Book chapter (18C)
Semstat Proceedings 2007 (Pages 429 – 472)
Article Title: Parameter estimation for multiscale diffusions: An overview
Article Authors: Pavliotis, Pokern & Stuart.

CHAPTER 7

Parameter estimation for multiscale diffusions: An overview

Grigorios A. Pavliotis, Yvo Pokern and Andrew M. Stuart

Department of Mathematics, Imperial College London
South Kensington Campus, London SW7 2AZ, United Kingdom
and
Department of Statistical Science, University College London
Gower Street, London WC1E 6BT, United Kingdom
and
Mathematics Department, University of Warwick
Coventry CV4 7AL , United Kingdom

7.1 Introduction

There are many applications where it is desirable to fit reduced stochastic descriptions (e.g. SDEs) to data. These include molecular dynamics (Schlick (2000), Frenkel and Smit (2002)), atmosphere/ocean science (Majda and Kramer (1999)), cellular biology (Alberts et al. (2002)) and econometrics (Dacorogna, Gençay, Müller, Olsen, and Pictet (2001)). The data arising in these problems often has a multiscale character and may not be compatible with the desired diffusion at small scales (see Givon, Kupferman, and Stuart (2004), Majda, Timofeyev, and Vanden-Eijnden (1999), Kepler and Elston (2001), Zhang, Mykland, and Aït-Sahalia (2005) and Olhede, Sykulski, and Pavliotis (2009)). The question then arises as to how to optimally employ such data to find a useful diffusion approximation.

The types of data available and the pertinent scientific questions depend on the particular field of application. While this chapter is about multiscale phenomena *common* to the above fields of application, we detail the type of data available and the pertinent scientific questions for the example of molecular dynamics.

Molecular dynamics data usually arises from large scale simulation of a high-dimensional Hamiltonian dynamical system, which stems from an approximation of molecular processes by, essentially, classical mechanics. The simulations can be deterministic or stochastic in nature, and applied interest usually focuses on some chemically interesting coordinates, the reaction coordinates, which are of much lower dimension than the simulated system. Such data typically evolves on a large range of timescales from fast and small vibrations of the distance between neighbouring atoms joined by a chemical bond (so-called bond-length vibrations) with characteristic timescale $t \approx 10^{-13}$s to large-scale conformational changes, like the folding of a protein molecule on timescales of at least $t \approx 10^{-6}$s. This creates an extremely challenging computational problem. See Schlick (2000) for an accessible overview of this application area.

Molecular dynamics data can be available at inter-observation times as low as $t \approx 10^{-15}$s but because the data may itself be deterministic it is clear that successfully fitting a stochastic model at those timescales is unlikely. At slightly larger timescales, fits to SDEs are routinely attempted and fitting the special class of hypoelliptic SDEs can be advantageous, as it allows for some smoothness (i.e. the paths being of greater regularity than that e.g of Brownian motion) of the input path as well as imposing physically meaningful structures, like that of a damped-driven Hamiltonian system.

Furthermore, as the diffusivity is most affected by information from small timescales, it is interesting to note that in non-parametric drift estimation, local time (or, more generally, the empirical measure) can be an almost sufficient statistic, so that time-ordering of the data is not relevant and hence, drift estimation performed in this way will be less affected by inconsistencies at small timescales.

It may also be advantageous to model the separation in timescales between e.g. bond-length vibrations and large scale conformational changes explicitly by a system of SDEs operating at different timescales. In the limit of infinite separation of these timescales, effective SDEs for the slow process can be derived through the mathematical techniques of averaging and homogenization for diffusions.

If the fitted SDEs are of convenient type, it is then possible to glean information of applied interest, concerning e.g. effective energy barriers to a conformational transition, relative weights of transition paths, number and importance of metastable states, etc.

We illustrate the issues arising from multiscale data, first through studying some illustrative examples in Section 7.2, including a toy-example from molecular dynamics, and then more generally in the context of averaging and homogenization for diffusions in Section 7.3. In Section 7.4 we show how subsampling may be used to remove some of the problems arising from multiscale data. Sections 7.5 and 7.6 treat the use of hypoelliptic diffusions and ideas stemming from non-parametric drift estimation, respectively.

The material in this overview is based, to a large extent, on the papers of Papaspiliopoulos, Pokern, Roberts, and Stuart (2009), Papavasiliou, Pavliotis, and Stuart (2009), Pavliotis and Stuart (2007), Pokern, Stuart, and Vanden-Eijnden (2009), and Pokern, Stuart, and Wiberg (2009). We have placed the material in a common framework, aiming to highlight the interconnections in this work. The details, however, are in the original papers, including the proofs where we do not provide them.

7.2 Illustrative examples

In this section we start with four examples to illustrate the primary issue arising in this context. To understand these examples it is necessary to understand the concept of the quadratic variation process for a diffusion. Consider the stochastic differential equation (SDE)

$$\frac{dz}{dt} = h(z) + \gamma(z)\frac{dW}{dt}, \quad z(0) = z_0. \tag{7.1}$$

Here $z(t) \in \mathcal{Z}$ with $\mathcal{Z} = \mathbb{R}^d$ or $\mathbb{T}^d$, the d-dimensional torus*. We assume that h, γ are Lipschitz on $\mathcal{Z}$. To make the notion of solution precise we let $\mathcal{F}_t$ denote the filtration generated by $\{W(s)\}_{0\leq s\leq t}$ and define $z(t)$ to be the unique $\mathcal{F}_t$-adapted process which is a semimartingale defined via the integral equation

$$z(t) = z_0 + \int_0^t h(z(s))ds + \int_0^t \gamma(z(s))dW(s). \tag{7.2}$$

The stochastic integral is interpreted in the Itô sense. It is an $\mathcal{F}_t-$martingale and we write

$$m(t) = \int_0^t \gamma(z(s))dW(s). \tag{7.3}$$

A matrix valued process $Q(t)$ is increasing if $Q(t) - Q(s)$ is non-negative for all $t \geq s \geq 0$. The quadratic variation process of z, namely $\langle z\rangle_t := Q(t)$ is defined as the unique adapted, increasing and continuous process for which

$$m(t)m(t)^T - Q(t)$$

is an $\mathcal{F}_t-$martingale, see Da Prato and Zabczyk (1992) for a definition. The quadratic variation is non-zero precisely because of the lack of regularity of sample paths of diffusion processes. It is given by the expression

$$Q(t) = \int_0^t \gamma(z(s))\gamma(z(s))^T ds.$$

* See Appendix 2 for a definition.

It is possible to give a corresponding differential statement, namely that

$$\lim_{\tau \to 0} \sum_{i=0}^{n-1} \left(z(t_{i+1}) - z(t_i)\right) \left(z(t_{i+1}) - z(t_i)\right)^T = Q(T) \quad \text{in prob.} \tag{7.4}$$

where the ordered times $t_0 = 0 < t_1 < \ldots < t_{n-1} < t_n = T$ on $[0, T]$ are such that their largest distance $\tau = \max_{i=1,\ldots n}(t_i - t_{i-1})$ decreases like $\mathcal{O}(1/n)$. An easily accessible treatment for one-dimensional continuous local martingales including this result (Theorems 2.3.1 and 2.3.8) is given in Durrett (1996). The issue of the required decay of τ is treated more carefully in Marcus and Rosen (2006).

If γ is constant then

$$Q(t) = t\gamma\gamma^T.$$

Notice that, for ordinary differential equations (ODEs) where $\gamma \equiv 0$ the quadratic variation is zero. The definition we have given here may be generalized from solutions of SDEs to Itô processes where the drift depends upon the past history of $z(t)$. In this fashion, it is possible to talk about the quadratic variation associated with a single component of a system of SDEs in several dimensions.

7.2.1 Example 1. SDE from ODE

This example is taken from Melbourne and Stuart (2011).

Consider the scale-separated system of ODEs

$$\begin{aligned} \frac{dx}{dt} &= \frac{1}{\epsilon} f_0(y) + f_1(x, y), \quad x(0) = \xi, \\ \frac{dy}{dt} &= \frac{1}{\epsilon^2} g_0(y), \quad y(0) = \eta, \end{aligned}$$

where $x \in \mathbb{R}^d$. We make some technical assumptions on y (detailed in Melbourne and Stuart (2011)) which imply that it is mixing with invariant measure μ. We assume that

$$\mathbb{E}^\mu f_0(y) = 0.$$

This ensures that the first term on the right-hand side of the equation for x gives rise to a well-defined limit as $\epsilon \to 0$. In fact this term will be responsible for the creation of white noise in the limiting equation for x. The technical assumptions imply that

$$\frac{1}{\epsilon} \int_0^t f_0(y(s))ds \Rightarrow \sqrt{2\Sigma} W(t)$$

for some covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, and W a standard d-dimensional Brownian motion. Here, $\Rightarrow$ denotes weak convergence in $C([0, T], \mathbb{R}^d)$ and we use uppercase letters for effective quantities (like diffusivity) arising from averaging or homogenization here and throughout the remainder of the chapter.

Intuitively, this weak convergence follows because y has correlation decay with timescale $1/\epsilon^2$ so that $\frac{1}{\epsilon}f_0(y)$ has an autocorrelation function which approximates a Dirac delta distribution.

Now define

$$F(x) = \mathbb{E}^{\mu} f_1(x, y),$$

which will become the mean drift in the limiting equation for x.

Theorem 7.1 *(Melbourne and Stuart (2011)) Let $\eta \sim \mu$. Then, under some technical conditions on the fast process y, $x \Rightarrow X$ in $C([0,T], \mathbb{R}^d)$ as $\epsilon \to 0$, where*

$$\frac{dX}{dt} = F(X) + \sqrt{2\Sigma}\frac{dW}{dt}, \quad X(0) = \xi.$$

The important point that we wish to illustrate with this example is that the limit of X, and x itself, have vastly different properties at small scales. In particular, the quadratic variations differ:

$$\langle x \rangle_t = 0; \quad \langle X \rangle_t = 2\Sigma t.$$

Any parameter estimation procedure which attempts to fit an SDE in X to data generated by the ODE in x will have to confront this issue. Specifically, any parameter estimation procedure which sees small scales in the data will have the potential to *incorrectly* identify an appropriate SDE fit to the data.

7.2.2 Example 2. Smoluchowski from Langevin

The situation arising in the previous example can also occur when considering scale-separated SDEs. We illustrate this with a physically interesting example taken from Papavasiliou et al. (2009): consider the *Langevin equation* for $x \in \mathbb{R}^d$:

$$\epsilon^2 \frac{d^2 x}{dt^2} + \frac{dx}{dt} + \nabla V(x) = \sqrt{2\sigma}\frac{dW}{dt}.$$

As a first order system this is

$$\begin{aligned} \frac{dx}{dt} &= \frac{1}{\epsilon} y, \\ \frac{dy}{dt} &= -\frac{1}{\epsilon}\nabla V(x) - \frac{1}{\epsilon^2} y + \sqrt{\frac{2\sigma}{\epsilon^2}}\frac{dW}{dt}. \end{aligned}$$

Using the method of homogenization the following may be proved:

Theorem 7.2 *(Pavliotis and Stuart (2008)) As $\epsilon \to 0$, $x \Rightarrow X$ in $C([0,T],\mathbb{R}^d)$ where X is the solution of the Smoluchowski equation* [†]

$$\frac{dX}{dt} = -\nabla V(X) + \sqrt{2\Sigma}\,\frac{dW}{dt}.$$

Thus, similarly to the previous example,

$$\langle x \rangle_t = 0; \quad \langle X \rangle_t = 2\Sigma t.$$

Again, any parameter estimation procedure for the SDE in X, and which sees small scales in the data x, will have the potential to incorrectly identify an appropriate homogenized SDE fit to the data.

7.2.3 Example 3. Butane

The two previous examples both possessed a known effective equation which may be untypical of many practical applications. In this example, an effective equation is not known; neither is the range of timescales at which such an equation would be approximately valid. This can, however, be assessed empirically to some extent, which is typical of practical applications of multiscale diffusions. The data presented in this example is taken from Pokern (2006). We consider a classical molecular dynamics model for butane. The model comprises the positions x and momenta of four (extended) atoms interacting with one another through various two, three and four body interactions, such as bond angle, bond stretch and dihedral angle interactions all combined in a potential $V(x)$. The equations of motion are

$$M\frac{d^2x}{dt^2} + \gamma M\frac{dx}{dt} + \nabla V(x) = \sqrt{2\gamma k_B T M}\,\frac{dB}{dt}. \tag{7.5}$$

The choice $\gamma = 0$ gives deterministic dynamics, whilst for $\gamma > 0$ stochastic dynamics are obtained. A typical configuration of the molecule is shown in Figure 7.1. The *dihedral angle* ϕ is the angle formed by intersecting the planes passing through the first three atoms and through the last three atoms respectively. The molecule undergoes conformational changes which can be seen in changes in the dihedral angle. This is shown in Figure 7.2 which exhibits a time series for the dihedral angle, as well as its histogram, for $\gamma > 0$. Clearly the dihedral angle has three preferred values, corresponding to three different molecular conformations. Furthermore, the transitions between these states is reminiscent of thermally activated motion in a three well potential. This fact concerning the time series remains true even when $\gamma = 0$.

