Parameter estimation for partially observed hypoelliptic diffusions

Yvo Pokern*
Andrew M. Stuart†
Petter Wiberg‡

Abstract. 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 inter-sample 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 display asymptotic consistency of the method when applied to simulated data. The paper concludes with application of the Gibbs sampler to molecular dynamics data.

Key words. parameter estimation, hypoelliptic diffusion, partial observation, Gibbs sampler, numerical methods.

AMS subject classifications. 60G35 Applications (signal detection, filtering, etc.), 62M05 Markov processes: estimation, 65C30 Stochastic differential and integral equations.

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,[1] and also where observations are only made of variables that are not directly forced by white noise. Such a structure arises naturally in a number of applications.

One application is the modeling of macro-molecular systems [16] and [18]. In its basic form the molecule is described by a large Hamiltonian system of ordinary differential equations (ODEs). 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. While this phrasing of the question is relatively recent, under the name of the "Kramers problem" it dates back to [22] with a brief summary in section 5.3.6a of [13]. Another application,

* Mathematics Institute, University of Warwick, Coventry CV4 7AL, England.
email: pokern@maths.warwick.ac.uk.
† Mathematics Institute, University of Warwick, Coventry CV4 7AL, England.
email: stuart@maths.warwick.ac.uk.
‡ Goldman-Sachs, London.

[1] Meaning that the covariance matrix of the noise is degenerate, but the probability densities are smooth.

audio signal analysis, is referred to in [15] where a continuous time ARMA model is used.

We consider SDE models of the form

$$\left\{ \begin{array}{rcl} \mathrm{d}x & = & \Theta A(x)\mathrm{d}t + C\mathrm{d}B \\ x(0) & = & x_0 \end{array} \right. \tag{1.1}$$

where B is an m-dimensional Wiener process and x a k-dimensional continuous process with $k > m$. $A : \mathbb{R}^k \longrightarrow \mathbb{R}^l$ 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, $\Theta \in \mathbb{R}^{k\times l}$, and the diffusivity matrix C which we assume to be of the form

$$C = \left[\begin{array}{c} 0 \\ \Gamma \end{array} \right] \in \mathbb{R}^{k\times m}$$

where $\Gamma \in \mathbb{R}^{m\times m}$ is nonsingular.

It is known that under the above hypotheses on A and C, a unique L^2-integrable solution $x(\cdot)$ exists almost-surely for all times $t \in \mathbb{R}^+$, see e.g. Theorem 5.2.1 in [27]. We also assume that the process defined by (1.1) is hypoelliptic as defined in [26]. Intuitively, this corresponds to the noise being spread into all components of the system (1.1) 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*[2]. To distinguish between rough and smooth variables, we introduce the notation $x(t)^T = (u(t)^T, v(t)^T)$ where $u(t) \in \mathbb{R}^{k-m}$ is smooth and $v(t) \in \mathbb{R}^m$ is rough. It is helpful to define linear functions $P : \mathbb{R}^k \to \mathbb{R}^{k-m}$ by $Px = u$ and $Q : \mathbb{R}^k \to \mathbb{R}^m$ by $Qx = v$.

We suppose that the smooth component of a sample path is observed at $N + 1$ equally spaced points in time, $\{x_n = x(n\Delta t)\}_{n=0}^N$, and we write $x_n^T = (u_n^T, v_n^T)$ 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. Our interest is in 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 (1.1); such a parameter estimation problem arises naturally in many applications and an example is given in section 7. It is natural to consider $N\Delta t = T \gg 1$ and $\Delta t \ll 1$. It is important to realize that, for continuous time observations, the diffusion coefficient Γ can be found from the quadratic variation of a single path on $[0, T]$, any $T > 0$, see e.g. Theorem 2.8.6 in [7]. For Θ, however, the estimates are strongly consistent only as $T \to \infty$. These two facts will be reflected in the parameter estimation for discrete time observations.

The sequence $\{x_n\}_{n=0}^N$ defined above is generated by a Markov chain. By expanding the random map $x_n \mapsto x_{n+1}$ in powers of Δt, and retaining the leading order contributions to the mean and to the variance in each component of the equation, one obtains

$$x_{n+1} \approx x_n + \Delta t\Theta A(x_n) + \sqrt{\Delta t}R(\Delta t; \Theta)\xi_n \tag{1.2}$$

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 invertible, but the zeros in C mean that is is highly

[2] We do not mean C^∞ here, but they are at least C^1

ill-conditioned for $0 < \Delta t \ll 1$. In fact we have:

$$R(0;\Theta) = \begin{bmatrix} 0 & 0 \\ 0 & \Gamma \end{bmatrix}. \tag{1.3}$$

We refer to expressions of the form (1.2) as statistical models and we will use them to approximate the exact likelihood, $\mathcal{L}(u, v|\Theta, \Gamma\Gamma^T)$, of the path u, v given parameter values Θ and $\Gamma\Gamma^T$.

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

$$\begin{aligned} \mathcal{L}(v, \Theta, \Gamma\Gamma^T) &= \tfrac{\mathcal{L}(v,\Theta,\Gamma\Gamma^T,u)}{\mathcal{L}(u)} \\ &= \mathcal{L}(u, v|\Theta, \Gamma\Gamma^T)\tfrac{p_0(\Theta,\Gamma\Gamma^T)}{\mathcal{L}(u)} \end{aligned} \tag{1.4}$$

In principle, this can be used as the basis for Bayesian sampling of $(\Theta, \Gamma\Gamma^T)$, viewing v as missing data. However, the exact likelihood of the path is typically unavailable. In this paper we will combine judicious approximations of this likelihood to solve the sampling problem. The approximations that we use, $\mathcal{L}_E$ and $\mathcal{L}_{IT}$, are found from (1.2), in the case of $\mathcal{L}_E$ by replacing $R(\Delta t; \Theta)$ with $R(0; \Theta)$ given by (1.3). Thus $\mathcal{L}_E$ is found from an Euler-Maruyama approximation of (1.1). The approximate likelihood $\mathcal{L}_{IT}$ arises from retaining further terms in the Itô-Taylor expansion to ensure that noise is propagated into each component of the map (1.2).

The questions we address in this paper are:

1. How does the ill-conditioning of the Markov chain $x_n \mapsto x_{n+1}$ affect parameter estimation for $\Gamma\Gamma^T$ and for the last m rows of Θ in the regime $\Delta t \ll 1$, $N\Delta t = T \gg 1$?
2. In many applications, it is natural that only the smooth data $\{u_n\}_{n=0}^N$ is 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 $N\Delta t = T \gg 1$?
3. The exact likelihood is usually not available; what approximations of the likelihood should be used, in view of the ill-conditioning?
4. How should the answers to these questions be combined to produce an effective method to sample the distribution of parameters Θ, $\Gamma\Gamma^T$ 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 central to the questions above. We will 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 behavior of other methods.

For the numerical simulations, we will use either exact discrete time samples of (1.1) in simple linear cases, or trajectories obtained by Euler-Maruyama simulation of the SDE on a temporal grid with a spacing considerably finer than the observation time interval Δt.

At this point, we introduce some notation to simplify the presentation. Firstly, 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 $x \in \mathbb{R}^n$. Also, we will occasionally refer to a lilkelihood $\mathcal{L}(B)$ as a function of some parameters B not mentioning the complementary parameter set C. This is understood to refer to the conditional likelihood $\mathcal{L}(B|C)$ whenever the parameter set C is clear from the context.

In section 2 we will introduce our three model problems and in section 3 we study the performance of $\mathcal{L}_E$ to estimate the diffusion coefficient. Observing and analysing its failure in the case with partial observation leads to the improved statistical model yielding $\mathcal{L}_{IT}$ which eliminates these problems; we introduce this in section 4. In section 5 we show that $\mathcal{L}_{IT}$ is inappropriate for drift estimation, but that $\mathcal{L}_E$ 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 a simple application to molecular dynamics and section 8 provides concluding discussion.

1.1. Literature review. The primary novelty of our work is that it concerns hypoelliptic diffusions where only smooth components are observed. We set our work in context. First, we review parameter estimation for (1.1) in continuous time. We assume that the observation is compatible with (1.1) in that, if the observed path is $x(t)^T = (u(t)^T, v(t)^T)$, then

$$\dot{u} = P\Theta A(x), \qquad u(0) = Px(0); \tag{1.5}$$

furthermore, if only $u(t)$ is observed, then we assume that (1.5) determines $v(t)$ uniquely. (In situations where compatibility fails it is necessary to add observational noise to the solution of (1.5) and to estimate it.)

Once v is determined uniquely we have

$$dv = Q\Theta A(x) + \Gamma dB, \quad v(0) = Qx(0). \tag{1.6}$$

The covariance matrix $\Gamma\Gamma^T$ can be estimated by noting that

$$\frac{1}{T}\sum_{n=0}^{N-1}(v_{n+1} - v_n)(v_{n+1} - v_n)^T \to \Gamma\Gamma^T \quad \text{as} \quad N \to \infty \tag{1.7}$$

with $T = N\Delta t$ fixed [7].

