty$; see Fischer (1997).

It should be stressed that, in each of these experiments, convergence diagnostics indicate convergence of the Gibbs sampler and the posterior distributions for the drift and diffusion parameters are very concentrated and hence posterior variances both for the drift and diffusion parameters as well as the induced invariant densities are low.

With increasing polynomial order c we find some qualitative change in the resulting invariant density and also (in particular moving from $c = 5$ to $c = 7$) a marked increase in posterior variance. This goes hand in hand with a marked increase in the condition number of the drift parameter matrix M_{E} in distribution (26). It is simply an ill-conditioned problem to derive increasingly higher order polynomial coefficients from a fixed length of observed path.

It is observed that, even though the empirically observed invariant density is smooth and close to the thermodynamical expectation, the fitted potentials induce an SDE whose invariant measure is not a good approximation of the empirical density. This may simply be attributed to the fact that the SDE that is being fitted does not represent a good model of the *dynamics* of the dihedral angle in the butane molecule with second-order Langevin heat bath model.

8. Conclusions

A hybrid algorithm for fitting drift and diffusion parameters of a hypoelliptic diffusion process, with constant diffusivity, from observation of smooth data at discrete times has been described. The method combines a Gibbs sampler together with differing approximate likelihoods employed in different steps of the Gibbs loop. Its performance has been validated numerically for several test cases and an application to molecular dynamics data has been given. Although parameter fitting can be viewed as an inverse problem for SDE solvers—and thus ill conditioning of some kind is always to be expected—a detailed understanding of the particular ill conditioning that is induced by hypoellipticity and partial observation has been attained.

Although only second-order hypoelliptic problems have been treated in this paper, the algorithm's applicability is expected to encompass order k hypoelliptic problems and it has been tested successfully on a third-order example. Furthermore, non-linear p-dependence in example (7) can be dealt with by using a Langevin sampler for the missing path and this has also been tested. Additionally, observations that are not exactly equispaced can also be processed provided that the maximal intersample time is sufficiently small.

Further avenues of investigation include the use of imputed data points between samples to diminish $\mathcal{O}(\Delta t)$ errors; however, there is a risk of bad mixing as σ is determined by the small-scale behaviour of the process which would then be dominated by the imputed data points. This has been analysed in the case of elliptic diffusion processes in Roberts and Stramer (2001) and an application of standard estimators to this problem in the hypoelliptic case was given in Godsill and Yang (2006).

Also, an extension to position-dependent diffusion coefficients may prove useful; in particular, it may render the algorithm more useful in molecular dynamics contexts such as those in Hummer (2005).

Acknowledgements

The authors express their gratitude to the referees and the Joint Editor for their careful reading of the paper and for their constructive suggestions.

References

Breton, A. L. and Musiela, M. (1985) Some parameter estimation problems for hypoelliptic homogeneous gaussian diffusions. *Seq. Meth. Statist.*, **22**, 337–356.

Brooks, B. R. (1983) Charmm: a program for macromolecular energy, minimization and dynamics calculations. *J. Computnl Chem.*, **4**, 187–217.

Catlin, D. E. (1989) *Estimation, Control and the Discrete Kalman Filter*. New York: Springer.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

Durrett, R. (1996) *Stochastic Calculus—a Practical Introduction*. London: CRC Press.

Fischer, A. (1997) Die hybride Monte-Carlo-Methode in der Molekülphysik. *Diplomarbeit*. Frei Universität, Berlin.

Gaines, J. G. and Lyons, T. J. (1997) Variable step size control in the numerical solution of stochastic differential equations. *SIAM J. Appl. Math.*, **57**, 1455–1484.

Gardiner, C. W. (1985) *Handbook of Stochastic Methods*. New York: Springer.

Giannopoulos, P. and Godsill, S. J. (2001) Estimation of car processes observed in noise using bayesian inference. In *Proc. Int. Conf. Acoustics, Speech and Signal Processing*. New York: Institute of Electrical and Electronics Engineers.

Godsill, S. and Yang, L. (2006) Bayesian inference for continuous-time ar models driven by non-gaussian lévy processes. *Proc. Int. Conf. Acoustics, Speech and Signal Processing*. New York: Institute of Electrical and Electronics Engineers.

Grubmüller, H. and Tavan, P. (1994) Molecular dynamics of conformational substates for a simplified protein model. *J. Chem. Phys.*, **101**, 5047–5057.

Hummer, G. (2005) Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, **7**, no. 34.
Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *J. Bas. Engng*, **82**, 35–45.
Kloeden, P. E. and Platen, E. (1992) *Numerical Solutions of Stochastic Differential Equations*. New York: Springer.
Kramers, H. A. (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, **7**, 284–304.
Lasota, A. and Mackey, M. C. (1994) *Chaos, Fractals and Noise*. New York: Springer.
Liu, J. S. (2001) *Monte Carlo Strategies in Scientific Computing*. New York: Springer.
Meng, X.-L. and van Dyk, D. (1997) The EM algorithm—an old folk-song sung to a fast new tune (with discussion). *J. R. Statist. Soc.* B, **59**, 511–567.
Nualart, D. (1991) *The Malliavin Calculus and Related Topics*. New York: Springer.
Øksendal, B. (2000) *Stochastic Differential Equations, an Introduction with Applications*. New York: Springer.
Pokern, Y. (2007) Fitting stochastic differential equations to molecular dynamics data. *PhD Thesis*. University of Warwick, Coventry.
Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. New York: Springer.
Roberts, G. O. and Stramer, O. (2001) On inference for nonlinear diffusion models using the hastings-metropolis algorithms. *Biometrika*, **88**, 603–621.
Roberts, G. O. and Tweedie, R. L. (1997) Exponential convergence of langevin diffusions and their discrete approximations. *Bernoulli*, **2**, 341–363.
Schlick, T. (2000) *Molecular Modeling and Simulation—an Interdisciplinary Guide*. New York: Springer.