[†] In molecular dynamics, this equation would be termed *Brownian Dynamics*, whereas in the statistical literature it is sometimes called the *Langevin equation*.

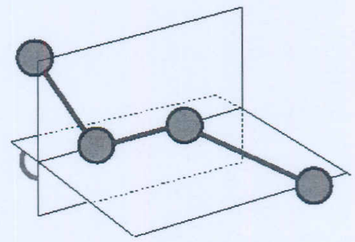


Figure 7.1 *The butane molecule*

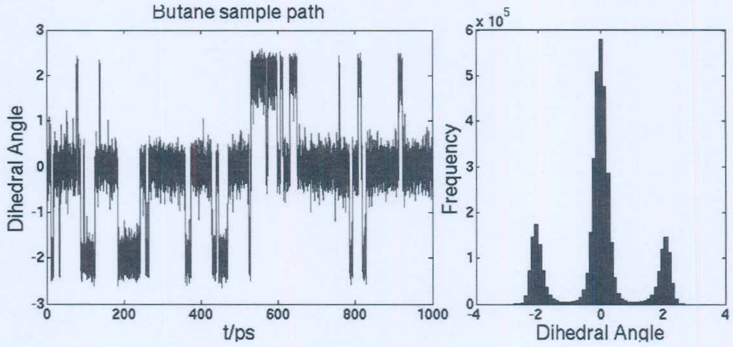


Figure 7.2 *The dihedral angle (time series and histogram)*

It is hence natural to try and fit an SDE to the dihedral angle, and we consider an SDE of the form

$$\frac{d\Phi}{dt} = -\Psi'(\Phi) + \sigma\frac{dW}{dt}, \tag{7.6a}$$

$$\Psi(\phi) = \sum_{j=1}^{5} \theta_j \Big(\cos(\phi)\Big)^j, \tag{7.6b}$$

where $\{\theta_j\}_{j=1}^{5}$ and σ are the parameters to be estimated. However, the fit to such an SDE is highly sensitive to the rate at which the data is sampled, as shown in Figure 7.3. Here the diffusion coefficient is estimated exploiting (7.4), and the maximum likelihood principle is used to estimate θ.

The behaviour of the estimator at different scales is caused by the fact that the data is again incompatible with the diffusion approximation at small scales. The true dihedral angle ϕ has zero quadratic variation because it is a function of the positions of x only, and not the momenta $M\dot{x}$; only the momenta are directly forced by noise. Thus $\langle\phi\rangle_t = 0$. In contrast, for the model (7.6), we have

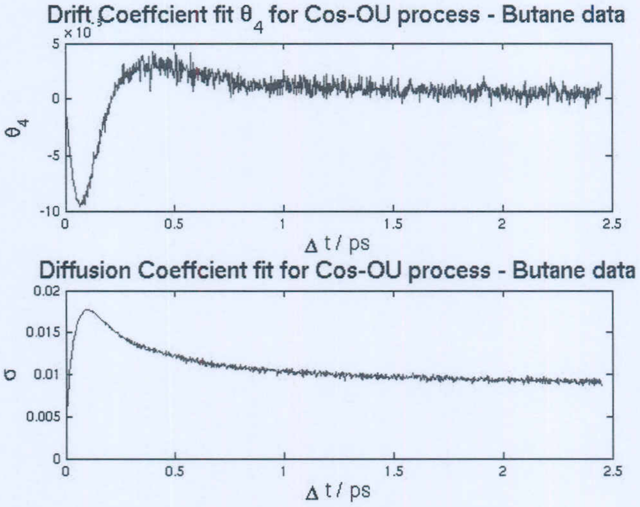


Figure 7.3 *Parameter fits to* θ_4 *and* σ *from* (7.6) *given data from Figure 7.2.*

$\langle\Phi\rangle_t = \sigma^2 t$. Notice that the estimated σ in Figure 7.3 tends to zero as the sampling rate increases, reflecting the smoothness of the data at small scales. When the invariant measure is to be preserved, a direct link between the maximum likelihood estimator for the drift to the maximum likelihood estimator for the diffusion arises for this diffusion, see Pavliotis and Stuart (2007). Therefore, the drift estimator is inconsistent between different scales of the data, too.

The situation is even more complex in the case where data is taken from (7.5) with $\gamma = 0$, a Hamiltonian ODE – see Figure 7.4 for a typical time series and Figure 7.5 for estimated diffusivities. Again the data is inconsistent at small scales, leading to estimates of drift and diffusion which tend to zero. But at intermediate scales oscillations caused by bond length stretches between atoms cause inflated, large diffusion coefficient estimates – a resonance effect.

7.2.4 Example 4. Thermal motion in a multiscale potential

All of the previous examples have the property that the quadratic variation of the data at finest scales is zero, and this is incompatible with the assumed diffusion. Here we present an example (taken from Pavliotis and Stuart (2007)) where, at small scales the quadratic variation of the data is *larger* than that of the desired model to be fitted to the data.

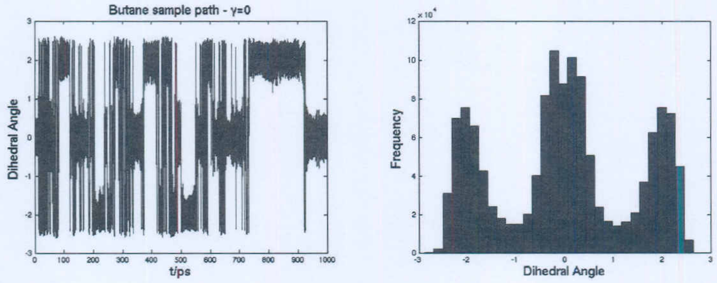
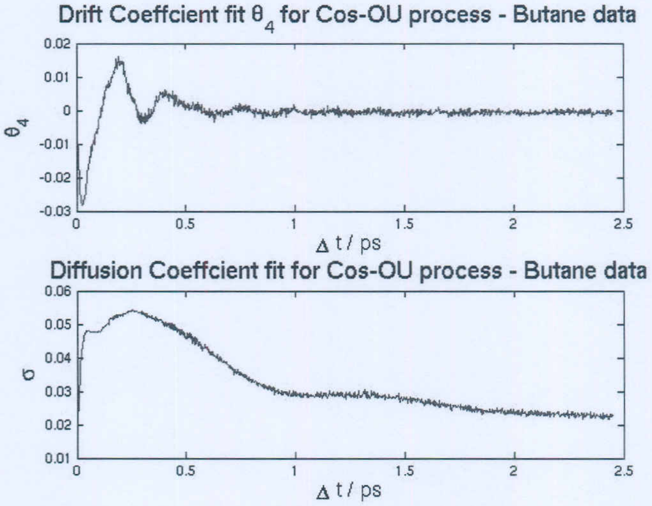

Butane sample path - γ=0

Figure 7.4 *The dihedral angle time series for $\gamma = 0$*


Drift Coeffcient fit θ$_4$ for Cos-OU process - Butane data

Diffusion Coeffcient fit for Cos-OU process - Butane data

Figure 7.5 *Parameter fits to σ from* (7.6) *given data from* (7.5) *with $\gamma = 0$*

Consider the equation

$$\frac{dx}{dt} = -K^{\epsilon}(x)\nabla V^{\epsilon}(x) + \sqrt{2K^{\epsilon}(x)}\,\frac{dW}{dt}, \tag{7.7}$$

where $x \in \mathbb{R}^d$. Here, for K, p 1−periodic functions we define

$$K^{\epsilon}(x) = K(x/\epsilon),$$
$$V^{\epsilon}(x) = V(x) + p(x/\epsilon).$$

We also assume that K is symmetric positive-definite, so that its square-root

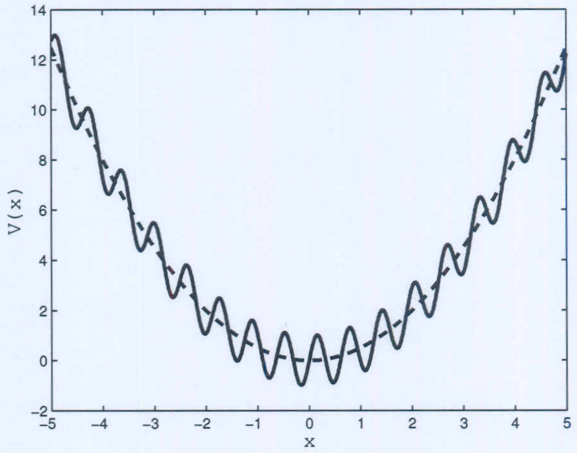


Figure 7.6 $V^\epsilon(x)$ *and* $V(x)$.

is defined, and divergence-free in the sense that each row of the matrix K is a vector with zero divergence.

By applying standard techniques from homogenization theory it is possible to prove the following theorem:

Theorem 7.3 *(Pavliotis and Stuart (2008)) There exist matrices* $\overline{K}$ *and* K^* *such that as* $\epsilon \to 0$, $x \Rightarrow X$ *in* $C([0,T],\mathbb{R}^d)$, *where* X *satisfies the equation*

$$\frac{dX}{dt} = -\overline{K}\nabla V(X) + \sqrt{2\overline{K}}\,\frac{dW}{dt}. \tag{7.8}$$

Furthermore

$$\lim_{\epsilon\to 0}\int_0^T \langle x\rangle_t dt = K^\star T, \qquad \int_0^T \langle X\rangle_t dt = \overline{K}T; \quad \overline{K} < K^\star.$$

Here, $\overline{K} < K^\star$ means that $\left(K^\star - \overline{K}\right)$ is positive definite. A typical potential is shown in Figure 7.6. Notice that the rapid oscillations persist at $\mathcal{O}(1)$ in the limit $\epsilon \to 0$. Thus, although the oscillatory part of the potential, p, is not present in the homogenized equation, its effect is felt in slowing down the diffusion process, as the particle must cross the energy barriers caused by this oscillation.

Once again we expect that parameter estimation which sees the small scales will fail. However, the situation differs from the three previous examples: here the quadratic variation of the data x is *larger* than that of the desired homogenized model X.

7.3 Averaging and homogenization

7.3.1 *Orientation*

In this section we probe the phenomenon exhibited in the preceding examples by means of the study of scale-separated SDEs. These provide us with a set of model problems which can be used to study the issues arising when there is a mismatch between the statistical model and the data at small scales. The last example in the preceding section illustrates such a model problem. Many coupled systems for (x, y) contain a parameter $\epsilon \ll 1$ which characterizes scale-separation. If y evolves more quickly than x, then it can sometimes be eliminated to produce an equation for $X \approx x$ alone. Two important situations where this arises are *averaging* and *homogenization*. We will be interested in fitting parameters in an effective averaged or homogenized equation for X, given data x from the coupled system for x, y. All proofs from this section may be found in the paper by Papavasiliou et al. (2009).

7.3.2 *Set-up*

In the following we set $\mathcal{X} = \mathbb{T}^l$ and $\mathcal{Y} = \mathbb{T}^{d-l}$. Let $\varphi_\xi^t(y)$ denote the Markov process which solves the SDE

$$\frac{d}{dt}\Big(\varphi_\xi^t(y)\Big) = g_0\big(\xi, \varphi_\xi^t(y)\big) + \beta\big(\xi, \varphi_\xi^t(y)\big)\frac{dV}{dt}, \quad \varphi_\xi^0(y) = y. \tag{7.9}$$

Here $\xi \in \mathcal{X}$ is a fixed parameter and, for each $t \geq 0$, $\varphi_\xi^t(y) \in \mathcal{Y}$, $g_0 : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{d-l}$, $\beta : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{(d-l)\times m}$ and V is a standard $m-$dimensional Brownian motion. The generator of the process is

$$\mathcal{G}_0(\xi) = g_0(\xi, y) \cdot \nabla_y + \frac{1}{2}B(\xi, y) : \nabla_y \nabla_y \tag{7.10}$$

with $B(\xi, y) := \beta(\xi, y)\beta(\xi, y)^T$ and equipped with periodic boundary conditions and : denotes the matrix inner product, see Appendix 2. Notice that $\mathcal{G}_0$ is a differential operator in y alone.