The Girsanov formula shows that the path space likelihood for (1.6) is proportional to

$$\exp\left(\int_0^T \Gamma^{-1}Q\Theta A(x(s))\Gamma^{-1}dv(s) - \frac{1}{2}\int_0^T \left\|\Gamma^{-1}Q\Theta A(x(s))\right\|^2 ds\right).$$

This can be used as the basis for various estimation procedures, one of them being the maximum likelihood estimator for the lower rows of Θ which is found by maximizing

$$\hat{\Theta} = \text{argmax}_\Theta\left(\int_0^T \Gamma^{-1}Q\Theta A(x(s))\Gamma^{-1}dv(s) - \frac{1}{2}\int_0^T \left\|\Gamma^{-1}Q\Theta A(x(s))\right\|^2 ds\right) \tag{1.8}$$

over Θ. Such estimators are consistent as $T \to \infty$. In the linear case, where A is just the identity, the maximum likelihood estimate for the whole of Θ is given by

$$\hat{\Theta} = [\int_0^T dx x^T][\int_0^T xx^T dt]^{-1}. \tag{1.9}$$

This is proved to be consistent as $T \to \infty$ in [3]. Note, then, that diffusion parameters can be estimated from arbitrarily short pieces of trajectory, whereas drift parameters

require long time intervals. A discussion of continuous time parameter estimation for linear hypoelliptic diffusions with multiplicative noise is given in [20].

In practice, observations are typically made in discrete time. There is substantial literature on parameter estimation in this context, much of it concerned with estimation of φ in problems of the form

$$dv = a(v, \varphi)dt + \Gamma \dot{w}, \qquad v(0) = v_0, \tag{1.10}$$

where $\Gamma\Gamma^T$ is everywhere invertible. In some cases, a is allowed to depend on the entire path $\{v(s)\}_{s\in[0,t]}$ and then the hypoelliptic problem (1.6) is a special case. We now discuss the literature available when only discrete time observations of v, the rough variable, are given. Note that, for most of this paper, we assume that the v-data is hidden and only u in (1.1) is observed. Thus although u can be eliminated from (1.1), and an equation written for v in the form (1.10) with a depending on the entire path of v on $[0, t]$, the existing literature on discrete time observations of (1.10) does not apply to the case we consider here, where v is not observed. Nonetheless we overview what is known.

One approach is to form continuous time estimators, using the generalization of (1.8) to (1.10). If φ appears linearly and only in a, not γ, then the continuous time estimator can be calculated from Riemann and stochastic integrals of $v(t)$. These continuous time estimators can be approximated by quadrature, assuming the time increment between observations, Δt, is small, and estimates of $\widehat{\varphi}$ obtained in this manner, see [21] for details. An alternative, when Δt is small, is to approximate the likelihood of the discrete time Markov chain generated by sampling (1.10) at rate Δt. This approach is considered in [32, 14, 5, 11] with several of these papers studying the Euler approximation, generating a Gaussian likelihood, as we do in this paper. Theorems about convergence of parameter estimates typically consider the limit $\Delta t \to 0$ with $N\Delta t \to \infty$ [11]. Alternatively one may consider $\Delta t \to 0$ with $N\Delta t = T \gg 1$ and estimate the bias due to finite T.

In [5] functionals of the Brownian bridge are used to build up the approximation; in [30] related ideas are used in a Bayesian approach to parameter estimation for discretely observed diffusions.

When the time increment between observations, Δt, is not small then $O(1)$ errors can enter parameter estimates unless the discrete time likelihood is carefully approximated. One way to do this is by fine Monte Carlo simulation between observation points, see [28]. Another approach, leading to closed formulas and using Hermite polynomials, may be found in [1]. In [5] functionals of the Brownian bridge are used to build up the approximation; in [30] related ideas are used in a Bayesian approach to parameter estimation for discretely observed diffusions. Recent work of Beskos et al uses exact sampling of a diffusion process to address this issue, see [8]. A review of estimation for discretely observed diffusion processes, and a discussion of martingale estimating functions, can be found in [2].

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 nonlinear oscillator of similar form. All these problems are hypoelliptic diffusions and we will present them in detail in the next three subsections. Their general form is given as the second order Langevin equation

$$\left\{ \begin{array}{rcl} dq &=& p dt, \\ dp &=& \left(-\gamma p + f(q)\right) dt + \sigma dB \end{array} \right. \tag{2.1}$$

where f is some (possibly nonlinear) force-function and the variables q and p are scalar.

2.1. Model Problem I: Stochastic Growth. Here, $x = (q, r)^T$ satisfies

$$\begin{cases} dq &= rdt \\ dr &= \sigma dB. \end{cases} \tag{2.2}$$

The process has one parameter, the diffusion parameter σ, that describes the size of the fluctuations. In the setting of (1.1) we have

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

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

$$\mu(t) = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} q_0 \\ r_0 \end{bmatrix} \qquad \text{and} \qquad \Sigma(t) = \sigma^2 \begin{bmatrix} t^3/3 & t^2/2 \\ t^2/2 & t \end{bmatrix}.$$

The exact discrete samples may be written as

$$\begin{cases} q_{n+1} &= q_n + r_n \Delta t + \sigma \frac{(\Delta t)^{3/2}}{\sqrt{12}} \zeta_n^{(1)} + \sigma \frac{(\Delta t)^{3/2}}{2} \zeta_n^{(2)}, \\ r_{n+1} &= r_n + \sigma \sqrt{\Delta t} \zeta_n^{(2)}, \end{cases} \tag{2.3}$$

with $\zeta_0 \sim \mathcal{N}(0, I)$ and $\{\zeta_n\}_{n=0}^N$ being i.i.d.

2.2. Model Problem II: Harmonic Oscillator. As our second model problem we consider a damped harmonic oscillator driven by a white noise forcing where $x = (q, p)^T$:

$$\begin{cases} dq &= pdt \\ dp &= -Dqdt - \gamma pdt + \sigma dB. \end{cases} \tag{2.4}$$

This model is obtained from the general SDE (1.1) for the choice

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

and $u = q$, $v = p$. The process is Gaussian and the mean and covariance of the solution can be explicitly calculated.

2.3. Model Problem III: Oscillator with trigonometric potential. In the third model problem, $x = (q, p)^T$ 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 [23]. This potential is sometimes used in molecular dynamics in connection with the dynamics of dihedral angles – see section 7. The model is

$$\begin{cases} dq &= pdt, \\ dp &= (-\gamma p - \sum_{j=1}^c D_j \sin(q) \cos^{j-1}(q))dt + \sigma dB. \end{cases} \tag{2.5}$$

This equation has parameters γ, D_i, $i = 1, \ldots, c$ and σ. It can be obtained from the general SDE (1.1) for the choice

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

and $u = q$, $v = p$. No explicit closed-form expression for the solution of the SDE is known in this case; the process is not Gaussian.

3. Euler Statistical Model. In this section, the Euler-Maruyama approximation to (1.1) is used to generate a statistical model and associated likelihood. Using this likelihood to estimate the diffusivity works whenever observations of both the smooth and the rough components are available. However, it yields $\mathcal{O}(1)$ errors in the partially observed case; this is demonstrated analytically for Model Problem I and the results are extended by means of numerical experiments.

3.1. Statistical Model. If the force function $A(\cdot)$ is nonlinear, closed-form expressions for the likelihood are in general unavailable. To overcome this obstacle, one can use a discrete time statistical model. The Euler model 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 (1.1) is

$$X_{n+1} = X_n + \Delta t \Theta A(X_n) + \sqrt{\Delta t} C \xi_n \tag{3.1}$$

where $\xi_n \sim \mathcal{N}(0, I)$ is an i.i.d. sequence of k-dimensional vectors with standard normal distribution. This corresponds to (1.2) with $R(\Delta t; \Theta)$ replaced by $R(0; \Theta)$ from (1.3). Thus we obtain

$$\left\{ \begin{array}{rcl} U_{n+1} & = & U_n + \Delta t P \Theta A(X_n) \\ V_{n+1} & = & V_n + \Delta t Q \Theta A(X_n) + \sqrt{\Delta t} \Gamma \xi_n \end{array} \right\} \tag{3.2}$$

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

$$\begin{array}{l} \mathcal{L}_{ND}(U, V|\Theta, \Gamma\Gamma^T) = \\ \prod_{n=0}^{N-1} \frac{\exp\left(-\frac{1}{2}\|\Delta V_n - \Delta t Q \Theta A(X_n)\|_\Gamma^2\right)}{\sqrt{2\pi|\Gamma\Gamma^T|}} \delta\left(\frac{U_{n+1}-U_n}{\Delta t} - P\Theta A(X_n)\right) \cdot \end{array} \tag{3.3}$$

The Dirac mass insists that the data is compatible with the statistical model, i.e. the V path must be given by numerical differentiation (ND) of the U path. To estimate parameters we will use the following expression:

$$\mathcal{L}_E(U, V|\Theta, \Gamma\Gamma^T) = \prod_{n=0}^{N-1} \frac{\exp\left(-\frac{1}{2}\|\Delta V_n - \Delta t Q \Theta A(X_n)\|_\Gamma^2\right)}{\sqrt{2\pi|\Gamma\Gamma^T|}} \ , \tag{3.4}$$

where we assume that $\{u_n\}$, $\{v_n\}$ are related through numerical differentiation when the Euler model is used to estimate missing components.

