Parameter estimation techniques for hypoelliptic ergodic diffusions

MSIAM M2 STAT

Research project performed at
Laboratoire Jean Kuntzmann

Student: **Anna Melnykova**
Supervisor: **Adeline Leclercq-Samson**

Grenoble 2017

Acknowledgements

I would like to start with expressing my deepest gratitude to my supervisor, Adeline Leclercq-Samson, for her continuous support and positive attitude toward people in general and me in particular.

I am also extremely thankful to everyone who has ever lead existentialist talks with me at the kitchen about university, literature, mathematics and the purpose of life in general. Among the others, special thanks to Sasha Burashnikova and Ilnura Usmanova for being such patient listeners, especially in the last few weeks.

Finally, I would like to thank all the Ensimag professors whom I have had pleasure to work with during the first semester and also Laboratoire Jean Kuntzmann and IDEX for their financial support.

And, of course, I feel grateful to my mother, Valentyna Melnykova, for having tremendous patience and being always by my side all these years.

Abstract

This thesis is devoted to the parameter estimation for discretely observed multidimensional hypoelliptic diffusions which naturally arise as models of neuronal activity. We treat the case when the multidimensional data is fully available. Chapters are organized as follows: in Chapter 2 we give important probabilistic results which are used in this work. In Chapter 3 we describe the model and set necessary assumptions. Chapter 4 is essentially divided in two parts: in Section 4.1 we propose a scheme which allows to establish a link between observations sampled from the continuous model and discrete-time model, and study its properties. In Section 4.2 we build an estimator based on the log-likelihood of the proposed discrete model. Then in Chapter 5 we show the link between our scheme and previous works dealing with the same problem. We conclude our study with numerical experiments (Chapter 6). As an epilogue we discuss yet unsolved problems and possible extensions of our method with the most important issue being the case of incomplete observations (Chapter 7). Proofs of theoretical results and formal derivations are gathered in Appendix.

Key words: hypoelliptic diffusion, statistical inference, discretization, FitzHugh Nagumo model

Contents

1

Introduction

Neurons are excitable cells which are considered to be the main information processing units in a brain. Body of a neuron is divided into three functionally different parts: *soma* (which is a central processing unit of the system), *dendrite* (can be considered an input device which transmits information into soma) and *axon* ("output device"). We can measure an activity of one single neuron by placing an electron in the soma. Then the voltage is measured relative to that at the reference electrode placed in the body fluids (**intracellular measurements**). What is measured in the discrete time is a membrane potential, that is difference between the interior and the exterior of the cell. We are interested in action potentials (also called **spikes**) — stereotypic events, when the difference between the voltage on the electrode inside and outside the cell becomes bigger than some threshold. They are of interest of neuroscientists as the spiking activity is directly related to how the information is processed in a brain.

The most natural way is to consider spike occurrences as a point process, often a Poisson process. This approach is widely studied in the literature and often used in neuroscience (see, in particular, [Gerstein and Mandelbrot, 1964]). However, it is observed that the spiking times within a network of neurons or even within one single neuron are not always independent on each other. In that case it is more natural to use a system of Hawkes processes. Problem of goodness-of-fit tests for point processes is treated, in particular, in [Reynaud-Bouret et al., 2014].

However, point processes can not capture all the physical processes inside a neuron that precede a spike. Thus, spike trains are often considered to be a realization of a stochastic diffusion of certain kind. A good reference to give here is [Tuckwell, 2005].

One of the most accurate descriptions of a firing mechanism of the one single neuron is a Hodgkin-Huxley model [Hodgkin and Huxley, 1952] (Nobel Prize in Physiology or Medicine 1963)). It couples the equation which describes a potential of a neuron with differential equations which represent open fractions of ion channels of different kind. It is highly-nonlinear model, which serves as base for several more complex or, on the contrary, relaxed models. Complex models also use the geometry of an axon, or introduce additional channel ion population aiming to describe the spike occurrences in the most accurate way. For other purposes, where computational cost plays more important role — for example, when we

need to model neuronal activity in a complex neuronal net which consists of billion of neurons, we can sacrifice the precision and work with relaxed models. In particular, we should mention a Morris-Lecar model [Morris and Lecar, 1981], which substitutes three equations related to ion channels in Hodgkin-Huxley model by only one which represents the membrane conductance. More information on simulation problems can be found in [Brette et al., 2007].

Among modification of the Hodgkin-Huxley model it is worth to mention a 2-dimensional FitzHugh-Nagumo model [Fitzhugh, 1961] [Nagumo et al., 1962], which was initially introduced as a deterministic system, and then expanded by adding noise to both coordinates. This model is less plausible from physical point of view in comparison to a Morris-Lecar model, it is however much easier to analyze due to a polynomial (and not trigonometrical) drift.