Our interest is in data generated by the projection onto the x coordinate of systems of SDEs for (x, y) in $\mathcal{X} \times \mathcal{Y}$. In particular, for U, a standard Brownian motion in $\mathbb{R}^n$, we will consider either of the following coupled systems of

SDEs:

$$\frac{dx}{dt} = f_1(x,y) + \alpha_0(x,y)\frac{dU}{dt} + \alpha_1(x,y)\frac{dV}{dt}, \tag{7.11a}$$
$$\frac{dy}{dt} = \frac{1}{\epsilon}g_0(x,y) + \frac{1}{\sqrt{\epsilon}}\beta(x,y)\frac{dV}{dt}; \tag{7.11b}$$

or the SDEs

$$\frac{dx}{dt} = \frac{1}{\epsilon}f_0(x,y) + f_1(x,y) + \alpha_0(x,y)\frac{dU}{dt} + \alpha_1(x,y)\frac{dV}{dt}, \tag{7.12a}$$
$$\frac{dy}{dt} = \frac{1}{\epsilon^2}g_0(x,y) + \frac{1}{\epsilon}g_1(x,y) + \frac{1}{\epsilon}\beta(x,y)\frac{dV}{dt}. \tag{7.12b}$$

Here $f_i : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^l, \alpha_0 : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^{l\times n}, \alpha_1 : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^{l\times m}$, $g_1 : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}^{d-l}$ and g_0, β are as above. Note that in (7.11) (resp. (7.12)) the equation for y with x frozen has solution $\varphi_x^{t/\epsilon}(y(0))$ (resp. $\varphi_x^{t/\epsilon^2}(y(0))$ with $g_1 = 0$). Of course x is not frozen, but since it evolves much more slowly than y, intuition based on freezing x and considering the process (7.9) is useful. In addition to the generator $\mathcal{G}_0$, we also define the operator $\mathcal{G}_1$ as follows:

$$\mathcal{G}_1 = f_0 \cdot \nabla_x + g_1 \cdot \nabla_y + C : \nabla_y \nabla_x,$$

where the matrix-valued function C is defined as

$$C(x,y) = \alpha_1(x,y)\beta(x,y)^T.$$

Assumptions 7.4 • *All the functions f_i, g_i, α_i and β are C^∞ on the torus $\mathbb{T}^d$.*

- *The equation*

$$-\mathcal{G}_0^*(\xi)\rho(y;\xi) = 0, \quad \int_{\mathcal{Y}} \rho(y;\xi)dy = 1$$

has a unique non-negative solution $\rho(y;\xi) \in L^1(\mathcal{Y})$ for every $\xi \in \mathcal{X}$; furthermore $\rho(y;\xi)$ is C^∞ in y and ξ. Here as throughout, we use $\cdot^*$ to denote the adjoint operator.

- *Define the weighted Hilbert space $L^2_\rho(\mathcal{Y})$ with inner-product*

$$\langle a, b\rangle_\rho := \int_{\mathcal{Y}} \rho(y;\xi)a(y)b(y)dy.$$

The Poisson equation

$$-\mathcal{G}_0(\xi)\Theta(y;\xi) = h(y;\xi), \quad \int_{\mathcal{Y}} \rho(y;\xi)\Theta(y;\xi)dy = 0$$

has a unique solution $\Theta(y;\xi) \in L^2_\rho(\mathcal{Y})$, provided that

$$\int_{\mathcal{Y}} \rho(y;\xi)h(y;\xi)dy = 0.$$

- *If $h(y;\xi)$ and all its derivatives with respect to y, ξ are uniformly bounded in $\mathcal{X} \times \mathcal{Y}$ then the same is true of Θ solving the Poisson equation above.*

The second assumption is an ergodicity assumption. It implies the existence of an invariant measure $\mu_\xi(dy) = \rho(y;\xi)dy$ for $\varphi_\xi^t(\cdot)$. From the Birkhoff ergodic theorem it follows that, for μ–almost all $y \in \mathcal{Y}$,

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \phi(\xi, \varphi_\xi^t(y))dt = \int_{\mathcal{Y}} \phi(\xi, y)\rho(y;x)dy.$$

The averaging and homogenization theorems we now state arise from the calculation of appropriate averages against the measure $\mu_x(dy)$.

7.3.3 *Averaging*

Starting from model (7.11), we define $F(x)$ by

$$F(x) = \int_{\mathcal{Y}} f_1(x,y)\rho(y;x)dy$$

and $\Sigma(x)$ to be the matrix satisfying

$$2\Sigma(x) = \int_{\mathcal{Y}} \big(\alpha_0(x,y)\alpha_0(x,y)^T + \alpha_1(x,y)\alpha_1(x,y)^T\big)\rho(y;x)dy.$$

We note that $\Sigma(x)$ is positive semidefinite and hence its square root is well-defined.

Theorem 7.5 *(Papavasiliou et al. (2009)) Let Assumptions 7.4 hold. Then $x \Rightarrow X$ in $C([0,T],\mathcal{X})$ where X solves the SDE*

$$\frac{dX}{dt} = F(X) + \sqrt{2\Sigma(X)}\frac{dW}{dt} \tag{7.13}$$

with W a standard $l-$dimensional Brownian motion.

7.3.4 *Homogenization*

In order for the equations (7.12) to produce a sensible limit as $\epsilon \to 0$ it is necessary to impose a condition on f_0. Specifically we assume the following which, roughly, says that $f_0(x,y)$ averages to zero against the empirical measure of the fast y process.

Assumptions 7.6

$$\int_{\mathcal{Y}} \rho(y;x)f_0(x,y)dy = 0.$$

Let $\Phi(x,y) \in L^2_\rho(\mathcal{Y})$ be the unique solution of the equation

$$-\mathcal{G}_0\Phi(y;x) = f_0(x,y), \quad \int_{\mathcal{Y}} \rho(y;x)\Phi(y;x)dy = 0.$$

This has a unique solution by Assumptions 7.4 and 7.6 (by the Fredholm Alternative, see Evans (1998) or a presentation in context in Pavliotis and Stuart (2008)). Define

$$F(x) = F_0(x) + F_1(x)$$

where

$$\begin{aligned} F_0(x) = \int_{\mathcal{Y}} \Big(& \big(\nabla_x \Phi f_0\big)(x,y) + \big(\nabla_y \Phi g_1\big)(x,y) \\ & + \big(\alpha_1 \beta^T : \nabla_y \nabla_x \Phi\big)(x,y)\Big)\rho(y;x)dy, \\ F_1(x) = \int_{\mathcal{Y}} & f_1(x,y)\rho(y;x)dy. \end{aligned}$$

Also define $\Sigma(x)$ to be the matrix satisfying

$$2\Sigma(x) = A_1(x) + A_2(x)$$

where

$$\begin{aligned} A_1(x) &= \int_{\mathcal{Y}} \Big(\big(\nabla_y \Phi \beta + \alpha_1\big)\big(\nabla_y \Phi \beta + \alpha_1\big)^T\Big)(x,y)\rho(y;x)dy, \\ A_2(x) &= \int_{\mathcal{Y}} \alpha_0(x,y)\alpha_0(x,y)^T \rho(y;x)dy. \end{aligned}$$

By construction $\Sigma(x)$ is positive semidefinite and so its square root is well-defined.

Theorem 7.7 *(Papavasiliou et al. (2009)) Let Assumptions 7.4, 7.6 hold. Then $x \Rightarrow X$ in $C([0,T],\mathcal{X})$ where X solves the SDE*

$$\frac{dX}{dt} = F(X) + \sqrt{2\Sigma(X)}\frac{dW}{dt} \tag{7.14}$$

with W a standard $l-$dimensional Brownian motion.

7.3.5 Parameter estimation

A statistical approach to multiscale data $\{x(t)\}_{t\in[0,T]}$ might consist of simply using equations of the form

$$\frac{dX}{dt} = F(X;\theta) + \sqrt{2\Sigma(X)}\frac{dW}{dt}. \tag{7.15}$$

(which is just (7.13) or (7.14) but with an unknown parameter θ and we assume $\Sigma(X)$ is uniformly positive definite on $\mathcal{X}$) to fit multiscale data that may not necessarily arise from that diffusion, so that the diffusion is a good description only at some timescales.

If we assume that the data is actually generated by the particular multiscale systems (7.11) or (7.12) we can analyze how classical maximum likelihood estimators behave in the presence of multiscale data. Naturally, this only covers one particular instance of model-misspecification due to the presence of multiscale data but it has the advantage of being amenable to rigorous analysis.

Suppose that the actual drift compatible with the data is given by $F(X) = F(X;\theta_0)$. We ask whether it is possible to correctly identify $\theta = \theta_0$ by finding the *maximum likelihood estimator* (MLE) when using a statistical model of the form (7.15), but given data from (7.11) or (7.12). We assume that (7.15) is ergodic with invariant measure $\pi(x)$ at $\theta = \theta_0$. This enables us to probe directly the question of how parameter estimators function when the desired model-fit is incompatible with the data at small scales.

Given data $\{z(t)\}_{t\in[0,T]}$, application of the Girsanov theorem shows that the log likelihood for θ satisfying (7.15) is given by

$$\mathcal{L}(\theta; z) = \int_0^T \langle F(z;\theta), dz\rangle_{\Sigma^{-1}(z)} - \frac{1}{2}\int_0^T |F(z;\theta)|^2_{\Sigma^{-1}(z)}dt, \qquad (7.16)$$

where

$$\langle r_1, r_2\rangle_{\Sigma(z)^{-1}} = \frac{1}{2}\langle r_1, \Sigma(z)^{-1} r_2\rangle.$$

The MLE is a random variable given by

$$\hat{\theta} = \text{argmax}_\theta \mathcal{L}(\theta; z).$$

Before analyzing the situation which arises when data and model are incompatible, we first recap the situation that occurs when data is taken from the model used to fit the data, in order to facilitate comparison. The following theorem shows how the log likelihood behaves, for large T, when the data is generated by the model used to fit the data itself.

Theorem 7.8 *(Papavasiliou et al. (2009)) Assume that* (7.15) *is ergodic with invariant density* $\pi(X)$ *at* $\theta = \theta_0$, *and that* $\{X(t)\}_{t\in[0,T]}$ *is a sample path of* (7.15) *with* $\theta = \theta_0$. *Then*

$$\lim_{T\to\infty} \frac{2}{T}\mathcal{L}(\theta; X) = \int_{\mathcal{Y}} |F(Z;\theta_0)|^2_{\Sigma^{-1}(Z)}\pi(Z)dZ$$
$$- \int_{\mathcal{Y}} |F(Z;\theta) - F(Z;\theta_0)|^2_{\Sigma^{-1}(Z)}\pi(Z)dZ,$$

where convergence takes place in $L^2(W)$ *(square integrable random variables on the probability space for the Brownian motion* W*) and is almost sure wrt.*

the initial condition $X(0)$. The above expression is maximized by choosing $\hat{\theta} = \theta_0$.

We make three observations: (i) for large T the likelihood is asymptotically independent of the particular sample path chosen — it depends only on the invariant measure; (ii) as a consequence we see that, asymptotically, time-ordering of the data is irrelevant to drift parameter estimation — this is something we will exploit in our non-parametric estimation in Section 7.6; (iii) the large T expression also shows that choosing data from the model which is to be fitted leads to the correct estimation of drift parameters, in the limit $T \to \infty$.

In the following we make:

Assumptions 7.9 *Equation* (7.11) *(resp.* (7.12)*) is ergodic with invariant measure $\rho^\epsilon(x, y)dxdy$. This measure converges weakly to the measure $\pi(x)\rho(y; x)$ $dxdy$ where $\rho(y; x)$ is the invariant density of the fast process* (7.9) *and $\pi(x)$ is the invariant density for* (7.13) *(resp.* (7.14)*).*

This assumption may be verified under mild assumptions on the drift and diffusion coefficients of the SDEs.

We now ask what happens when the MLE for the averaged equation (7.15) is confronted with data from the original multiscale equation (7.11). The following result explains what happens if the estimator sees the small scales of the data and shows that, in the averaging scenario, there is no problem arising from the incompatibility. Specifically the large T and small ϵ limit of the log-likelihood with multiscale data converges to the likelihood arising with data taken from the statistical model itself.

Theorem 7.10 *(Papavasiliou et al. (2009)) Let Assumptions 7.4 and 7.9 hold. Let $\{x(t)\}_{t\in[0,T]}$ be a sample path of* (7.11) *and $X(t)$ a sample path of* (7.15) *at $\theta = \theta_0$.*

$$\lim_{\epsilon\to 0}\lim_{T\to\infty}\frac{1}{T}\mathcal{L}(\theta; x) = \lim_{T\to\infty}\frac{1}{T}\mathcal{L}(\theta; X),$$

where convergence takes place in the same sense as in Theorem 7.8.

We now ask what happens when the MLE for the homogenized equation (7.15) is confronted with data from the original multiscale equation (7.12). In contrast to the situation with averaging, here there is a problem arising from the incompatibility at small scales. Specifically the large T and small ϵ limit of the log-likelihood with multiscale data differs from the likelihood arising with data taken from the statistical model at the correct parameter value.