3.2. Model Problem I. The Euler statistical model for this model problem is

$$\begin{cases} Q_{n+1} = Q_n + R_n \Delta t, \\ R_{n+1} = R_n + \sigma \sqrt{\Delta t} \xi_n. \end{cases} \tag{3.5}$$

Here, $\{\xi_n\}$ is an i.i.d. $\mathcal{N}(0,1)$ sequence. The root cause of the phenomena that we discuss in this paper is manifest in comparing (2.3) and (3.5). The difference is that the $O((\Delta t)^{3/2})$ white noise contributions in the exact time series (2.3) do not appear in the equation for Q_n. We will 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 (3.5) gives the formula

$$R_n = \frac{Q_{n+1} - Q_n}{\Delta t} \tag{3.6}$$

for the missing data. In the following numerical experiment we generate exact data from (2.3) using the parameter value $\sigma = 1$. We substitute R_n given by (3.6) into (3.4) 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 $\{R_n\}$, from (2.3), and again use a maximum likelihood estimator for σ from (3.4).

Using $N = 100$ timesteps for a final time of $T = 10$ with $\sigma = 1$ the histograms for the estimated difffusion coefficient presented in the middle column of Figure 3.2 are obtained. The top row contains histograms obtained in the case of complete observation where good agreement between the true σ and the estimates is observed. The bottom row contains the histograms obtained for partial observation using (3.6). The observed mean value of $\mathbb{E}\widehat{\sigma} = 0.806$ indicates that the method yields biased estimates. Increasing the final time to $T = 100$ (see left column of graphs in Figure 3.2) or increasing the resolution to $\Delta t = 0.01$ do not remove this bias.

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

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. The log-likelihood function given in (3.4) yields the following expression in the case of stochastic growth:

$$\log \mathcal{L}_E(\sigma) = -2N \log \sigma - \frac{1}{\sigma^2 \Delta t} \sum_{n=0}^{N-1} (\Delta r_n)^2$$

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

$$\widehat{\sigma}^2 = \frac{1}{N\Delta t} \sum_{n=0}^{N-1} (\Delta r_n)^2. \tag{3.7}$$

In the case of complete data, (2.3) gives

$$\widehat{\sigma}^2 = \frac{\sigma^2}{N} \sum_{n=0}^{N-1} (\zeta_n^{(2)})^2. \tag{3.8}$$

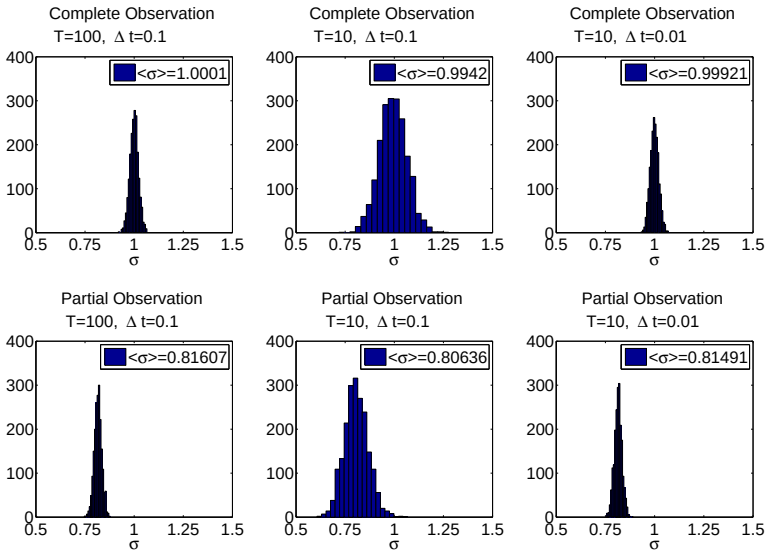

FIG. 3.1. *Estimates of σ using Euler Model for Model Problem I. Top row: fully observed process; bottom row: partially observed process.*

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

Let us consider what happens when r is hidden. In this case, r_n is estimated by

$$\widehat{r}_n = \frac{q_{n+1} - q_n}{\Delta t}.$$

But since q_n is generated by (2.3) we find that

$$\widehat{r}_n = \frac{r_{n+1} + r_n}{2} + \sigma \frac{\sqrt{\Delta t}}{\sqrt{12}} \zeta_n^{(1)}$$

and

$$\begin{aligned} \Delta \widehat{r}_n &= \frac{\Delta r_{n+1}}{2} + \frac{\Delta r_n}{2} + \sigma \frac{\sqrt{\Delta t}}{\sqrt{12}} \left(\zeta_{n+1}^{(1)} - \zeta_n^{(1)} \right) \\ &= \frac{\sigma \sqrt{\Delta t}}{2} \left(\zeta_{n+1}^{(2)} + \zeta_n^{(2)} + \frac{1}{\sqrt{3}} \zeta_{n+1}^{(1)} - \frac{1}{\sqrt{3}} \zeta_n^{(1)} \right) \end{aligned}$$

When $\Delta \widetilde{r}_n$ is inserted in (3.7) it follows that

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

The random variables $\{\zeta_n\}_{n=0}^N$ are i.i.d with $\zeta_0 \sim N(0, I)$. So, by the law of large numbers, $\widehat{\sigma}^2 \to \frac{2}{3}\sigma^2$ almost surely as $N \to \infty$. Furthermore, the limits hold in either of the cases where either $N\Delta t = T$ 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 $O(1)$ errors in the diffusion coefficient. [3]

4. Improved statistical model. The failure of the Euler model to estimate paths having the correct quadratic variation is caused by not propagating the noise to the smooth component of the solution. A new statistical model is thus proposed which propagates the noise 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 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 using our common framework the Model Problems I, II and III, namely (2.1). The improved statistical model is based on the observation that in the second row of an Itô-Taylor expansion of (2.1) 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 (2.1):

$$\begin{bmatrix} Q_{n+1} \\ P_{n+1} \end{bmatrix} = \begin{bmatrix} Q_n \\ P_n \end{bmatrix} + \Delta t \begin{bmatrix} P_n \\ f(Q_n) - \gamma P_n \end{bmatrix} + \sigma \begin{bmatrix} \int_0^{\Delta t} B(s)ds \\ B(\Delta t) \end{bmatrix}$$

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:

$$\begin{bmatrix} Q_{n+1} \\ P_{n+1} \end{bmatrix} = \begin{bmatrix} Q_n \\ P_n \end{bmatrix} + \Delta t \begin{bmatrix} P_n \\ f(Q_n) - \gamma P_n \end{bmatrix} + \sigma\sqrt{\Delta t} R \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \tag{4.1}$$

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

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

This is a specific instance of (1.2). It should be noted that this model is in agreement with the Ito-Taylor approximation up to error terms of order $\mathcal{O}(\Delta t^2)$ in the first row and $\mathcal{O}(\Delta t^{\frac{3}{2}})$ in the second row.

If complete observations are available, this model performs satisfactorily for the 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 the first row (referring to complete observation) of Figure 4.2 for Model Problem I and from the first row of Figure 4.2 for Model Problem II. In both cases the true value is given by $\sigma = 1$. See subsection 4.2 for a full discussion of these numerical experiments.

[3] There is similarity here with work of Gaines and Lyons [12] showing that adaptive methods for SDEs get the quadratic variation wrong if the adaptive strategy is not chosen carefully.

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/smoother [19, 4] which could then be combined with the expectation-maximisation (EM) algorithm [6, 25] to estimate parameters. In this paper, however, we employ a Bayesian approach sampling directly from the posterior distribution for the rough component, p, without factorising the sampling into forward and backward sweeps.

4.1. Path Sampling. The log likelihood functional for the missing data induced by the statistical model (1.2) can be written as follows:

$$\log \mathcal{L}_{IT}(p) = -\frac{1}{2\sigma^2} \sum_{l=0}^{N} \|\Delta X_l - \Theta A(X_l)\Delta t\|_R^2 + \text{const.} \tag{4.2}$$

We will apply this in the case (4.1) which is a specific instance of (1.2).

One way to sample from this likelihood $\mathcal{L}_{IT}(p)$ for rough paths $\{p_i\}_{i=0}^N$ is via the Langevin equation (see section 6.5.2 in [29]) and, in general, we expect this to be effective in view of the high dimensionality of p. However, when p is Gaussian it is possible to generate independent samples, and we explain how this may be implemented below.

The Langevin equation is:

$$\frac{dp}{ds} = \nabla_p \log \mathcal{L}_{IT}(p) + \sqrt{2}\frac{dW_s}{ds} \tag{4.3}$$

The required derivatives of $\log \mathcal{L}_{IT}(p)$ with respect to the rough path p are computed in the Appendix. For our oscillator framework, they can be expressed using a tridiagonal, negative definite matrix P_{mat} with highest order stencil $-1\ -4\ -1$ acting on the p-vector plus a possibly nonlinear contribution $Q(q)$ acting on the q-vector only. The gradient of $\mathcal{L}_{IT}$ can then be written as follows:

$$\nabla_p \log \mathcal{L}_{IT}(q,p) = P_{\text{mat}} p + Q(q).$$

The suggested sampler for p-paths is simply:

$$p_n = -P_{\text{mat}}^{-1} Q(q) + U^{-1}\xi_n \tag{4.4}$$

Here $U^T U = -P_{\text{mat}}$ is a Cholesky factorization.

4.2. Estimating diffusion coefficient and missing path. The approximate likelihood $\mathcal{L}_{IT}(P, Q|\sigma, \Theta)$ can be used to estimate both the missing path p and the diffusion coefficient σ for our Model Problems I, II and III.