Despite the number of existing models, it is still very hard to treat real data. The main problem is that it is possible to observe experimentally only the membrane potential, and not the conductance — thus we always have to deal with a problem of missing data.

Further, it is not clear, which of the variables are directly perturbed by noise. The question of noise placement in Hodgkin-Huxley model and its derivatives received a lot of attention in the last years and is of great importance [Goldwyn and Shea-Brown, 2011]. If, for example, only a membrane potential is given by a deterministic equation, then we face a very specific class of diffusions with a degenerate variance coefficient, namely **hypoelliptic diffusions**. Hypoelliptic diffusions also naturally occur in multidimensional models of a neuron population (such as stochastic approximation of a Hawkes process [Ditlevsen and Löcherbach, 2015]).

Hypoellipticity can be intuitively explained in the following way: though the covariance matrix of noise is singular due to the fact that only one coordinate is driven by noise directly, smooth transition density with respect to the Lebesque measure still exists. That is the case when the noise is propagated to the remaining coordinate through the drift term.

In this work we face the problem of the parameter estimation for hypoelliptic diffusions. More precisely, we consider a two-dimensional system of stochastic differential equations (SDE) of the form:

$$\begin{cases} dX_t = a_1(X_t, Y_t; \theta)dt \\ dY_t = a_2(X_t, Y_t; \theta)dt + b(X_t, Y_t; \sigma)dW_t, \end{cases} \tag{1.1}$$

where $(X_t, Y_t)^T \in \mathbb{R} \times \mathbb{R}$, $(a_1(X_t, Y_t; \theta),\ a_2(X_t, Y_t; \theta))^T$ is a drift term, $(0, b(X_t, Y_t; \sigma))^T$ is the variance, W_t is a standard Gaussian noise defined on some probability space $(\Omega, \mathcal{F}, P)$, (θ, σ) is a vector of the unknown parameters, taken from some compact set $\Theta_1 \times \Theta_2$.

The main objective of our work is to propose a method of estimation of the unknown parameters from discretely observed data, restricting our attention to the case of (1.1), when $b(X_t, Y_t; \theta) \equiv \sigma = const$. We assume that both variables are discretely observed on some time interval $[0, T]$, not necessarily fixed.

As it was already mentioned, properties of the process described by hypoelliptic SDE significantly differ from those of elliptic diffusion, when all coordinates are driven by Gaus-

sian noise. They are more difficult to study. First problem is that each coordinate has the variance of different order. It is the main cause why classical numerical approximation methods do not work well with hypoelliptic diffusions: in particular, it is proven that for hypoelliptic systems classical low-order schemes of approximation do not preserve ergodic properties of the true process [Mattingly et al., 2002]. Second problem is the singularity of the covariance matrix. Most popular methods of parametric inference are based on the approximation of the transition density with the piece-wise Gaussian processes (see, for example [Kessler, 1997]), and in our case they cannot be applied directly because we cannot find an inverse of an original covariance matrix.

Now let us present some solutions that have been proposed in the literature. First hypoelliptic models have arised as a stochastic expansion of 2-dimensional deterministic dynamical systems (for example, perturbed by noise Van der Pol oscillator [Van der Pol, 1920]). Thus it is natural to begin our study with the class of stochastic Damping Hamiltonian systems, also known as Langevin equations (see, among others, [Gardiner and Collett, 1985]). They are defined as the solution of the following SDE:

$$\begin{cases} dX_t = Y_t dt \\ dY_t = a_2(X_t, Y_t; \theta)dt + b(X_t, Y_t; \sigma)dW_t. \end{cases} \tag{1.2}$$

Particular case of Hamiltonian system with $b(X_t, Y_t; \sigma) \equiv \sigma$ and $a_2(X_t, Y_t; \theta) = g_1(X_t; \theta)X_t + g_2(X_1; \theta)Y_t$ is considered in [Ozaki, 1989], where the link between continuous solution of (1.2) and corresponding discrete model is obtained with the help of so-called Local Linearization scheme. The idea of this scheme is the following: for a system of SDE with a non-constant drift and constant variance, its solution can be interval-wise approximated by a system with a linear drift, and original covariance matrix being expanded by adding higher-order terms. It allows to construct an estimator based on the Maximum Likelihood Estimator for a corresponding discretized process.

A more recent work [Pokern et al., 2007] attempts to solve the problem of non-invertibility of the covariance matrix for the particular case of system (1.2) with constant variance with the help of Itô-Taylor expansion of the transition density. It propagates noise into the first coordinate with order $\Delta^{\frac{3}{2}}$. However the drift term is approximated up to the terms of order Δ, which results in poor numerical performance.