In order to state the theorem we introduce the Poisson equation

$$-\mathcal{G}_0\Gamma = \langle F(x;\theta), f_0(x,y)\rangle_{\Sigma^{-1}(x)}, \qquad \int_{\mathcal{Y}}\rho(y;\xi)\Gamma(y;x)dy = 0 \qquad (7.17)$$

which has a unique solution $\Gamma(y;\xi) \in L^2_\rho(\mathcal{Y})$ (as for Φ, by the Fredholm Alternative). Note that

$$\Gamma = \langle F(x;\theta), \Phi(x,y)\rangle_{\Sigma^{-1}(x)}.$$

Then define

$$E = \int_{\mathcal{X}\times\mathcal{Y}} \Big(\mathcal{G}_1\Gamma(x,y) - \langle F(x;\theta), \big(\mathcal{G}_1\Phi(x,y)\big)\rangle_{\Sigma^{-1}(x)}\Big)\pi(x)\rho(y;x)dxdy.$$

Theorem 7.11 *Let Assumptions 7.4, 7.6 and 7.9 hold. Let $\{x(t)\}_{t\in[0,T]}$ be a sample path of* (7.12) *and $X(t)$ a sample path of* (7.15) *at $\theta=\theta_0$. Then*

$$\lim_{\epsilon\to 0}\lim_{T\to\infty}\frac{1}{T}\mathcal{L}(\theta;x) = \lim_{T\to\infty}\frac{1}{T}\mathcal{L}(\theta;X) + E,$$

where convergence is in the sense given in Theorem 7.8 and the order in which limits are taken is, of course, crucial.

This theorem shows that the correct limit of the log likelihood is not obtained unless $\mathcal{G}_1$ is a differential operator in y only, in which case we recover the averaging situation covered in the Theorem 7.5. The paper Papavasiliou et al. (2009) contains examples in which E can be calculated explicitly. These examples demonstrate that E is non-zero and leads to a bias. This bias indicates that fitting multiscale data to an effective homogenized model equation can lead to incorrect identification of parameters if the multiscale data is interrogated at the fastest scales. We now investigate methods designed to overcome this problem.

7.4 Subsampling

In the previous section we demonstrated that, in the situation where homogenization pertains, using classical MLE on multiscale data may result in convergent estimates of the homogenized coefficients, but the estimated homogenized coefficients can be incorrect!

In this section we illustrate the first of three ideas which can be useful in overcoming the fact that data may be incompatible with the desired diffusion at small scales. In other words, the basic idea is to use subsampling of the data, at an appropriate rate, to ensure that the data is interrogated on a scale where it "behaves like" data from the homogenized equation. This section is based on Pavliotis and Stuart (2007). Similar ideas relating to the role of subsampling are encountered in the market microstructure noise models discussed in Chapter 2.

We present results of an analysis for the special case of linear dependence of the drift on the unknown parameter θ, i.e. we assume that the vector field has the form $F(z;\theta) = \theta F(z)$ for some scalar $\theta \in \mathbb{R}$. This simplifies the results considerably, but the observation that subsampling at the correct rate leads to correct estimates of the homogenized coefficients is valid more generally; see Papavasiliou et al. (2009).

In the numerical experiments that we will present, the data will be in discrete time: it will be in the form of a sequence $z = \{z_n\}_{n=0}^N$ which we will view as approximating a diffusion process, whose parameters we wish to estimate, at time increment δ. The maximum likelihood estimator derived from (7.16) gives

$$\hat{\theta} = \frac{\int_0^T \langle F(z), dz\rangle_{\sigma(z)}}{\int_0^T |F(z)|^2_{\sigma(z)} dt}.$$

A natural discrete time analogue of this estimator, which we will use in this paper, is

$$\hat{\theta}_{N,\delta}(z) = \frac{\sum_{n=0}^{N-1} \langle F(z_n), z_{n+1} - z_n\rangle_{\sigma(z_n)}}{\sum_{n=0}^{N-1} \delta |F(z_n)|^2_{\sigma(z_n)}}. \qquad (7.18)$$

Although we concentrated in the previous section on drift parameter estimation, in numerical experiments presented here we will also investigate the estimation of the diffusion coefficient. Specifically, in the case where Σ is constant, given a discrete time-series $\{z_n\}_{n=0}^N$, we estimate Σ by

$$\hat{\Sigma}_{N,\delta}(z) = \frac{1}{2T} \sum_{j=0}^{N-1} \big(z_{j+1} - z_j\big)\big(z_{j+1} - z_j\big)^T, \qquad (7.19)$$

where $z_j = z(j\delta)$ and $N = \lfloor \frac{T}{\delta} \rfloor$. This is derived from equation (7.4).

For our numerical investigations we revisit Example 4 in one dimension. Consider the equation

$$\frac{dx}{dt} = -\alpha \nabla V^\epsilon(x) + \sqrt{2\sigma}\,\frac{dW}{dt}. \qquad (7.20)$$

Here $V^\epsilon(x) = V(x) + p(x/\epsilon)$ and p is $1-$periodic. To write this in the form to which homogenization applies notice that setting $y = x/\epsilon$ we obtain

$$\begin{aligned}\frac{dx}{dt} &= -\alpha \nabla V(x) - \frac{\alpha}{\epsilon}\nabla p(y) + \sqrt{2\sigma}\,\frac{dW}{dt},\\ \frac{dy}{dt} &= -\frac{\alpha}{\epsilon}\nabla V(x) - \frac{\alpha}{\epsilon^2}\nabla p(y) + \sqrt{\frac{2\sigma}{\epsilon^2}}\,\frac{dW}{dt}.\end{aligned}$$

This is now a specific case of (7.12).

Theorem 7.3 shows that the homogenized equation is

$$\frac{dX}{dt} = -\theta \nabla V(X) + \sqrt{2\Sigma}\,\frac{dW}{dt}. \qquad (7.21)$$

Furthermore, the theorem shows that

$$\theta < \alpha, \quad \Sigma < \sigma.$$

In fact $\theta/\Sigma = \alpha/\sigma$. It may be shown that Σ is exponentially small in $\sigma \to 0$ Campillo and Pitnitski (2002). Thus the relative discrepancy between the original and homogenized diffusion coefficients is enormous in the small diffusion limit.

The numerical experiments that we now describe concern the case where

$$V(x) = \frac{1}{2}x^2, \quad p(y) = \cos(y).$$

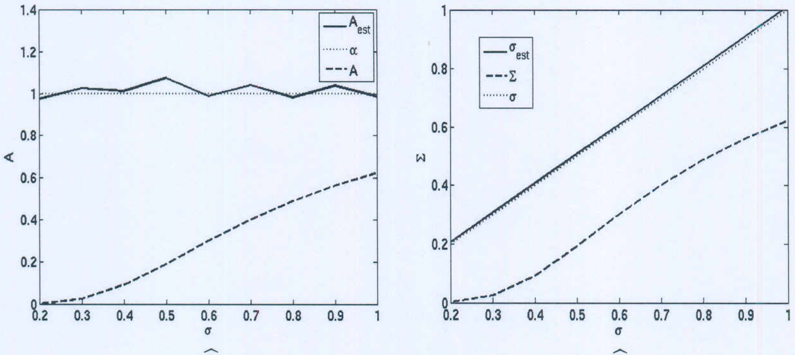
The experiments are conducted in the following way. We generate the data by simulating the multiscale process x using a time-step Δt which is small compared to ϵ^2 so that the data is a fully resolved approximation of the multiscale process. We then use this data in the estimators (7.18), (7.19) which are based on a homogenized model. We study two cases: in the first we take data sampled at time-step $\delta = \Delta t$ so that the data is high frequency relative to the small scales in the equation; we anticipate that this scenario should be close to that covered by the theory in the previous section where we take continuous time data as input. We then show what happens if we subsample the data and take a step δ which is comparable to ϵ; as the fast process has time-scale ϵ^2 the hope is that, on the scale ϵ, which is long compared with ϵ^2, the data will "look like" that of the homogenized process.

The time-interval used is $t \in [0, 10^4]$ and the data is generated with time-step $\Delta t = 5 \cdot 10^{-4}$. Figure 7.7 shows the maximum likelihood and quadratic variation estimators (7.18) and (7.19), for the drift and diffusion coefficients respectively, with data $z = x$ at the fine-scale $\delta = \Delta t$. The figure clearly shows that the estimators fail to correctly identify coefficients in the homogenized equation (7.14) for X when employing multiscale data x. Indeed the estimator finds α and σ, from the unhomogenized equation, rather than θ and Σ. Hence it overestimates.

Figure 7.8 shows that, if subsampling is used, this problem may be overcome by interrogating the data at scale ϵ; this is shown for $\epsilon = 0.1$ and using the time interval $t \in [0, 2 \cdot 10^4]$ and again generating data with a timestep of $\Delta t = 5 \cdot 10^{-4}$, by choosing $\delta = 256 \times \Delta t$, $512 \times \Delta t$ and $\delta = 1024 \times \Delta t$ and showing that, with these choices, the correct parameters are estimated for both drift and diffusion, uniformly over a wide range of σ.

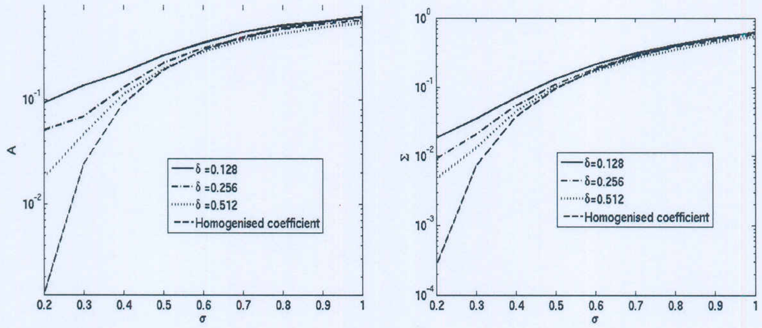
The following theorem justifies the observation that subsampling, at an appropriate rate, results in correct estimators.

Theorem 7.12 *(Papavasiliou et al. (2009)) Consider the parameter estimators* (7.18) *and* (7.19) *for drift and diffusion parameters in the statistical model* (7.21). *Define* $x = \{x(n\delta)\}_{n=0}^{N-1}$ *where* $\{x(t)\}$ *is a sample path of* (7.20).



a. $\widehat{\theta}$ b. $\widehat{\Sigma}$

Figure 7.7 *$\widehat{\theta}$ and $\widehat{\Sigma}$ vs σ for $\alpha = 1$, $\epsilon = 0.1$*



a. $\widehat{\theta}$ b. $\widehat{\Sigma}$

Figure 7.8 *$\widehat{\theta}$ and $\widehat{\Sigma}$ vs σ for $a = 1$, $\epsilon = 0.1$. $\delta \in \{0.128, 0.256, 0.512\}$*

- *Let $\delta = \epsilon^\alpha$ with $\alpha \in (0,1)$, $N = \lfloor \epsilon^{-\gamma} \rfloor$, $\gamma > \alpha$. Then*

$$\lim_{\epsilon \to 0} \hat{\theta}_{N,\delta}(x) = \theta \quad \textit{in distribution.}$$

- *Fix $T = N\delta$ with $\delta = \epsilon^\alpha$ with $\alpha \in (0,1)$. Then*

$$\lim_{\epsilon \to 0} \widehat{\Sigma}_{N,\delta}(x) = \Sigma \quad \textit{in distribution.}$$

7.5 Hypoelliptic diffusions

In this section we return to the Butane molecule considered in Section 7.2.3. In that example we showed problems arising from trying to fit a scalar diffusion to the raw time series data from the dihedral angle. Amongst several problems with the attempted fit, we highlighted the fact that the data came from a time series with zero quadratic variation, derived as a nonlinear function of the time-series $x(t)$ from (7.5), whilst the equation (7.6) which we attempted to fit had non-zero quadratic variation. In this section we show how we may attempt to overcome this problem by fitting a *hypoelliptic diffusion process* to the dihedral angle data. Technical details can be found in the paper Pokern, Stuart, and Wiberg (2009).

Specifically we attempt to fit to the data $\{\phi_n\}_{n=0}^N$, where observations are made at small but fixed inter-sample time δ so that $\phi_n = \phi(n\delta)$, a model of the form

$$\frac{d^2q}{dt^2} + \gamma\frac{dq}{dt} - V'(q) = \sigma\frac{dW}{dt}, \tag{7.22a}$$

$$V(q) = \sum_{j=1}^{5} \frac{\theta_j}{j}\Big(\cos(q)\Big)^j, \tag{7.22b}$$

q here plays the same role as Φ in (7.6): it describes the dihedral angle in a postulated lower dimensional stochastic fit to data derived from a higher dimensional dynamical model. The statistical task at hand is then to use the data ϕ_n to infer the value of the parameters $\{\theta_j\}_{j=1}^5$ as well as γ and σ.

This task is problematic because

1. observations of only ϕ rather than ϕ and its time derivative $\frac{d\phi}{dt}$ are assumed to be available, so a missing data problem is to be dealt with;
2. the two components of the process, namely ϕ and $\frac{d\phi}{dt}$ have different smoothness rendering straightforward statistical models ill-conditioned.

Velocities for ϕ may be available in practice (molecular dynamics codes can certainly produce such output if desired) but it may be preferable to ignore them — such data may be incompatible with SDE models.