In order to estimate σ, the derivative of the log likelihood

$$\log \mathcal{L}_{IT}(\sigma) = \log \mathcal{L}_{IT}(P, Q|\sigma, \Theta) + \log\left(\frac{p_0(\Theta,\sigma)}{\mathcal{L}(P,Q,\Theta)}\right)$$

(where priors $p_0(\Theta,\sigma)$ are assumed to be given) with respect to σ is computed:

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

Here, we have used the abbreviation

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

In this case no prior distribution was felt necessary in this example, as when $N \to \infty$ its importance would diminish rapidly. Thus we set $p_0 \equiv 1$. The resulting maximum likelihood estimator is:

$$\widehat{\sigma^2} = \frac{Z}{2N\Delta t} \tag{4.5}$$

Instead of providing just the maximum of the likelihood it may be more desirable to sample from the distribution of σ given observations p and q and the drift parameters. As the derivative of the log-likelihood conditional on these obsevations is available we can write a Langevin type sampler for this distribution in the following form:

$$\begin{aligned} d\sigma &= \frac{\partial \mathcal{L}_{IT}}{\partial \sigma} ds + \sqrt{2} dW \\ &= \left(-\frac{2N}{\sigma} + \frac{1}{\sigma^3} Z \right) ds + \sqrt{2} dW \end{aligned}$$

Empirically, the singularity at $\sigma = 0$ is seen to be more amenable to numerical solution if the transformation $\zeta(\sigma) = \sigma^4$ is used. Using the Itô formula, this yields the Langevin sampler:

$$d\zeta = \left((12 - 8N)\sqrt{\zeta} + 4Z \right) ds + 4\sqrt{2}\zeta^{\frac{3}{4}} dW. \tag{4.6}$$

A simple explicit Euler-Maruyama discretisation in s is used to simulate paths for this SDE.

This Langevin-type sampler (4.6) can then be alternated in a Systematic-Scan Gibbs Sampler (as described on p.130 of [24]) using N_{Gibbs} iterations with the direct sampler for the paths, (4.4). The yields estimates of the missing path and the diffusion coefficient which is estimated by averaging over the N_{Gibbs} samples. We illustrate this with an example. For Model Problem I we use the following parameters:

$$\sigma = 1 \quad T \in \{10, 100\} \quad \Delta t \in \{0.1, 0.01\} \quad N_{\text{Gibbs}} = 10$$

The sample paths used for the fitting are generated from exact samples using (2.3) and the resulting plot is given in Figure 4.2 where the first row corresponds to the behaviour when complete observations are available and the second row corresponds to only the smooth component being observed. For Model Problem II we use the following parameters:

$$\begin{array}{lll} \sigma = 1 & D = 4 & \gamma = 0.5 \\ T \in \{10, 100\} & \Delta t \in \{0.02, 0.002\} & N_{\text{Gibbs}} = 10 \end{array}$$

The sample paths used for the fitting are generated using a subsampled Euler-Maruyama method with temporal grid $\frac{\Delta t}{k}$ where $k = 30$. This experiment results in the plot given in Figure 4.2.

It appears from these figures that the estimator for this joint problem performs well for Model Problem I. While the bias observed in Model Problem II can be considerable, it decays under Δt refinement. 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.

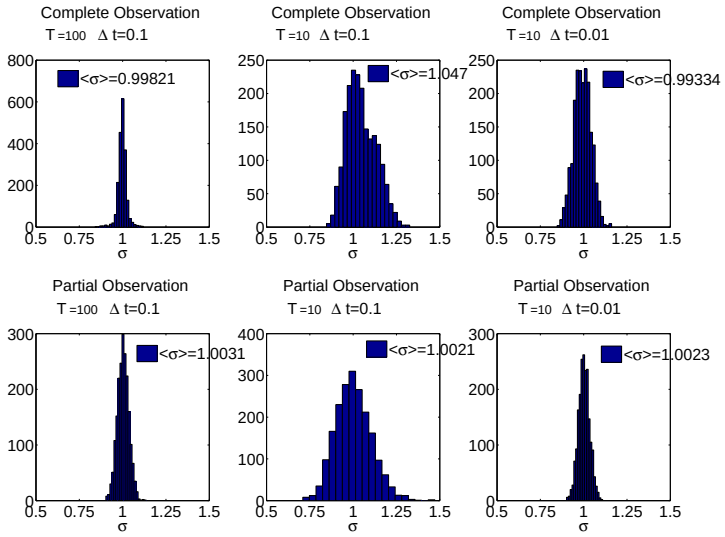

FIG. 4.1. *Estimates of σ using the $\mathcal{L}_{IT}$ Model for Model Problem I. Top row: fully observed process; bottom row: partially observed process.*

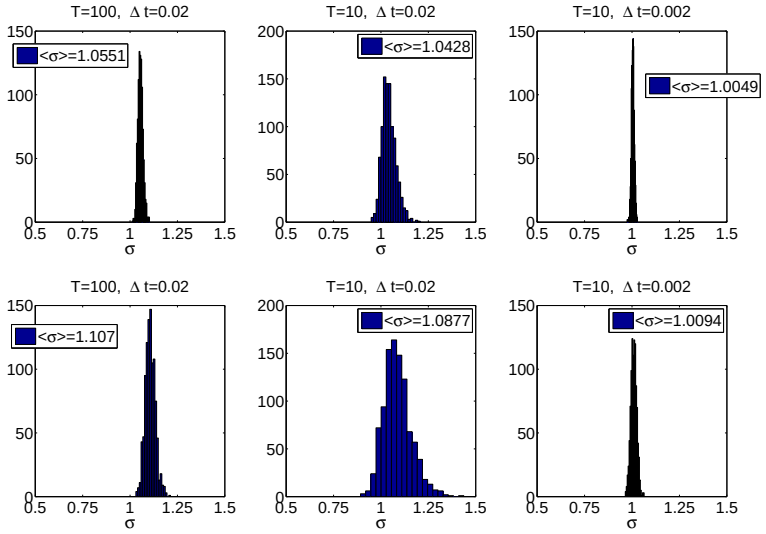

FIG. 4.2. *Estimates of σ using the $\mathcal{L}_{IT}$ Model for Model Problem II. Top row: fully observed process; bottom row: partially observed process.*

5.1. Overview. With the approximate likelihoods $\mathcal{L}_E$ and $\mathcal{L}_{IT}$ 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 $\mathcal{L}_E$ based maximum likelihood estimator performs well. In contrst, ill-conditioning due to hypoellipticity leads to error amplification and affects the performance of the $\mathcal{L}_{IT}$ based estimator. Alternatively, the estimator (1.9) suggested by Le Breton and Musiela can be used, but this is inap-

propriate if a harmonic oscillator fit is sought, as it means that all entries of Θ must be estimated and known entries of Θ cannot be fixed a priori. While it is possible to use a cut-back version of this estimator applying it to only those rows of Θ whose entries need to be estimated, it is unclear how to obtain an approximate likelihood corresponding to this estimator that is amenable to Langevin sampling of the drift parameters and – at the same time – avoids the error amplification observed in the $\mathcal{L}_{IT}$-based case.

Hence, since the $\mathcal{L}_E$-based estimators also cover Model Problem III, and since they are amenable to Langevin sampling, they are our choice for estimating drift parameters.

5.2. Drift parameters from $\mathcal{L}_E$. In order to simplify analysis, we illustrate the estimator using the Model Problem II, (2.4), only. Nonetheless, we start from equation (2.1) for which the Euler statistical model is given as follows:

$$\begin{cases} Q_{n+1} &= Q_n + \Delta t P_n \\ P_{n+1} &= P_n + \Delta t \sum_{i=1}^{c} D_i f_i(Q_n) - \Delta t \gamma P_n + \sqrt{\Delta t}\sigma \xi_n \end{cases} \tag{5.1}$$

Here, we assume that the force functions $\{f_i\}_{i=1}^c$ are prescaled by parameters $D_i \in \mathbb{R}$. The likelihood functional in this case is given by:

$$\mathcal{L}_E(\gamma, D|Q, P, \sigma) \propto \frac{1}{\sqrt{2\pi\sigma^2}^N} \exp\left(-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(\Delta P_n - \Delta t \sum_{i=1}^{c} D_i f_i(Q_n) + \Delta t \gamma P_n)^2\right)$$

Differentiating this likelihood with respect to the parameters $\{D_i\}_{i=1}^c$ and γ and equating to zero yields a linear system of equations which we denote by

$$M_E \begin{bmatrix} D_1 \\ \vdots \\ D_c \\ \gamma \end{bmatrix} = b_E \tag{5.2}$$

In the harmonic oscillator case, where $c = 1$ and $f_1(q) = -Dq$ we obtain the following linear system:

$$\begin{bmatrix} \sum_{n=0}^{N-1} \Delta t Q_n^2 & \sum_{n=0}^{N-1} \Delta t Q_n P_n \\ \sum_{n=0}^{N-1} \Delta t Q_n P_n & \sum_{n=0}^{N-1} \Delta t P_n^2 \end{bmatrix} \begin{bmatrix} \hat{D} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} -\sum_{n=0}^{N-1} Q_n \Delta P_n \\ -\sum_{n=0}^{N-1} P_n \Delta P_n \end{bmatrix} \tag{5.3}$$

The continuum limit for $\Delta t \longrightarrow 0$ with $N\Delta t = T$ of this system is simply:

$$\begin{bmatrix} \int_0^T q(t)^2 dt & \int_0^T p(t)q(t)dt \\ \int_0^T p(t)q(t)dt & \int_0^T p(t)^2 dt \end{bmatrix} \begin{bmatrix} \hat{D} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} -\int_0^T q(t) dp_t \\ -\int_0^T p(t) dp_t \end{bmatrix}$$

This corresponds to the estimator of D and γ alone given by (1.9). Casting aside issues about the discretisation error (finite Δt), the proof of asymptotic consistency given in [3] still applies to this estimator in the linear case.

Using the same likelihood, $\mathcal{L}_E$, a Langevin sampler can also be used for the drift parameters. Since the resulting distribution for Θ is Gaussian, direct sampling can be used in the spirit of subsection 4.1:

$$\widehat{\Theta} \sim \mathcal{N}\left(M_E^{-1} b_E, M_E^{-1}\right) \tag{5.4}$$

5.3. Drift parameters from $\mathcal{L}_{IT}$. As the approximate model based on $\mathcal{L}_{IT}$ 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 log-likelihood function is given by (4.2). To illustrate the problems arising from the use of $\mathcal{L}_{IT}$ we use Model Problem II, so that (4.2) becomes

$$\log \mathcal{L}_{IT}(\Theta) = \frac{1}{2\sigma^2 \Delta t} \sum_{n=0}^{N-1} \|(X_{n+1} - X_n - \Delta t \Theta A(X_n))\|_R^2 + \text{const} \tag{5.5}$$

where $R = \begin{bmatrix} \frac{\Delta t}{\sqrt{12}} & \frac{\Delta t}{2} \\ 0 & 1 \end{bmatrix}$, irrelevant constants have been omitted and we have

$$A\left(\begin{bmatrix} Q_n \\ P_n \end{bmatrix}\right) = \begin{bmatrix} Q_n \\ P_n \end{bmatrix}, \quad \theta = \begin{bmatrix} 0 & 1 \\ -D & -\gamma \end{bmatrix}.$$

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

$$\begin{bmatrix} \sum_n Q_n^2 \Delta t & \sum_n P_n Q_n \Delta t \\ \sum_n P_n Q_n \Delta t & \sum_n P_n^2 \Delta t \end{bmatrix} \begin{bmatrix} \hat{D} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} -\sum_n Q_n \Delta P_n \\ -\sum_n P_n \Delta P_n \end{bmatrix} + \begin{bmatrix} \sum_n \frac{3}{2} Q_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \\ \sum_n \frac{3}{2} P_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \end{bmatrix} \tag{5.6}$$

Comparing this linear system to the successful estimator (5.2) we note the presence of an additional term on the right hand side. This term leads to the failure of the above estimator.

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 of the tests share these parameters:

$$D = 4 \quad \gamma = 0.5 \quad \sigma = 0.5 \quad k = 30$$

Data for the tests are again generated using an Euler-Maruyama method on a finer temporal grid with resolution $\Delta t/k$. In the plot given in Figure 5.1 the top row contains histograms for the drift parameter D whereas the second row contains histograms for the drift parameter γ in any case using the full sample path for inference. It is clear from these experiments summarised in Figure 5.1 that both D and γ are grossly underestimated.

5.5. Why the Model fails for the drift parameters. The key is to analyse the error term on the right hand side of (5.6) comparing it to the consistent estimator (5.2). Using the 2nd order Itô-Taylor approximation

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

we can compute the error term on the right hand side of (5.6):

$$\begin{bmatrix} \sum_n \frac{3}{2} Q_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \\ \sum_n \frac{3}{2} P_n \left(\frac{\Delta Q_n}{\Delta t} - P_n \right) \end{bmatrix} = \begin{bmatrix} -\frac{3}{4}\gamma \sum_n Q_n P_n \Delta t - \frac{3}{4} D \sum_n Q_n^2 \Delta t \\ -\frac{3}{4} D \sum_n Q_n P_n \Delta t - \frac{3}{4}\gamma \sum_n P_n^2 \Delta t \end{bmatrix} + I_s + \mathcal{O}(\Delta t). \tag{5.7}$$

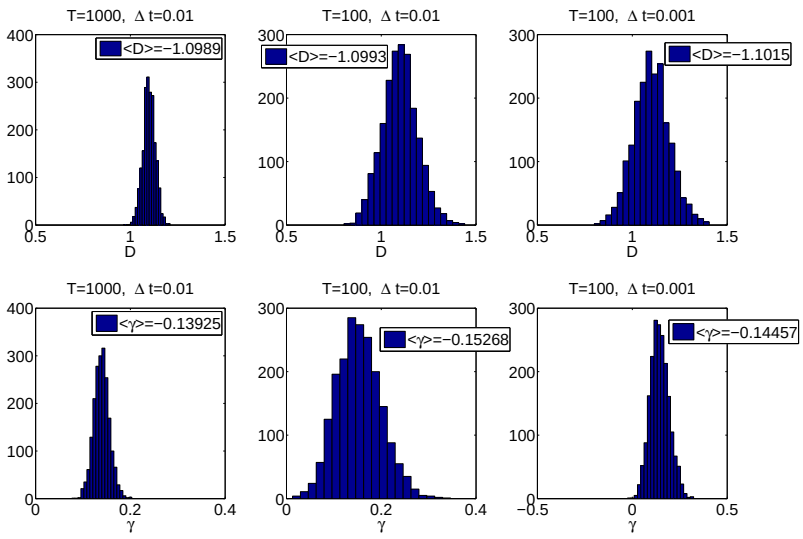

FIG. 5.1. *Drift estimation for Model Problem II, using* $\mathcal{L}_{IT}$

Here, D and γ refer to the exact drift parameters used to *generate* the sample path, whereas $\hat{D}$ and $\hat{\gamma}$ in (5.6) and (5.7) are the drift parameters estimated using the improved statistical model. The term I_s on the right hand side contains stochastic integrals whose expected value is zero.

As the mean error terms can be written in terms of the matrix elements themselves, (5.7) can be substituted in (5.6) to obtain:

$$\mathbb{E}\hat{D} = \frac{1}{4}D + \mathcal{O}(\Delta t) \tag{5.8}$$

$$\mathbb{E}\hat{\gamma} = \frac{1}{4}\gamma + \mathcal{O}(\Delta t). \tag{5.9}$$

This seems to be corroborated by the numerical tests.

5.6. Conclusion for Drift Estimation. It has been observed numerically that the likelihood $\mathcal{L}_E$ associated with an Euler model for the SDE (1.1) yields asymptotically consistent Langevin and maximum likelihood estimators for Model Problem II. For the case of continuous time the proof of asymptotic consistency in the limit $T \to \infty$ given in [3] can be adapted in the linear case (i.e. $A = id$) and it would be expected to carry over to the discretised problem in the limit $\Delta t \to 0$ and $N\Delta t \to \infty$.

While it is aesthetically desirable to base the estimation of all parameters as well as the missing data on the same approximation $\mathcal{L}_{IT}$ of the true likelihood $\mathcal{L}$, 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 get amplified to $\mathcal{O}(1)$ errors in the second row. Estimating all entries of Θ, while being outside the specification of the problem under consideration, also yields $\mathcal{O}(1)$ errors if $\mathcal{L}_{IT}$ is used and so does not remedy the problem. This problem is not shared by the discretised version of the

diffusion independent estimator (1.9), but this is not a maximum likelihood estimator for $\mathcal{L}_{IT}$.

In summary, for the purposes of fitting our model problems to observed data we employ the Euler estimator (5.4) for the drift parameters.

6. The Gibbs Loop. In this section, we combine the insights 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, drift parameters and diffusion parameters. Subsection 6.1 describes the approach in the general case, when applied to (1.1), wheras subsection 6.2 describes the application to Model Problem III.

6.1. Overview. In this section, the estimators for the hidden rough path V, the covariance $\Gamma\Gamma^T$ and the those rows of the drift parameters Θ which are to be estimated are combined in a Gibbs sampler. Given a likelihood $\mathcal{L}(U, V|\Theta, \Gamma\Gamma^T)$, a prior $p_0(\Theta, \Gamma\Gamma^T)$ and observation U, a Systematic Scan Gibbs Sampler would normally work as follows:

1. Sample V from $\mathcal{L}(V|U, \Theta, \Gamma\Gamma^T)$.
2. Sample Θ from $\mathcal{L}(\Theta|U, V, \Gamma\Gamma^T)$.
3. Sample $\Gamma\Gamma^T$ from $\mathcal{L}(\Gamma\Gamma^T|U, V, \Theta)$.
4. Restart from step 1 unless sufficiently equilibrated.

Of course, the exact likelihood for the problem at hand is unavailable and thus approximate likelihoods are chosen. Exactly which approximations are chosen depends on the problem at hand. We have outlined how to construct $\mathcal{L}_{IT}$ approximations to estimate V and $\Gamma\Gamma^T$ by propagating the highest order noise to every row and $\mathcal{L}_E$ approximations for the drift parameter estimation. Numerical and analytical evidence indicates that these approximations work well.

The algorithm to be put in practice thus reads:

1. Sample V from $\mathcal{L}_{IT}(V|U, \Theta, \sigma)$.
2. Sample Θ from $\mathcal{L}_E(\Theta|U, V, \sigma)$.
3. Sample σ from $\mathcal{L}_{IT}(\sigma|U, V, \Theta)$.
4. Restart from step 1 unless sufficiently equilibrated.