It should be emphasized that both the missing data aspect (1. above) and the different degrees of differentiability (2. above) are serious problems *regardless* of whether a frequentist or a Bayesian approach is developed. For simplicity, we will explain these problems further using the maximum likelihood principle and only later introduce a full (Bayesian) estimation algorithm.

In order to better understand these problems we rewrite (7.22) as a damped-driven Hamiltonian system as follows:

$$dq = pdt, \tag{7.23a}$$

$$dp = \left(-\gamma p - \sum_{j=1}^{5} \theta_j \sin(q) \cos^{j-1}(q) \right) dt + \sigma dB, \tag{7.23b}$$

where we have used the new variable $p(t) = \frac{dq}{dt}$. To understand how the two problems 1. and 2. enumerated above come about, consider one of the most widespread discrete time approximations to this SDE, the Euler-Maruyama approximation:

$$Q_{i+1} = Q_i + \delta P_i \tag{7.24a}$$

$$P_{i+1} = P_i - \left(\gamma P_i + \sum_{j=1}^{5} \theta_j \sin(Q_i) \cos^{j-1}(Q_i) \right) \delta + \sigma \sqrt{\delta} \xi_i \tag{7.24b}$$

where $\xi_i \sim \mathcal{N}(0,1)$ is a sequence of iid. standard normal random variables and we have used capital variable names to indicate that this is the discretised version of a continuous time system. It is possible to estimate all desired parameters using this model, by first using (7.24a) to obtain $P_i = \frac{1}{\delta}(Q_{i+1} - Q_i)$ and then estimating σ from the quadratic variation of the path $\{P_i\}_{i=0}^{N-1}$ and the drift parameters can then be estimated by applying the maximum likelihood principle to (7.24b).

Using this approximation to estimate σ given the data $\{\phi_n\}_{n=0}^{N}$ for $\{Q_i\}_{i=0}^{N}$ and no data for $\{P_i\}_{i=0}^{N}$ leads to gross mis-estimation. In fact for the simpler example of stochastic growth

$$\begin{aligned} dq &= pdt \\ dp &= \sigma dB \end{aligned}$$

it is straightforward to show that in the limit of infinitely many observations $N \to \infty$ (both in the case when δ is fixed and in the case when $T = N\delta$ is fixed!) the maximum likelihood estimator $\hat{\sigma}$ converges to an incorrect estimate almost surely:

$$\hat{\sigma}^2 \longrightarrow \frac{2}{3}\sigma^2 \quad \text{a.s.}$$

This failure can be traced back to the fact that (7.24) effectively uses numerical differentiation of the time series Q_i to solve the missing data problem, i.e. to estimate P_i. This approximation neglects noise contributions of order $\mathcal{O}(\delta^{\frac{3}{2}})$ in (7.24a) which are of the same order as the contributions obtained via numerical differentiation of Q_i.

Understanding the source of the error suggests that replacing the Euler-Muruyama scheme with a higher order discretization scheme that propagates noise to both rows of the equation (7.23) results in successful estimators for σ. One such scheme is given by

$$\begin{bmatrix} Q_{i+1} \\ P_{i+1} \end{bmatrix} = \begin{bmatrix} Q_i \\ P_i \end{bmatrix} + \delta \begin{bmatrix} P_i \\ \sum_{j=1}^{5} \theta_j \sin(Q_i) \cos^{j-1}(Q_i)) - \gamma P_i \end{bmatrix} + \sigma\sqrt{\delta} R \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \tag{7.25}$$

where the matrix R is given as

$$R = \begin{bmatrix} \frac{\delta}{\sqrt{12}} & \frac{\delta}{2} \\ 0 & 1 \end{bmatrix} \tag{7.26}$$

and ξ_1 and ξ_2 are again independent standard normal random variables. Generally, Itô-Taylor expansions of sufficiently high order should be used to propagate noise to all components of the process.

The approximation (7.25) can be used not only to infer σ but also to infer the missing component of the path.

Finally, it remains to estimate the drift parameters $\{\theta_j\}_{j=1}^{5}$ and γ and it turns out that the approximation (7.25) yields results with a large bias that does not decay as δ decreases or the observation time T increases. In fact, for the simpler case of a harmonic oscillator

$$\left\{ \begin{array}{lcl} dq & = & pdt \\ dp & = & -\theta q dt - \gamma p dt + \sigma dB. \end{array} \right. \tag{7.27}$$

it is possible to compute by Itô-Taylor-expansion that the maximum likelihood estimator for θ and γ based on an analogous model to (7.25) satisfies:

$$\mathbb{E}\hat{\theta} = \frac{1}{4}\theta + \mathcal{O}(\delta)$$
$$\mathbb{E}\hat{\gamma} = \frac{1}{4}\gamma + \mathcal{O}(\delta).$$

This can be traced back to the fact that such an estimator assumes the drift parameters in the first row of (7.27) to be known exactly whereas the discrete time path only satisfies (7.25) (or the analogous model for the harmonic oscillator) approximately. The ill-conditioning of the inverse of the matrix R introduced in (7.26) as $\delta \to 0$ causes small errors to be amplified to $\mathcal{O}(1)$ deviations in the drift parameter estimates. Using the Euler-Maruyama approximation (7.24) instead delivers satisfactory results.

Having used a maximum likelihood framework to highlight both the fact that the missing data problem adds significant difficulty and that the different degrees of differentiability of ϕ and $\frac{d\phi}{dt}$ produce ill-conditioning, we now proceed to a Bayesian algorithm to infer the missing data $\{P_j\}_{j=0}^{N}$ as well as the diffusion and drift parameters σ and $\{\theta_j\}_{j=1}^{5}$ and γ. The sequential Gibbs algorithm suggested to produce approximate samples $\gamma^{(i)}$, $\{\theta_j^{(i)}\}_{j=1}^{5}$, $\sigma^{(i)}$ and $P_n^{(i)}$ indexed by $i = 1, 2, \ldots$ thus reads as follows:

1. Sample $\theta^{(i+1)}, \gamma^{(i+1)}$ from $\mathbb{P}(\{\theta_j\}_{j=1}^5, \gamma | \{Q_j\}_{j=0}^N, \{P_j^{(i)}\}_{j=0}^N, \sigma^{(i)})$ using (7.24).
2. Sample $\sigma^{(i+1)}$ from $\mathbb{P}(\sigma | \{Q_j\}_{j=0}^N, \{P_j^{(i)}\}_{j=0}^N, \{\theta_j^{(i+1)}\}_{j=1}^5, \gamma^{(i+1)})$ using (7.25).
3. Sample $\{P_j^{(i+1)}\}_{j=0}^N$ from $\mathbb{P}(\{P_j\}_{j=0}^N | \{Q_j\}_{j=0}^N, \{\theta_j^{(i+1)}\}_{j=1}^5, \gamma^{(i+1)}, \sigma^{(i+1)})$ using (7.25).

Note that we have omitted the initialization stage and that sampling the missing path, stage 3, is simplified by the fact that p only ever enters the SDE linearly, so that a direct Gaussian sampler can be used. Stage 1 is also Gaussian, whereas stage 2 is not, and an MCMC method, for example, can be used.

The status of Gibbs samplers, such as this one, combining different approximate likelihoods, especially in the presence of ill-conditioning which renders some approximations unsuitable for some of the estimation tasks, is not yet theoretically understood. In the particular case of the SDE being fitted, the method has been subjected to very careful numerical studies detailed in Pokern, Stuart, and Wiberg (2009) and found to be convergent.

Finally, the method can be applied to the data given as Example 3 in Section 7.2.3 and Figure 7.9 shows posterior mean parameter estimates as a function of sampling interval δ.

We have shown in this section how to extend parameter estimation from the elliptic case (7.6) to the hypoelliptic case (7.22) and we have highlighted how to do this in the case of missing velocities. This is useful e.g. when neglecting the velocities is viewed as a means to decrease the fitting process' sensitivity to incompatibility between the model and the data at the short timescales. Still, Figure 7.9 shows that the fitted drift and diffusion parameters are far from independent of the timescale on which we look at the data. It is natural to ask why this is so. Since the dihedral angle is a nonlinear transformation of the Cartesian coordinates in (7.5) and hence (as a brief application of the Itô formula readily shows) will exhibit multiplicative rather than additive noise, it will *not* be well-described by a hypoelliptic diffusion with constant diffusivity. It is this problem with model fit that results in the timescale dependence of fitted parameters evidenced in Figure 7.9.

7.6 Non-parametric drift estimation

Theorem 7.8 illustrates the fact that, for large times, drift parameter estimation does not see path properties of the data, but rather just sees the invariant measure. This suggests an approach to drift parameter estimation which exploits this property directly and uses only the empirical measure of the data, thereby

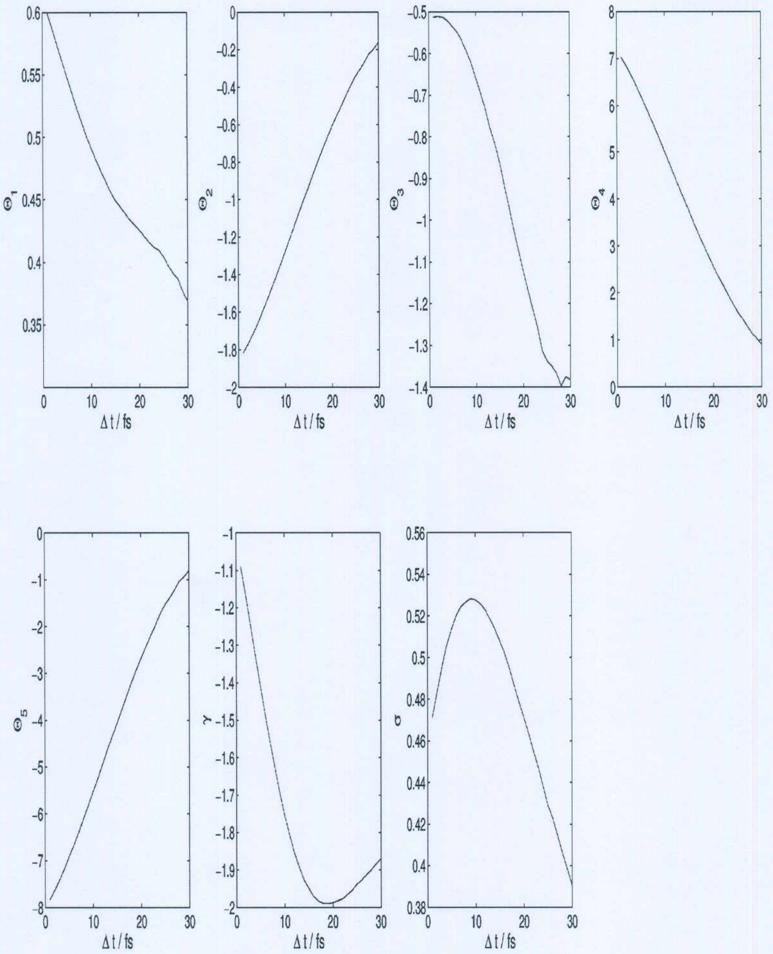


Figure 7.9 *Mean posterior estimates for drift and diffusion parameters using Butane data as a function of sampling interval* δ

avoiding issues relating to incompatibility at small scales. These ideas are pursued in Pokern, Stuart, and Vanden-Eijnden (2009) and then, building on this in a Bayesian context, in Papaspiliopoulos et al. (2009).

Here we illustrate the basic ideas in the context of the one-dimensional equation

$$\frac{dx}{dt} = -V'(x) + \frac{dW}{dt}. \tag{7.28}$$

In Pokern, Stuart, and Vanden-Eijnden (2009) we treat the multidimensional case, matrix diffusion coefficients and the second order Langevin equation. We note for later use that the SDE (7.28) has invariant measure with density

$$\rho \propto \exp\left(-2V\right), \tag{7.29}$$

provided $\exp\left(-2V\right) \in L^1(\mathbb{R})$.

Our strategy is a non-parametric one. We write down the log likelihood for this equation which, from (7.16), has the form

$$\mathcal{L}(V;x) = -\int_0^T V'(x)dx - \frac{1}{2}\int_0^T |V'(x)|^2 dt. \tag{7.30}$$

Notice that now we view the log likelihood as being a functional of the unknown drift with potential V. This reflects our non-parametric stance. Applying the Itô formula to $V(x(t))$ we deduce that

$$-\int_0^T V'(x)dx = \frac{1}{2}\int_0^T V''(x)dt + \Big(V(x(0)) - V(x(T))\Big).$$

Thus we obtain

$$\mathcal{L}(V;x) = V(x(0)) - V(x(T)) + \frac{1}{2}\int_0^T \left(V''(x) - |V'(x)|^2\right) dt. \tag{7.31}$$

With the goal of expressing the likelihood independently of small-timescale structure of the data we now express the log likelihood in terms of the local time L_T^a of the process. Recall that the local time L_T^a measures the time spent at a up to time T, so L_T^a/T is proportional to an empirical density function for the path history up to time T. Note that use of the local time L_T^a of the process indeed removes any time-ordering in the data and thus makes drift estimation independent of the dynamical information contained in the data. Since time-ordered data at small scales is at the root of the problems we are confronting in this paper, taking this point of view is likely to be beneficial.