In practice, we find that for Model Problem II and III, equilibration is fast. Furthermore, convergence of the estimates to the true parameter values is observed numerically for Model Problems II and III with $\mathcal{O}(\Delta t)$ discretisation errors and $\mathcal{O}\left(\frac{1}{T}\right)$ truncation errors if the sample paths do not start in the equilibrium measure. The overall bias is therefore of order $\mathcal{O}(\Delta t + \frac{1}{T})$ and the observed variance is of order $\mathcal{O}(\frac{1}{T})$. We now show this in detail.

6.2. The Algorithm. The proposed algorithm will be illustrated using Model Problem III.

ALGORITHM 6.1. *Given observations q_i, $i = 1, \ldots, N$, the initial p-path is obtained using numerical differentiation:*

$$p_i^{(0)} = \frac{\Delta q_i}{\Delta t}. \tag{6.1}$$

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

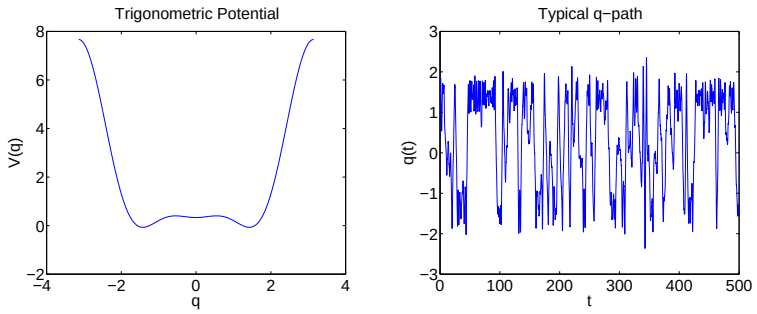

FIG. 6.1. *Typical sample path for Model Problem III, $T = 500$*

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

1. *Estimate the drift parameters $\gamma^{(k)}$ and $\{D_j^{(k)}\}_{j=1}^c$ using sampling from $\mathcal{L}_E$ given $\left\{p_i^{(k-1)}\right\}_{i=0}^N$ via a (5.4).*
2. *Estimate the diffusivity $\sigma^{(k)}$ using the Langevin sampler (4.6) based on $\mathcal{L}_{IT}$ given $\left\{p^{(k)}\right\}_{i=0}^N$ and $\gamma^{(k)}$, $\left\{D^{(k)}\right\}_{j=1}^c$.*
3. *Get an independent sample of the p-path, $\left\{p_i^{(k)}\right\}_{i=0}^N$ using (4.4) derived from $\mathcal{L}_{IT}$ given parameters $\gamma^{(k)}$, $\left\{D_j^{(k)}\right\}_{j=1}^c$ and $\sigma^{(k)}$.*

This algorithm is tested numerically where sample paths of (2.5) are generated using a sub-sampled Euler-Murayama approximation of the SDE. The data is generated using a timestep 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 large enough. The parameters used for these simulations are as follows:

$$\begin{array}{llll} D_0 = 1 & D_1 = -8 & D_2 = 8 & \gamma = 0.5 \quad \sigma = 0.7 \\ T = 500 & \Delta t \in \left\{\frac{1}{2}, \ldots, \frac{1}{128}\right\} & N_{\text{Gibbs}} = 10 & \end{array}$$

The trigonometric potential resulting from this choice of drift parameters is depicted on the left of Figure 6.1 and a typical samplepath is given on the right side of Figure 6.1. It should be noted that all sample paths are started at $(q, p) = (1, 1)$. As the potential is inspired by dihedral angle potentials used in molecular dynamics it seems appropriate that σ is chosen such that metastability occurs. This can be observed in the typical q-path given in Figure 6.1.

Using up to 64000 sample paths we obtain estimates of the drift parameters by averaging over the latter half of $N_{Gibbs} = 50$ Gibbs iterations. We label these as $\langle\widehat{D_i}\rangle$ and $\langle\widehat{\gamma}\rangle$. We then compute their deviation from the true values, $\Delta D_i = \langle\widehat{D_i}\rangle - D_i$ and plot ΔD_i and $\Delta\gamma$ versus Δt in a doubly logarithmic plot given in Figure 6.2.

A similar plot which is given in Figure 6.3 is obtained for the shorter final time $T = 50$ which will be helpful in understanding the influence of finite time resolution Δt and finite final time T on the observed bias of the estimators.

A straight line fit for the doubly logarithmic plot is desired to numerically ascertain the order of convergence. First attempts at obtaining such a fit using a standard least squares procedure yield a slope close to 1 indicating $\mathcal{O}(\Delta t)$ errors in the fitted parameters. However, since the Monte Carlo standard deviations around each data-

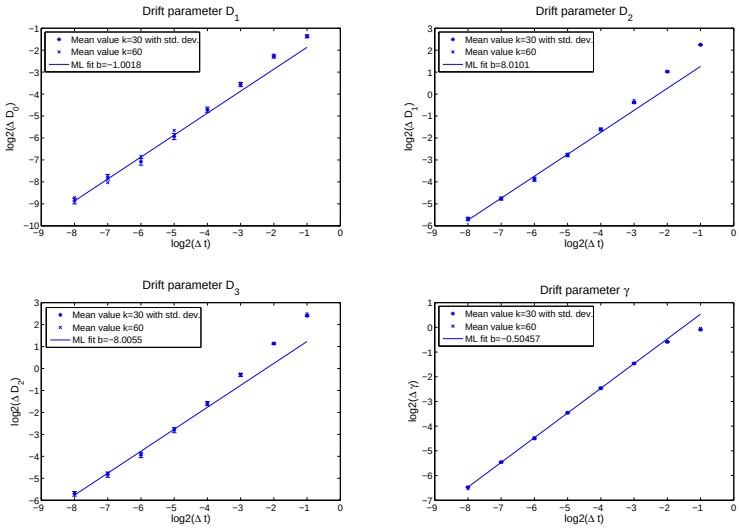

FIG. 6.2. *Whole loop estimation for Model Problem III:* $T = 500$

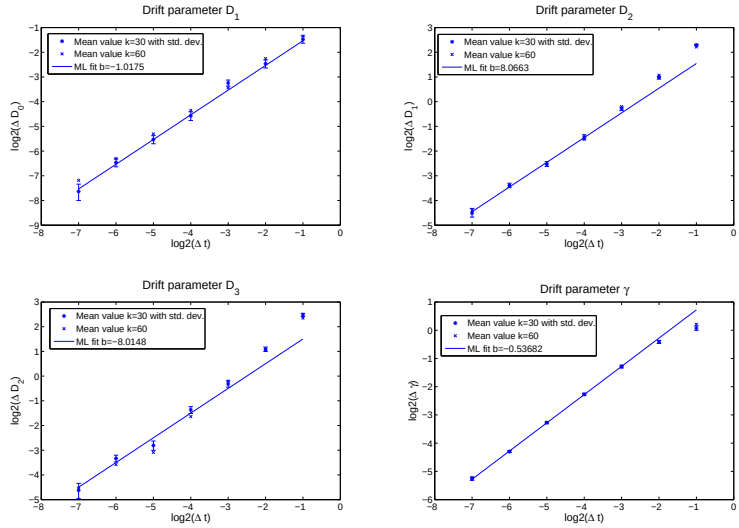

FIG. 6.3. *Whole loop estimation for Model Problem III:* $T = 50$

point get magnified due to the logarithmic transformation, the fact that the apparent variance increases as Δt is decreased has to be taken into account. As the observed transformed standard deviations cannot be assumed to be small in comparison to the observed mean error, a more sophisticated method than the standard least squares fit is suggested.

Given averaged numerically observed parameter estimates y_i and their numerically observed Monte Carlo standard deviations α_i obtained at timesteps Δt_i we fit b

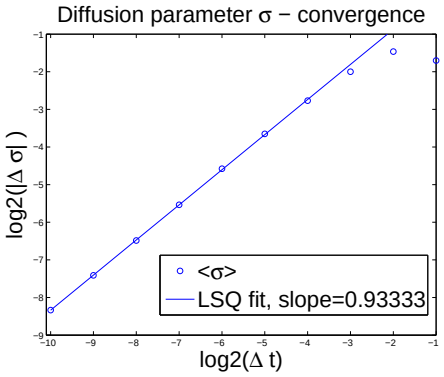

FIG. 6.4. *Whole loop estimation for Model Problem III:* $T = 500$

and c in the following model:

$$\alpha_i \xi_i = y_i - b - c\Delta t_i. \tag{6.2}$$

Assuming that the errors ξ_i are normally distributed (which is empirically found to be the case) 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 reported in Figures 6.2 and 6.3. Note that this fit constrains the slope of the fitted line in the doubly logarithmic plot to one. This is to minimise 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 Figures 6.2 and 6.3 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 (6.2)) from the true parameter value is consistent with it being $\mathcal{O}\left(\frac{1}{T}\right)$. This error is attributed to all sample paths having been started at $(q, p) = (1, 1)$ rather than from a point sampled from the equilibrium measure.

From these coniderations it is apparent that the numerical experiments' outcome is consistent with an $\mathcal{O}(\Delta t) + \mathcal{O}\left(\frac{1}{T}\right)$ bias, making the Algorithm 6.1 an asymptotically consistent estimator of the drift and diffusion parameters.

7. Application to Molecular Conformational Dynamics. As a simple application of fitting hypoelliptic diffusions using partial observations we consider data arising from molecular dynamics simulations of a Butane molecule using a simple heat bath approximation. After describing the origin of the data to be fitted, we observe that for small Δt, fitting an elliptic diffusion process is inappropriate as the fitted diffusion coeffcient $\hat{\sigma}$ tends to zero as $\Delta t \longrightarrow 0$.