To make the notion of local time as a scaled empirical density rigorous, consider Theorem 2.11.7 in Durrett (1996) which states that for x being a 1-d continuous semimartingale with local time L_T^a and quadratic variation process $\langle x\rangle_t$, the following identity holds for any Borel-measurable, bounded function g:

$$\int_{-\infty}^{\infty} L_T^a g(a)da \quad = \quad \int_0^T g(x_s)d\langle x\rangle_s. \tag{7.32}$$

Note that for the process (7.28) we have

$$dt \quad = \quad d\langle x\rangle_t$$

so that

$$\int_{-\infty}^{\infty} L_t^a g(a)da = \int_0^t g(x_s)ds.$$

In terms of the local time we have

$$\mathcal{L}(V;x) = -\left(V(x(T)) - V(x(0))\right) + \frac{1}{2}\int_{\mathbb{R}} \left(V''(a) - |V'(a)|^2\right) L_T^a da. \tag{7.33}$$

To make the mathematical structure of this functional more apparent, we re-express the likelihood in terms of the drift function $b = -V'$. To do this, we first introduce the signed indicator function

$$\tilde{\chi}(a; X_0, X_T) = \begin{cases} 1 & \text{if } X_0 < a < X_T \\ -1 & \text{if } X_T < a < X_0 \\ 0 & \text{otherwise} \end{cases}.$$

The expression for the likelihood then takes the following form:

$$\mathcal{L}(V;x) = -\frac{1}{2}\int_{\mathbb{R}} \left(\left(|V'(a)|^2 - V''(a)\right) L_T^a + 2\tilde{\chi}(a; X_0, X_T)V'(a)\right) da.$$

Having eliminated V by expressing everything in terms of its derivatives, we replace those derivatives by the drift function b as planned. We abuse notation and now write $\mathcal{L}$ as a functional of b instead of V:

$$\mathcal{L}(b;x) = -\frac{1}{2}\int_{\mathbb{R}} \left(b^2(a)L_T^a + b'(a)L_T^a - 2\tilde{\chi}(a; X_0, X_T)b(a)\right) da. \tag{7.34}$$

We would like to apply the likelihood principle to $\mathcal{L}(b;x)$ to estimate b. **Purely formally**, seeking a critical point of the functional (7.34) is possible, i.e. one asks that its functional derivative with respect to b be zero:

$$\frac{\delta \mathcal{L}(b;x)}{\delta b} \overset{!}{=} 0$$

To carry this out, integrate by parts to obtain

$$\mathcal{L}(b;x) = -\frac{1}{2}\int_{\mathbb{R}} \left(b^2(a)L_T^a - b(a)L_T'(a) - 2\tilde{\chi}(a; X_0, X_T)b(a)\right) da.$$

Expand this expression at $b(\epsilon) = b + \epsilon u$ where u is an arbitrary smooth function to compute the functional derivative:

$$\frac{\delta \mathcal{L}(b;x)}{\delta b} u = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\left(\mathcal{L}(b + \epsilon u; x) - \mathcal{L}(b;x)\right)$$

Equating the functional derivative to zero yields the **formal** maximum likelihood estimate

$$\hat{b} = \frac{L_T'}{2L_T} - \frac{\tilde{\chi}(\cdot; X_0, X_T)}{L_T}. \tag{7.35}$$

Note the derivative of the local time figures prominently in this estimate – however, it is not defined since $L_T^{\cdot}$ is not differentiable. It can be shown that, in one dimension, the local time L_t^a is jointly continuous in (t, a), but that it is not in general differentiable; it is only α-Hölder continuous up to but excluding exponent $\alpha = 1/2$. Therefore, the likelihood functional in (7.34) would not be expected to be bounded above and the estimate (7.35) is *not* a proper maximiser of the likelihood.

To get an idea of why this is, consider the case where local time is replaced by a Brownian bridge (which has essentially the same Hölder regularity as local time) where it is possible to show the following theorem (see Appendix 1):

Theorem 7.13 *Let $w \in C([0,1], \mathbb{R})$ be a realisation of the standard Brownian bridge on $[0,1]$. Then with probability one, the functional*

$$I(b; w) = -\frac{1}{2}\int_0^1 \left(b^2(s)w(s) + b'(s)w(s)\right) ds$$

is not bounded above for $b \in H^1([0,1])$.

If $\mathcal{L}(b; x)$ given by (7.34) is not bounded above, then application of the maximum likelihood principle will fail. To remedy this problem, several options are available. One can introduce a parametrization $b(x, \theta)$ for $\theta \in \Theta \subset \mathbb{R}^m$ for some finite m with the attendant problems of choosing a set of basis functions that make the parameters well-conditioned and easy to interpret. Alternatively, one can work with a mollified version of the local time, $\tilde{L}_T^a$, which is smooth enough to ensure existence of a maximizer of the likelihood functional. Finally, it is possible to use Tikhonov regularization and then, taking this further, to adopt a Bayesian framework and use a prior to ensure sufficient regularity. In Section 7.6.1, we will investigate the use of mollified local time in detail and in Section 7.6.2 we briefly introduce the Bayesian non-parametric approach.

7.6.1 Mollified local time

To start using mollified local time, we proceed in three steps adopting a traditional regularization and truncation approach. Firstly, in (7.34) we replace L_T^a by a mollification $\tilde{L}_T^a$ which is assumed to be compactly supported and non-negative just like the original local time. Additionally, we assume that it has Sobolev regularity $L_T^{\cdot} \in H^1(\mathbb{R})$. Secondly, we integrate by parts exploiting the smoothness and compact support of $\tilde{L}_T^a$:

$$\mathcal{L}(b; x) \approx -\frac{1}{2}\int_{\mathbb{R}} \left(b^2(a)\tilde{L}_T^a - b(a)\left(\tilde{L}_T^a\right)' - 2\tilde{\chi}(a; X_0, X_T)b(a)\right) da \tag{7.36}$$

Thirdly, we restrict attention to a bounded open interval $U \subset \mathbb{R}$ on the real line which is chosen such that

$$\exists \epsilon > 0\, \forall a \in U : \tilde{L}_T^a > \epsilon. \tag{7.37}$$

This leads to the final approximation of $\mathcal{L}(b;x)$ by the following functional:

$$\tilde{\mathcal{L}}(b;x) = -\frac{1}{2}\int_U \left(b^2(a)\tilde{L}_T^a - b(a)\left(\tilde{L}_T^a\right)' - 2\tilde{\chi}(a;X_0,X_T)b(a)\right) da \tag{7.38}$$

This functional is quadratic in b and it is straightforward to prove the following theorem:

Theorem 7.14 *The functional $\tilde{\mathcal{L}}(b;x)$ is almost surely bounded above on $b \in L^2(U)$ and its maximum is attained at*

$$\hat{b} = \frac{1}{2}\left(\log \tilde{L}_T^a\right)' - \frac{\tilde{\chi}(a;X_0,X_T)}{\tilde{L}_T^a} \tag{7.39}$$

Proof. We rewrite the mollified likelihood functional by completing the square as follows:

$$\tilde{\mathcal{L}}(b;x) = -\frac{1}{2}\int_U \left[\left(b - \frac{\left(\tilde{L}_T^a\right)'}{2\tilde{L}_T^a} - \frac{\tilde{\chi}(a;X_0,X_T)}{\tilde{L}_T^a} \right)^2 \tilde{L}_T^a \right. \tag{7.40}$$

$$\left. - \left(\frac{\left(\tilde{L}_T^a\right)'}{2\tilde{L}_T^a} + \frac{\tilde{\chi}(a;X_0,X_T)}{\tilde{L}_T^a} \right)^2 \tilde{L}_T^a \right] da \tag{7.41}$$

Observe that the first summand in the integrand is always non-negative. To avoid potentially subtracting two infinite terms from each other, we verify that the second summand in the integrand has a finite integral:

$$\begin{aligned}
&\int_U \left(\frac{\left(\tilde{L}_T^a\right)'}{2\tilde{L}_T^a} + \frac{\tilde{\chi}(a;X_0,X_T)}{\tilde{L}_T^a} \right)^2 \tilde{L}_T^a da \\
&\qquad = \int_U \frac{1}{\tilde{L}_T^a} \left(\frac{\left(\tilde{L}_T^a\right)'}{2} + \tilde{\chi}(a;X_0,X_T) \right)^2 da \\
&\qquad \le \frac{1}{\epsilon} \int_U \left(\frac{1}{2}\left(\tilde{L}_T^a\right)' + \tilde{\chi}(a;X_0,X_T) \right)^2 da < \infty
\end{aligned}$$

where we have used condition (7.37) in the penultimate inequality and the fact that U is compact and that $\tilde{L}^{\cdot}_T$ is smooth (and hence it and its derivative are square-integrable on U). All that remains is to read off the maximizer (7.39) from the brackets in (7.40). □

The last term in (7.39) is integrable on U (since $\tilde{L}$ is bounded away from zero on U and U is bounded), the integrated version of that MLE thus reads

$$\widehat{V}(a) = -\frac{1}{2}\log \tilde{L}^a_T + \int_{\inf(U)}^{a} \frac{\tilde{\chi}(s; X_0, X_T)}{\tilde{L}^s_T} ds, \quad a \in U.$$

Furthermore, we expect that $\tilde{L}^a_T$ scales like $\mathcal{O}(T)$. Thus the first term gives the dominant term in the estimator for large T. Retaining only the first term gives the approximation

$$\widehat{V}(a) = -\frac{1}{2}\log \tilde{L}^a_T.$$

If we make the reasonable assumption that $\frac{1}{T}\tilde{L}^a_T \to \rho(a)$ as $T \to \infty$ where ρ is the invariant measure for the process we deduce that, for this approximation,

$$\lim_{T\to\infty} \widehat{V}(a) = -\frac{1}{2}\log \rho(a).$$

But the invariant density is given by (7.29) and so we deduce that, under these reasonable assumptions,

$$\lim_{T\to\infty} \widehat{V}(a) = V(a)$$

as expected.

7.6.2 Regularized likelihood functional

Another way to regularize the functional (7.33) is to add a penalty function to the logarithm of the likelihood to obtain

$$\mathcal{L}_p(b; x) = \mathcal{L}(b; x) - \|b - b_0\|^2_H .$$

Here H is a suitable Hilbert subspace and we call b_0 the centre of regularization. We refer to this procedure as Tikhonov regularization; in the case $b_0 = 0$ it coincides with the standard usage – see Kaipio and Somersalo (2005). If the additional term $\|b - b_0\|^2_H$ is chosen to penalize roughness it is possible to ensure that the combined logarithm has a unique maximum. We will outline that in the sequel that since the penalization is quadratic, this may be linked to the introduction of a Gaussian prior. The posterior will be seen to be Gaussian, too, because the likelihood is quadratic in b, and all densities and probabilities arising can be given a fully rigorous interpretation. We leave technical and implementation details to Papaspiliopoulos, Pokern, Roberts, and Stuart (2011) and Papaspiliopoulos et al. (2009) and merely outline the key calculations first concentrating on the viewpoint of the Tikhonov regularization $\mathcal{L}_p$ of the likelihood functional $\mathcal{L}$, then introducing the Bayesian viewpoint at the end.

Tikhonov regularization

We first state a theorem to show that regularization using the Hilbert space norm

$$\|b\|_H^2 = 2 \int_{\mathbb{R}} b(a)^2 + (b'(a))^2 + (b''(a))^2 \, da$$

on the Sobolev space $H = H^2(\mathbb{R})$ is indeed possible:

Theorem 7.15 *For any fixed $c > 0$, the functional*

$$\mathcal{L}_p(b; x) = \mathcal{L}(b; x) - c\|b - b_0\|_{H^2(\mathbb{R})}^2$$

is almost surely bounded above on $b \in H^2(\mathbb{R})$ for any $b_0 \in H^2(\mathbb{R})$.

Proof. The proof follows along the same lines as the proof of Theorem 7.17 which will be given in full. □

Showing that a maximizer exists and that it is the solution of the accompanying Euler-Lagrange equations requires a more detailed analysis which is easier to carry out in the periodic setting. Thus, we henceforth consider Itô SDEs with constant diffusivity on the circle parametrized by $[0, 2\pi]$,

$$dx = b(x)dt + dW, \quad x(0) = x_0 \quad \text{on } [0, 2\pi].$$

The state space of the diffusion process is now compact. Also note that using the circle as a state space introduces another linear term into $\mathcal{L}(b; x)$ so that we now have

$$\mathcal{L}(b; x) = -\frac{1}{2} \int_0^{2\pi} \Big(b^2(a) L_T^a + b'(a) L_T^a - 2(M + \tilde{\chi}(a; X_0, X_T)) b(a) \Big) da, \tag{7.42}$$

where $M \in \mathbb{Z}$ is the winding number of the process x, i.e. the number of times x has gone around the circle in $[0, T]$.

Let us now consider the Tikhonov regularization of $\mathcal{L}$ by the H^2-seminorm as follows:

$$\mathcal{L}_p(b; x) = \mathcal{L}(b; x) - \frac{1}{2} |b - b_0|_{H^2}^2 \qquad b \in H^2_{\text{per}}([0, 2\pi]), \tag{7.43}$$

where $H^2_{\text{per}}([0, 2\pi])$ refers to the Sobolev space of twice weakly differentiable periodic functions on $[0, 2\pi]$ and $|b|_{H^2} = \int_0^{2\pi} (b''(a))^2 \, da$. Note that the prefactor $\frac{1}{2}$ in front of the seminorm can be replaced by an arbitrary positive constant, allowing an adjustment of the strength of regularization. To analyze this regularized functional, we separate off its quadratic terms by introducing the bilinear form $q(u,v)$ defined for $u, v \in H^2_{\text{per}}([0,2\pi])$ as follows:

$$q(u,v) = \frac{1}{2}\int_0^{2\pi} \Delta u(a)\Delta v(a) + u(a)L_T^a v(a)da, \tag{7.44}$$

where we denote second derivatives by the Laplace operator, $\Delta = \frac{d^2}{da^2}$. Important properties of this bilinear form are given in the following Lemma whose proof is slightly technical and can be found in Papaspiliopoulos et al. (2009).

Lemma 7.16 *If the local time L_T is not identically zero on $[0,2\pi]$, then the form q, defined in (7.44), is a continuous, coercive, symmetric bilinear form, i.e. there are constants $\alpha, C \in \mathbb{R}_+$ which may depend on L_T but not on u, v such that the following relations hold:*

$$\begin{aligned} \alpha\|u\|^2_{H^2} \le q(u,u) \quad & \forall u \in H^2_{\text{per}}([0,2\pi]) \qquad (7.45)\\ q(u,v) \le C\|u\|_{H^2}\|v\|_{H^2} \quad & \forall u,v \in H^2_{\text{per}}([0,2\pi])\\ q(u,v) = q(v,u) \quad & \forall u,v \in H^2_{\text{per}}([0,2\pi]) \end{aligned}$$

We now state an analogous theorem to Theorem 7.14:

Theorem 7.17 *The functional $\mathcal{L}_p(b)$ defined in (7.43) is almost surely bounded above on $b \in H^2_{\text{per}}([0,2\pi])$ and its maximum is attained at $\widehat{b} \in H^2_{\text{per}}([0,2\pi])$ which is given by the unique weak solution of the boundary value problem*

$$\left(\Delta^2 + L_T\right)\widehat{b} = \frac{1}{2}L_T' + M + \tilde{\chi}(\cdot; X_0, X_T) + \Delta^2 b_0. \tag{7.46}$$

Proof. We present a heuristic calculation first that simply proceeds by completing the square and reading off the answer. We then indicate how this can be approached rigorously. To simplify notation we assume that the centre of regularization is identically zero: $b_0 = 0$.

$$\mathcal{L}_p(b;x) = -\frac{1}{2}\int_0^{2\pi}\Bigg(b^2(a)L_T^a + (\Delta b(a))^2 + b'(a)L_T^a \\ -2\left(\tilde{\chi}(a;X_0,X_T) + M\right)b(a)\Bigg)da$$

To simplify the notation we drop the arguments. To formally derive the maximizer we repeatedly integrate by parts, pretending that the local time is sufficiently regular.

$$\mathcal{L}_p = -\frac{1}{2}\int\Big(b(\Delta^2 + L_T)b - b(L_T' + 2(\tilde{\chi} + M)\Big)da$$

We now introduce the abbreviations

$$\mathcal{D} = \Delta^2 + L_T \tag{7.47}$$

$$c = -\frac{1}{2}L_T' - (\tilde{\chi} + M) \tag{7.48}$$

and the following notational convention for the square root of the operator $\mathcal{D}$ and its inverse:

$$\begin{aligned} \left|\mathcal{D}^{\frac{1}{2}}u\right|^2 &= \langle u, \mathcal{D}u\rangle \qquad u \in H^2_{\mathrm{per}}([0, 2\pi]) \\ \left|\mathcal{D}^{-\frac{1}{2}}u\right|^2 &= \langle u, \mathcal{D}^{-1}u\rangle \qquad u \in H^2_{\mathrm{per}}([0, 2\pi]), \end{aligned}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product on the Hilbert space $L^2([0, 2\pi])$. These conventions are intended to make the following calculation more transparent, so we rewrite $\mathcal{L}_p$ first of all:

$$\mathcal{L}_p(b; x) = -\frac{1}{2}\langle b, \mathcal{D}b\rangle - \langle b, c\rangle$$

Now we complete the square

$$\mathcal{L}_p(b; x) = -\frac{1}{2}\left|\mathcal{D}^{\frac{1}{2}}\left(b + \mathcal{D}^{-1}c\right)\right|^2 + \frac{1}{2}\left|\mathcal{D}^{-\frac{1}{2}}c\right|^2$$

and note that the maximizer of the regularized functional can be seen to be given by

$$\widehat{b} = -\mathcal{D}^{-1}c.$$

This is identical to (7.46) which can be seen by inserting the terms from (7.47) and (7.48); boundedness from above is also apparent.

We rigorously establish boundedness from above (again in the case $b_0 = 0$) and leave the rest of the proofs to Papaspiliopoulos et al. (2009). To do this, we first rewrite the Tikhonov-regularized likelihood using the quadratic form (7.44) as follows:

$$\mathcal{L}_p(b; x) = \frac{1}{2}\int_0^{2\pi} \left(-b'(a)L_T^a + 2\left(M + \tilde{\chi}(a; X_0, X_T)\right)b(a)\right)da - q(b, b)$$

Now bound the linear term in $\mathcal{L}_p$ as follows:

$$\begin{aligned} &\left|\int_0^{2\pi} b'(a)L_T^a - 2b(a)(M + \tilde{\chi}(a; X_0, X_T))da\right| \\ \leq &\int_0^{2\pi} \epsilon_1\left(b'(a)\right)^2 + \frac{1}{\epsilon_1}\left(L_T^a\right)^2 + \epsilon_2 b^2(a) + \frac{1}{\epsilon_2}\left(M + \tilde{\chi}(a; X_0, X_T)\right)^2 da \end{aligned}$$

which holds for any $\epsilon_1, \epsilon_2 \in (0, \infty)$. Now choose $\epsilon = \epsilon_1 = \epsilon_2 < \alpha$ where

α is given by (7.45) from Lemma 7.16. Exploit coercivity (i.e. (7.45)) in the following way:

$$\begin{aligned}\mathcal{L}_p(b) \leq & -q(b,b) + \epsilon\|b\|_{H^1}^2 + \frac{1}{\epsilon}\|\dot{L}_T\|_{L^2}^2 + \frac{1}{\epsilon}\|M + \tilde{\chi}(\cdot; X_0, X_T)\|_{L^2}^2 \\ \leq & -(\alpha - \epsilon)\|b\|_{H^2}^2 + \text{const}.\end{aligned}$$

The case $b_0 \neq 0$ presents mainly notational complications, whereas deriving the PDE (7.46) requires a little variational calculus and showing existence and uniqueness of its solutions is an application of standard PDE theory given in Papaspiliopoulos et al. (2009). □

Bayesian viewpoint

In this subsection we show that the Tikhonov regularization given above underpins the adoption of a Bayesian framework and we briefly outline some speculation concerning the limit $T \to \infty$.

To introduce a prior, we decompose the space $H^2_{\text{per}}([0, 2\pi])$, into the direct sum of the (one-dimensional) space of constant functions and the space of Sobolev functions with average zero (denoted by $\dot{H}^2_{\text{per}}([0, 2\pi])$:

$$H^2_{\text{per}}([0, 2\pi]) = \{\alpha\mathbf{1} | \alpha \in \mathbb{R}\} \bigoplus \dot{H}^2_{\text{per}}([0, 2\pi]),$$

where we use $\mathbf{1}$ to denote the constant function with value one. We now define the prior measure as the product measure found from Lebesgue measure on the space of constant functions and the Gaussian measure $\mathcal{N}(b_0, A)$ on the space $\dot{H}^2_{\text{per}}([0, 2\pi])$ where $A = \left(-\frac{d^2}{da^2}\right)^{-2}$ subject to periodic boundary conditions:

$$p_0 = \lambda \otimes \mathcal{N}(b_0, A).$$

This is the prior measure. Note that as this is a degenerate Gaussian, we are using an improper prior. Purely formally, this prior can be written as a density with respect to (non-existing) Lebesgue measure on $H^2_{\text{per}}([0, 2\pi])$:

$$p_0(b) \sim \exp\left(-\frac{1}{2}\int_0^{2\pi} (b - b_0)(a)(\Delta^2(b - b_0))(a)da\right). \qquad (7.49)$$

Having defined a prior we now use the Radon Nikodym derivative $\mathcal{L}(b; x)$ as the likelihood just as before. The posterior measure then follows the Bayes formula:

$$\begin{aligned} \mathbb{P}\left(b|\{x_t\}_{t=0}^T\right) &\propto \mathbb{P}(\{x_t\}_{t=0}^T|b)p_0(b) \\ = \exp\Big(-\frac{1}{2}\int_0^{2\pi} & L_T^a b^2(a) - b(a)\,(L_T^a)' - 2b(a)\,(M + \tilde{\chi}(a; X_0, X_T))\, da \\ & -\frac{1}{2}\int_0^{2\pi} (b(a) - b_0(a))\Delta^2(b(a) - b_0(a))da\Big) \end{aligned}$$

where M again corresponds to the number of times the path $\{x_s\}_{s=0}^T$ winds around the circle. Note that this is just the straightforward product of the likelihood (7.42) with the prior measure (7.49). We simplify this expression by considering the case $b_0 = 0$ and by dropping the arguments:

$$\mathbb{P}\left(b|\{x_t\}_{t=0}^T\right) \propto \exp\left(-\frac{1}{2}\int_0^{2\pi} |\Delta b|^2 + L_T b^2 - b\,(L_T' + 2M + 2\tilde{\chi})\, da\right).$$

Formally completing the square in the exponent as in the proof of Theorem 7.17 one finds that this posterior measure is again a Gaussian with formal density

$$\mathbb{P}\left(b|\{x_t\}_{t=0}^T\right) \sim \exp\left(-\frac{1}{2}\left|\mathcal{D}^{-\frac{1}{2}}\left(b + \mathcal{D}^{-1}c\right)\right|^2 + \frac{1}{2}\left|\mathcal{D}^{-\frac{1}{2}}c\right|^2\right),$$

where we have used the abbreviations (7.47) and (7.48) to shorten notation. Its mean is given by (7.46) and its covariance is

$$C_o = \left(\Delta^2 + L_T\right)^{-1}. \tag{7.50}$$

This establishes the usual connection between regularization of the likelihood and the mean of an appropriate Bayesian posterior. It is possible to prove existence and robustness of these measures against small errors in the local time, including stability of a numerical implementation, see Papaspiliopoulos et al. (2009) for details.

Finally, let us rewrite (7.46) in a suggestive form:

$$\widehat{b} = \left(\Delta^2 + L_T\right)^{-1}\left(\frac{1}{2}L_T' + M + \tilde{\chi}(\cdot; X_0, X_T) + \Delta^2 b_0\right). \tag{7.51}$$

Heuristically we expect that L_T is $\mathcal{O}(T)$ for large T, and since Δ^2 is $\mathcal{O}(1)$ on low frequencies we expect that equation (7.50) defines a small operator, at least on low frequencies, and that the covariance of the posterior tends to the zero operator as $T \to \infty$. Likewise the mean, given by (7.47), will approach

$$\frac{1}{2}(\log L_T^a)'$$

for large T. Note that we have a similar result for the maximizers of the likelihood for regularized local times, see (7.39).

In summary, these heuristics indicate that the Bayesian framework should be amenable to a posterior consistency result with the posterior measure converging to a Dirac distribution centred at the true drift function.

7.7 Conclusions and further work

In this overview we have illustrated the following points:

- At small scales, data is often incompatible with the diffusion process that we wish to fit.
- In Section 7.3 we saw that
  1. this situation can be understood in the context of fitting averaged/homogenized equations to multiscale data,
  2. in the *averaging* situation fine-scale data produces the *correct* averaged equation,
  3. in the *homogenization* situation fine-scale data produces an *incorrect* homogenized equation.
- In Section 7.4 we saw that
  1. to estimate the drift and diffusion coefficients accurately in the homogenization scenario it is necessary to *subsample*,
  2. there is an optimal subsampling rate, between the two characteristic timescales of the multiscale data,
  3. the optimal subsampling rate may differ for different parameters.
- In Section 7.5 we observed that in the case where the data is smooth at small scales a useful approach can be to fit hypoelliptic diffusions; such models are often also dictated by physical considerations.
- In Section 7.6 we observed that when fitting the drift, another approach is to use estimators which do not see time-ordering of the data and use, instead, the *local time* (or empirical measure) of the data.