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 inter-sample intervals Δt. Also, stabilisation of the fitted force function $f(q) = \sum_{j=1}^{c} D_j f_j(q)$ as the number of terms to be included, c, increases, is observed. Thus the hybrid Algorithm 6.1 is shown to be effective on real data. It is also clear, though, that the resulting fit has only limited predictive capabilities as it fails to fit the invariant measure of the data at all well. However, this is a *modeling* issue which is not central to this paper.

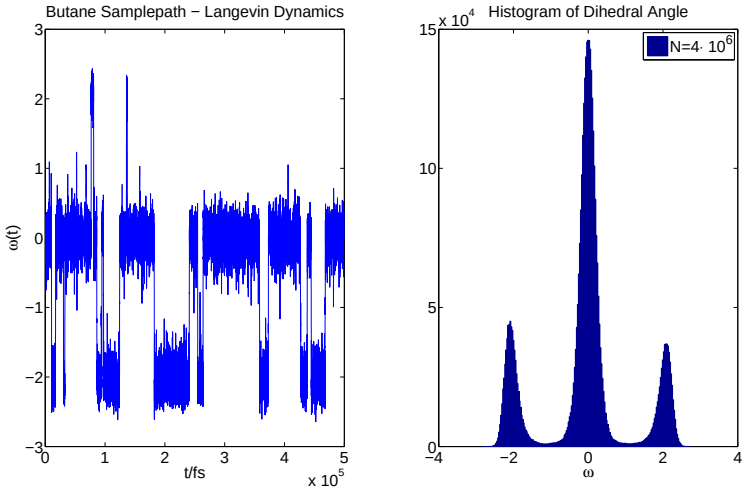

FIG. 7.2. *MD Samplepath: Butane*

7.1. Molecular Dynamics. The data used for this fitting example are generated using a molecular dynamics (MD) simulation for a single molecule of Butane. In order to avoid explcit computations for solvent molecules, several *ad hoc* approximate algorithms have been developed in molecular dynamics. One of the more sweeping approximation that is nonetheless fairly popular, at least as long as electrostatic effects of the solvent can be neglected or treated otherwise, is Langevin dynamics. The butane molecule is modelled as a damped-driven Hamiltonian system of the form

$$\ddot{x} = \nabla V(x) + \gamma \dot{x} + \sigma \dot{B}. \tag{7.1}$$

The coordinate x in this equation stands for cartesian coordinates of the four extended atoms making up the butane molecule, see [9] for details of the CHARMM forcefield used here.

From a chemical point of view interest is focused on the dihedral angle, which is the angle between the two planes in $\mathbb{R}^3$ formed by atoms $1, 2, 3$ and atoms $2, 3, 4$ respectively; see the sketch in figure 7.1. Conformational change is manifest in this angle, and the cartesian coordinates themselves are of little direct chemical interest. Hence it is natural to try and describe the stochastic dynamics of the dihedral angle in a self-contained fashion.

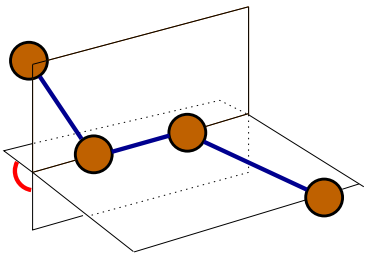

FIG. 7.1. *Sketch of Dihedral Angle*

One MD run is produced using a timestep of $\Delta t = 10^{-16}$s (one tenth of a femtosecond) and a Verlet variant (see p.435 in [31]) covering a total time of $T = 4 \cdot 10^{-9}$s (4 nanoseconds). A section of path of the dihedral angle versus time can be seen on the left of figure 7.2; the corresponding histogram is depicted to the right of that figure. It is known ([10]) that the stationary distribution of (7.1) is given by the canonical distribution associated with the torsional potential, so that an explicit analytical representation can easily be obtained.

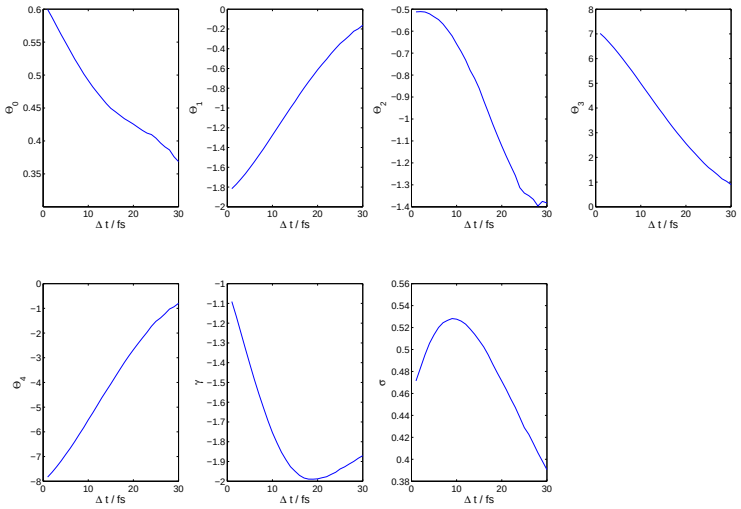

FIG. 7.3. *Convergence for fitted MD path with subsampling*

It should be stressed that the effective stochastic differential equation governing the behaviour of the dihedral angle ω is *not* of the form (2.5), in particular, it will have a non-constant diffusivity σ. So, fitting to this data tests the robustness of the fitting algorithm in a way that the experiments in previous sections did not.

7.2. Fitting. The physical time-units in seconds are miniscule and do not lead to SDE parameter fits of order one. It transpires that, in order to obtain parameter values of order one, re-scaling 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.

In order to assess consistency, the MD data is subsampled, at timesteps $\Delta t \in \{1{\cdot}10^{-15}\text{s}, 2{\cdot}10^{-15}\text{s}, 3{\cdot}10^{-15}\text{s} \ldots\}$ in physical time units, corresponding to $\{k0.02\}_{k\in\mathbb{N}}$ in the rescaled time units. The Deterministic Scan Gibbs sampler is then run for $N_{\text{Gibbs}} = 40$ outer iterations on each path using a potential ansatz

$$V(\omega) = \sum_{k=1}^{c} \theta_k cos^k(\omega)$$

where $c \in \{3, 5, 7\}$ is used. This corresponds to a choice of the force function in (2.5). The obtained drift parameters under subsampling at timestep Δt can be seen from figure 7.3. This plot shows the behaviour of the drift parameters averaged over $N_{\text{Gibbs}} = 100$ Monte-Carlo samples $\theta_1, \ldots, \theta_5, \gamma$ as the subsampling rate is increased. Below a subsampling rate $k = 20$, behaviour consistent with $\mathcal{O}(\Delta t)$ errors is observed indicating convergence of the algorithm as Δt is decreased. This is exactly the behaviour observed on simulated data and it is a measure of the robustness of the proposed algorithm.

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 MD data gives a good or a bad fit, it merely indicates that, using the algorithm suggested in this paper, it is possible to perform such fitting.

To show limitations of the model in this particular application and see how the performance can be assessed using the fitting algorithm from section 6.2, we show posterior invariant probability densities resulting from the fitted trigonometric potentials. In order to do this, we convert the drift parameter samples $\{D_j^{(m)}\}_{j=1}^c$ obtained at step m using input data subsampled at rate $k = 1$ to an invariant density, $\varrho^{(m)}$ 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 the plot given in figure 7.4. There, we display results for three orders of trigonometric potential c to be fitted and contrast this with the empirically observed invariant density and the density arising from the classic canonical thermodynamic ensemble which is proportional to $\exp\left(-\frac{V(\omega)}{kT}\right)$. For the parameterisation used here, it is known that the latter two agree in the limit $T \to \infty$, see [10].

With increasing polynomial order c we find some qualitative change in the resulting probability 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 (5.4). It is simply an ill-conditioned problem to derive higher and 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.

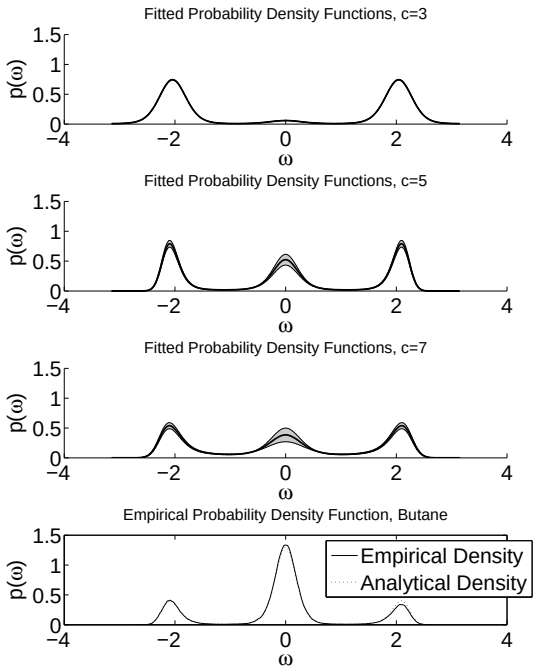

FIG. 7.4. *PDFs resulting from fitted potentials for different orders of trigonometric potential - Shaded regions display posterior variance*

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. Its performance has been validated numerically for a number of test cases and an application to molecular dynamics data has been given. While parameter fitting can be viewed as an inverse problem for SDE solvers – and thus ill-conditioning of some kind is to be expected – a detailed understanding of the ill-conditioning induced by hypoellipticity and partial observation has been attained.