There are many open questions for further investigation:

Section 7.3: How to identify multiscale character from time-series?

Section 7.4:

1. If subsampling is used, then what is the optimal subsampling rate?
2. Is subsampling at random helpful?
3. Is it possible to optimize the data available by combining shifts?

4. How to estimate diffusion coefficients from low frequency data?

Section 7.5:

1. A theoretical understanding of the conditions under which hybrid Gibbs samplers, using different approximate likelihoods for different parts of the sampling problem, yield approximately correct samples from the true posterior is yet to be attained.
2. While the recipe described in this section extends to higher order hypoellipticity, the method is still to be tested in this region.

Section 7.6:

1. How to obtain estimates of the local time L_T^a for all a which are good in a suitable norm, e.g. L^2?
2. What is the convergence behaviour as $T \to \infty$ for the mollified maximum likelihood and the Bayesian estimators?
3. How to extend the Bayesian approach to higher dimensions where the empirical measure is even less regular than in the one dimensional case?

Acknowledgements We gratefully acknowledge crucial contributions from all our co-authors on work we cited in this chapter.

7.8 Appendix 1

Let us consider the random functional

$$\mathcal{I}(b; w) = \int_0^1 b^2(x)w(x) + b'(x)w(x)dx. \tag{7.52}$$

where $b(\cdot) \in H^1(0,1)$ and $w(x)$ is a standard Brownian bridge. We claim that this functional is not bounded below and state this as a theorem:

Theorem 7.18 *There almost surely exists a sequence $b^{(n)}(\cdot) \in H^1(0,1)$ such that*

$$\lim_{n\to\infty} \mathcal{I}(b^{(n)}; w) = -\infty \quad \text{a.s.}$$

Proof. For the Brownian bridge we have the representation

$$w(x) = \sum_{i=1}^{\infty} \frac{\sin(i\pi x)}{i}\xi_i \tag{7.53}$$

where the $\{\xi_i\}_{i=1}^{\infty}$ are a sequence of iid normal $\mathcal{N}(0,1)$ random variables. This series converges in $L^2(\Omega; L^2((0,1),\mathbb{R}))$ and almost surely in $C([0,1],\mathbb{R})$, see Kahane (1985).

Now consider the following sequence of functions $b^{(n)}$:

$$b^{(n)}(x) = \sum_{i=1}^{n} \frac{\xi_i}{i} \cos(i\pi x). \tag{7.54}$$

We think of a fixed realization $\omega \in \Omega$ of (7.53) for the time being and note that $\{w(x) : x \in [0,1]\}$ is almost surely bounded in $L^\infty((0,1),\mathbb{R})$, so if there exists a $C > 0$ (which may depend on $\{\xi_i\}_{i=0}^\infty$) such that

$$\|b^{(n)}\|_{L^2} < C \quad \forall n \in \mathbb{N} \tag{7.55}$$

the first integral in (7.52) will stay finite. By Parseval's identity, it is clear that for the sequence of functional (7.54) this will be the case if the coefficients $\frac{\xi_i}{i}$ are square-summable.

Computing the second summand in (7.52) is straightforward, since the series terminates due to orthogonality:

$$\int_0^1 \left(\sum_{i=1}^{\infty} \frac{\sin(i\pi x)}{i} \xi_i \right) \cdot \left(\sum_{j=1}^{n} \frac{\xi_j}{j} \cos(j\pi x) \right)' dx = -\frac{\pi}{2} \sum_{j=1}^{n} \frac{\xi_j^2}{j}.$$

It can now be seen that (7.52) is unbounded from below if the following two conditions are fulfilled:

$$\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{j} \xi_j^2 = \infty \tag{7.56}$$

$$\lim_{n\to\infty} \sum_{j=1}^{n} \frac{1}{j^2} \xi_j^2 < \infty \tag{7.57}$$

We finally allow ω to vary and seek to establish that the conditions (7.56) and (7.57) are almost surely fulfilled. To do this, first note that the random variables being summed are independent. Thus, by the Kolmogorov 0-1 law the probability for convergence is either zero or one. We proceed by applying Kolmogorov's Three-Series Theorem (Theorem 12.5 in Williams (1991)) to each of the three sequences to establish (7.56) and (7.57).

We start by treating (7.56). Denote by $X_j \mid^K$ the truncation of the random variable for some $K > 0$ in the sense:

$$X_j \mid^K (\omega) = \begin{cases} X_j(\omega) & \text{if } |X_j(\omega)| \leq K \\ 0 & \text{if } |X_j(\omega)| > K \end{cases} .$$

To abbreviate notation, define the following two sequences of random variables:

$$
\begin{aligned}
X_j &= \frac{1}{j}\xi_j^2 \\
Y_j &= \frac{1}{j^2}\xi_j^2
\end{aligned}
$$

Now consider the summability of expected values for the sequence X_j: since ξ_j^2 follows a χ-squared distribution with one degree of freedom, its expected value is one. For the truncated variable $X_j\ |^K$, for any $K > 0$, there will be some j^* so that for all $j \geq j^*$ we have that

$$
\mathbb{E}(X_j\ |^K) = \mathbb{E}\left[\frac{1}{j}\left(\xi^2\ |^{jK}\right)\right] > \frac{1}{2j}
$$

Therefore, the expected value summation fails as follows:

$$
\begin{aligned}
\sum_{j=1}^{\infty}\mathbb{E}(X_j\ |^K) &= \sum_{j=1}^{\infty}\frac{1}{j}\mathbb{E}\left(\xi^2\ |^{jK}\right) \\
&\geq \sum_{j=j^*}^{\infty}\frac{1}{2j} = \infty
\end{aligned}
$$

Therefore, the series $\sum_{j=1}^{\infty} X_j$ diverges to infinity almost surely, thus (7.56) is established.

Now let us establish (7.57) using the Three-series theorem. First check the summability of the expected values:

$$
\sum_{j=1}^{\infty}\mathbb{E}(Y_j\ |^K) \leq \sum_{j=1}^{\infty}\mathbb{E}Y_j = \sum_{j=1}^{\infty}\frac{1}{j^2} < \infty
$$

Now let us establish the summability of the variances:

$$
\begin{aligned}
\sum_{j=1}^{\infty}\mathrm{Var}(Y_n\ |^K) &\leq \sum_{j=1}^{\infty}\mathrm{Var}Y_n \\
&= \sum_{j=1}^{\infty}\frac{1}{j^4}\mathrm{Var}\xi_j^2 \\
&= 2\sum_{j=1}^{\infty}\frac{1}{j^4} < \infty
\end{aligned}
$$

where we used that ξ_j^2 follows a χ-squared distribution with one degree of freedom and hence has variance $\mathrm{Var}\xi_j^2 = 2$. Finally, to establish the summability of the tail probabilities we use the following argument for any $K > 0$:

$$\begin{aligned}\sum_{j=1}^{\infty} P(|Y_j| > K) &\leq \sum_{j=1}^{\infty} \frac{1}{K}\mathbb{E}|Y_j| \\ &\leq \frac{1}{K}\sum_{j=1}^{\infty}\frac{1}{j^2} < \infty\end{aligned}$$

where we have used the Markov inequality and the previous calculation of the expected value of $Y_j = |Y_j|$.

To put everything together, let us reconsider the functional $I_B[b]$:

$$\begin{aligned}I_B[b^{(n)}] &= \int_0^1 \left(b^{(n)}\right)^2(x)w(x) + \left(b^{(n)}\right)'(x)w(x)dx \\ &\leq \left(\sup_{x\in[0,1]} w(x)\right)\int_0^1 \left(b^{(n)}\right)^2(x)dx - \frac{\pi}{2}\sum_{j=1}^{n}\frac{1}{j}\xi_j^2 \\ &\leq \left(\sup_{x\in[0,1]} w(x)\right)\frac{1}{2}\sum_{j=1}^{n} X_j - \frac{\pi}{2}\sum_{j=1}^{n} Y_j\end{aligned}$$

Now use the almost surely true convergence and divergence statements (7.56) and (7.57) to conclude:

$$\lim_{n\to\infty} I_B[b^{(n)}] = -\infty \quad \text{a.s.}$$

□

7.9 Appendix 2

Torus

We denote by $\mathbb{T}^d$ the d-dimensional torus. We parameterise the torus by d variables $z_i \in [0, 2\pi]$ where we identify the end points 0 and 2π so as to obtain periodicity in each direction z_i.

Matrix inner product

Given two matrices $A, B \in \mathbb{R}^{n\times m}$ we define their inner product as

$$A : B = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{i,j}B_{i,j}.$$

This defines a positive-definit symmetric bilinear form on $\mathbb{R}^{n\times m}$ and turns this space into an inner product space (also known as a finite dimensional Hilbert space).

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell*. New York: Garland Science.

Campillo, F., & Pitnitski, A. (2002). Effective diffusion in vanishing viscosity. In *Nonlinear Partial Differential Equations and Their Applications* (pp. 133–145). Amsterdam: North-Holland. (France Seminar, Vol. XIV (Paris 1997/1998), volume 31 of Stud. Math. Appl.)

Da Prato, G., & Zabczyk, J. (1992). *Stochastic Equations in Infinite Dimensions*. Cambridge: Cambridge University Press.

Dacorogna, M. M., Gençay, R., Müller, U., Olsen, R. B., & Pictet, O. (2001). *An Introduction to High-Frequency Finance*. San Diego: Academic Press.

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

Evans, L. C. (1998). *Partial Differential Equations*. American Mathematical Society.

Frenkel, D., & Smit, B. (2002). *Understanding Molecular Simulation. From algorithms to applications*. London: Academic Press.

Givon, D., Kupferman, R., & Stuart, A. M. (2004). Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity*, *17*, R55–R127.

Kahane, J.-P. (1985). *Some Random Series of Functions*. Cambridge: Cambridge University Press.

Kaipio, J., & Somersalo, E. (2005). *Statistical and Computational Inverse Problems*. Berlin: Springer.

Kepler, T., & Elston, T. (2001). Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.*, *81*, 3116–3136.

Majda, A. J., & Kramer, P. R. (1999). Simplified models for turbulent diffusion: Theory, numerical modelling and physical phenomena. *Physics Reports*, *314*(4-5), 237–574.

Majda, A. J., Timofeyev, I., & Vanden-Eijnden, E. (1999). Models for stochastic climate prediction. *Proc. Natl. Acad. Sci. USA*, *96*(26), 14687–14691.

Marcus, M. B., & Rosen, J. (2006). *Markov Processes, Gaussian Processes, and Local Times*. Cambridge: Cambridge University Press.

Melbourne, I., & Stuart, A. M. (2011). A note on diffusion limits of chaotic skew-product flows. *Nonlinearity*, *24*, 1361-1367.

Olhede, S. C., Sykulski, A., & Pavliotis, G. A. (2009). Frequency domain estimation of integrated volatility for Itô processes in the presence of market-microstructure noise. *SIAM J. Multiscale Model. Simul.*, *8*, 393–427.

Papaspiliopoulos, O., Pokern, Y., Roberts, G. O., & Stuart, A. M. (2009). *Nonparametric Bayesian drift estimation for one-dimensional diffusion processes* (Tech. Rep.). CRiSM report no. 09-29, Warwick.

Papaspiliopoulos, O., Pokern, Y., Roberts, G. O., & Stuart, A. M. (2011). Nonparametric estimation of diffusions: a differential equations approach. (Submitted, available from
http://www.econ.upf.edu/~omiros/papers/submission_arxiv.pdf)

Papavasiliou, A., Pavliotis, G. A., & Stuart, A. M. (2009). Maximum likelihood drift estimation for multiscale diffusions. *Stochastic Process. Appl.*, *119*, 3173–3210.

Pavliotis, G. A., & Stuart, A. M. (2007). Parameter estimation for multiscale diffusions. *J. Stat. Phys.*, *127*, 741–781.

Pavliotis, G. A., & Stuart, A. M. (2008). *Multiscale Methods: Averaging and Homogenization*. New York: Springer.

Pokern, Y. (2006). *Fitting stochastic differential equations to molecular dynamics data*. Unpublished doctoral dissertation, Warwick University.

Pokern, Y., Stuart, A. M., & Vanden-Eijnden, E. (2009). Remarks on drift estimation for diffusion processes. *Multiscale Modeling & Simulation*, *8*, 69-95.

Pokern, Y., Stuart, A. M., & Wiberg, P. (2009). Parameter estimation for partially observed hypoelliptic diffusions. *J. Roy. Statist. Soc. B*, *71*, 49–73.

Schlick, T. (2000). *Molecular Modeling and Simulation – an Interdisciplinary Guide*. New York: Springer.

Williams, D. (1991). *Probability with Martingales*. Cambridge: Cambridge University Press.

Zhang, L., Mykland, P. A., & Aït-Sahalia, Y. (2005). A tale of two time scales: Determining integrated volatility with noisy high frequency data. *J. Amer. Statistical Assoc.*, *100*, 1394-1411.