While only second order hypoelliptic problems have been treated in this article, the algorithm's applicability is expected to encompass order k hypoelliptic problems and it has been tested successfully on a third order example. Furthermore, nonlinear p-dependence in the example (2.1) can be dealt with using a Langevin sampler for the missing path and this has also been tested.

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 [30].

Also, an extension to position dependent diffusion coefficients may prove useful, in particular, in may render the algorithm more useful in molecular dynamics contexts such as those in [18].

9. Appendix. We explain how the exact sampler (4.4) is derived. The Langevin equation used to sample from the distribution of p (given drift parameters and σ) is:

$$\frac{dp}{ds} = P_{\text{mat}}p + Q(q) + \sqrt{2}\frac{dW}{ds} \tag{9.1}$$

Here, W consists of N independent standard white noise processes and $p = p(s)$ is thought of as a function

$$p : [0, \infty) \longrightarrow \mathbb{R}^N$$

This equation is continuous in time but discrete in space. Given that the derivative of $\log \mathcal{L}_{IT}$ is linear in the p_i, (9.1) is recognised as an Ornstein-Uhlenbeck process, so

that the equilibrium measure is expressible as follows:

$$p \sim \mathcal{N}(-P_{\text{mat}}^{-1}Q(q), -P_{\text{mat}}^{-1}) \tag{9.2}$$

Given a computer-generated pseudo-random i.i.d. sequence of normally distributed random variables, $\{\xi_n\}$, one can generate independent samples with the desired distribution, if the root of the covariance matrix is available, simply by setting:

$$p_n = -P_{\text{mat}}^{-1}Q_{\text{mat}}q + \sqrt{-P_{\text{mat}}^{-1}}\xi_n.$$

As noted above, $-P_{\text{mat}}^{-1}$ is positive definite symmetric. We may thus compute the Cholesky factorisation $U^TU = -P_{\text{mat}}$ and use the following observation which yields

$$\begin{aligned}\mathbb{E}\left(U^{-1}\xi\left(U^{-1}\xi\right)^T\right) &= U^{-1}IU^{-T}\\ &= U^{-1}U^{-T}\\ &= -P_{\text{mat}}^{-1}\end{aligned}$$

as desired.

The suggested sampler for p-paths is then (4.4). Since a Cholesky factorisation of P_{mat} is an efficient way to compute the mean, the application of U^{-1} is a just a backsubstitution using the already computed Cholesky factor.

A cautionary note from Trefethen ([33], p.177) shows that while solving the linear system for P^{-1} is backward stable, the computation of the factor U is not forward-stable, i.e. the errors in U might be large for a generic positive definite matrix. In our case, P is very well-conditioned (Gershgorin yields an upper bound for its condition number with respect to the 2-norm of $\kappa(P) < 3 + \mathcal{O}(\Delta t)$) so that we expect U to be computed accurately. Employing a combination of Theorem 10.5 for stability and Theorem 10.8 for conditioning of the Cholesky factor from [17] this can be substantiated.

Now we compute derivatives of the approximate likelihood $\mathcal{L}_{IT}$ needed for a Langevin sampler of the missing path p and for the resulting exact sampler (4.4). We have

$$\begin{aligned}-\sigma^2\frac{\partial\mathcal{L}}{\partial p_i} = {} & q_{i+1}\left(\frac{6}{b\Delta t}(\gamma - \Delta t^{-1})\right)\\ & + q_i\left(-(1+\Delta ta)\frac{6}{b\Delta t}(\gamma - \frac{1}{\Delta t}) - \Delta tD(2\Delta t^{-1} - 4\gamma) - 6b^{-1}\Delta t^{-2}\right)\\ & + q_{i-1}\left((1+\Delta ta)6b^{-1}\Delta t^{-2} - 4D\right)\\ & + p_{i+1}\left(2\Delta t^{-1} - 4\gamma\right)\\ & + p_i\left(6(\Delta t^{-1} - \gamma)(2\Delta t^{-1} - 4\gamma)(-1 - \Delta t\gamma) + 4\Delta t^{-1}\right) + p_{i-1}\left(2\Delta t^{-1} - 4\gamma\right)\end{aligned}$$

at inner points $0 < i < N$. At the boundary points one gets:

$$\begin{aligned}-\sigma^2\frac{\partial\mathcal{L}}{\partial p_0} = {} & q_0\left(-(1+\Delta ta)(6b^{-1}\Delta t^{-1}\gamma - 6b^{-1}\Delta t^{-2}) - 2D + 4\gamma D\Delta t\right)\\ & + q_1\left(6b^{-1}\Delta t^{-1}\gamma - 6b^{-1}\Delta t^{-2}\right)\\ & + p_0\left(-\Delta tb(-6b^{-1}\Delta t^{-2} + 6b^{-1}\Delta t^{-1}\gamma) - (1+\Delta\gamma)(2\Delta t^{-1} - 4\gamma)\right)\\ & + p_1\left(2\Delta t^{-1} - 4\gamma\right)\end{aligned}$$

And for $i = N$:

$$-\sigma^2 \frac{\partial \mathcal{L}}{\partial p_N} = q_{N-1}\left((1+\Delta t a)6b^{-1}\Delta t^{-2} - 4D\right) + q_N\left(-6b^{-1}\Delta t^{-2}\right) + p_{N-1}\left(6\Delta t^{-1} - (1+\Delta t\gamma)4\Delta t^{-1}\right) + p_N\left(4\Delta t^{-1}\right)$$

REFERENCES

[1] Y. Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation. *Econometrica*, 70(1):223–262, 2002.

[2] B. M. Bibby and M. Sørensen. On estimation for discretely observed diffusions: A review. *Theory of Stochastic Processes*, 2(18):49–56, 1996.

[3] A. Le Breton and M. Musiela. Some parameter estimation problems for hypoelliptic homogeneous gaussian diffusions. *Seq. Meth. in Stat.*, 22:337–356, 1985.

[4] D. E. Catlin. *Estimation, Control and the Discrete Kalman Filter*. Springer-Verlag, 1989.

[5] D. Dacunha-Castelle and D. Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19(4):263–284, 1986.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data with the EM algorithm. *J. R. Stat. Soc., Ser. B*, 39(1):1–38, 1977.

[7] R. Durrett. *Stochastic Calculus - A practical Introduction*. CRC Press, London, 1996.

[8] A. Beskos et al. Exact and computationally efficient estimation for discretely observed diffusion processes. *J. R. Statist. Soc. B*, 68(2):1–29, 2006.

[9] B. R. Brooks et al. Charmm: A program for macromolecular energy, minmization and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[10] A. Fischer. *Die Hybride Monte-Carlo-Methode in der Molekülphysik*. FU Berlin, Diplomarbeit, 1997.

[11] D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989.

[12] J. G. Gaines and T. J. Lyons. Variable step size control in the numerical solution of stochastic differential equations. *SIAM J. Appl. Math.*, 57(5):1455–1484, 1997.

[13] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer, 1985.

[14] V. Genon-Catalot and J. Jacod. On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. Henri Poincaré*, 29(1):119–151, 1993.

[15] P. Giannopoulos and S. J. Godsill. Estimation of car processes observed in noise using bayesian inference. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, available from http://www-com-serv.eng.cam.ac.uk/%7Esjg/pubs/pubs_noabst.html*, 2001.

[16] H. Grubüller and P. Tavan. Molecular dynamics of conformational substates for a simplified protein model. *J. Chem. Phys.*, 101(6):5047–5057, 1994.

[17] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 1996.

[18] G. Hummer. Position-dependent diffusion coefficients and free energies from bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7(34), 2005.

[19] R. E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, March 1960.

[20] R. Khasminskii, N. Krylov, and N. Moshchuk. On the estimation of parameters for linear stochastic differential equations. *Probab. Theory Related Fields*, 113(3):443–472, 1999.

[21] P. E. Kloeden, E. Platen, H. Schurz, and M. Sørensen. On effects of discretization on estimators of drift parameters for diffusion processes. *J. Appl. Prob.*, 33:1061–1076, 1996.

[22] H. A. Kramers. *Physica*, 7(284), 1940.

[23] A. Lasota and M. C. Mackey. Springer, 1994.

[24] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Verlag, 2001.

[25] X.-L. Meng and D. van Dyk. The EM algorithm—an old folk-song sung to a fast new tune. *J. R. Stat. Soc., Ser. B*, 59(3):511–567, 1997.

[26] D. Nualart. *The Malliavin Calculus and Related Topics*. Springer-Verlag, 1991.

[27] B. Oksendal. *Stochastic Differential Equations, An Introduction with Applications*. Springer Verlag, 2000.

[28] A. R. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, 22:55–71, 1995.

[29] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 1999.

[30] G. O. Roberts and O. Stramer. On inference for nonlinear diffusion models using the hastings-metropolis algorithms. *Biometrika*, 88(3):603–621, 2001.
[31] T. Schlick. *Molecular Modeling Simulation - An Interdisciplinary Guide*. Springer, 2000.
[32] I. Shoji and T. Ozaki. Comparative study of estimation methods for continuous time stochastic processes. *J. Time Ser. Anal.*, 18(5):485–506, 1997.
[33] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM, 1997.