In [Samson and Thieullen, 2012], it is shown that a consistent estimator for fully and partially observed data can be constructed using only the discrete approximation of the second equation of the system (1.2). This method works reasonably good in practice even for some more general models when it is possible to convert a system (1.1) to a simpler form (1.2). However, transformation of the observations sampled from the continuous model (1.1) in that way requires the knowledge about the parameters involved in the first equation, which is not possible in real-world applications.

Untill recently no estimation techniques were available for the systems of more general form (1.1). First step in this direction is the preprint [Ditlevsen and Samson, 2017]. They construct a consistent estimator for arbitrary n-dimensional system of SDE with no noise in the first coordinate using a 1.5 strong order discretization scheme based on a Itô-Taylor

expansion.

In this work we propose a new estimation method for models of type (1.1), adjusting the local linearization scheme described in [Ozaki, 1989] to more general class of SDEs. This scheme propagates the noise to both coordinates of the system and allows to obtain an invertible covariance matrix. We start with describing the scheme, approximating the transition density and proposing a contrast estimator based on the discretized log-likelihood. While we attempt to estimate the parameters included in drift and diffusion simultaneuously, we also explain in which cases and how the estimator can be splitted. Then we compare our approach with the other schemes: in particular, Pokern estimator [Pokern et al., 2007], Euler contrast for stochastic Damping Hamiltonian system described in [Samson and Thieullen, 2012], and the estimator based on 1.5 strong order Euler scheme for multidimensional systems suggested in [Ditlevsen and Samson, 2017]. We finish our study with numerical experiments, testing our method on the hypoelliptic FitzHugh-Nagumo model and the stochastic approximation of the Hawkes process suggested in [Ditlevsen and Löcherbach, 2015].

2

Definitions and preliminary results

This chapter is given exclusively for reference. It is organized as follows: section **Probability theory** contains definitions and useful results which are later used in Chapter 3, section **Discrete time models** is a preface to Chapters 4 and 5.

2.1 Probability theory

2.1.1 Stochastic differential equations

As we want to restrict ourselves only to the results which are directly used in the following chapters, we refer to [Karatzas and Shreve, 1987, Oksendal, 2003] for details, formal proofs and missing definitions. We also take for granted some notions from measure theory without giving specifying. Throughout the chapters we consider filtered complete probability space $(\Omega, \mathcal{F}_t, P)$.

Stochastic integral of some measurable function $f(t)$ with respect to a Brownian motion W_t is defined in a similar way to a Lebesque integral:

$$\int_0^t f(s)dW_s = \lim_{\Delta \to 0} \sum_{i=0}^{N} f(t_i^*) \left(W_{t_{i+1}} - W_{t_i} \right),$$

where $0 = t_0 < t_1 < \cdots < t_N = t$ is a partition of a time interval such that $\forall i : t_{i+1} - t_i = \Delta$, and t_i^* is some point on interval $[t_i, t_{i+1})$. On the contrary to a deterministic case, choice of point t_i^* results in stochastic integrals with different properties. In particular, when $t_i^* = t_i$, we speak about **Itô integral** (denoted by $\int_0^t f dW_s$), and when $t_i^* = \frac{t_{i+1} - t_i}{2}$ — about **Stratonovich integral** (denoted by $\int_0^t f \circ dW_s$). Further under **stochastic integral** we refer to an Itô integral, unless otherwise specified.

First important property of an integral in Itô form is that it is a martingale. Second, also known as **Itô isometry**, states that the variance of an Itô integral equals to the expectation of a deterministic integral:

$$\mathbb{E}\left[\int_0^t f(s, \omega) dW_s \right]^2 = \mathbb{E}\left[\int_0^t f^2(s, \omega) ds \right],$$

where ω is a measurable random variable.

Notion of the stochastic integral allows us to introduce a **stochastic differential equation** (SDE). Consider a d-dimensional process $Z_t = (Z_t^{(1)}, \dots Z_t^{(1)})$, where each element is described by the following equation:

$$Z_t^{(i)} = z_i + \int_0^t a_i(t, Z_s)ds + \sum_{j=1}^{r} \int_0^t b_{ij}(t, Z_s)dW_s^{(j)}, \quad 0 \le t < \infty,\ 1 \le i \le d, \tag{2.1}$$

where $W = \{W_t, \mathcal{F}_t : 0 \le t < \infty\}$ is a Brownian motion in $\mathbb{R}^r$ and the coefficients $a_i,\ b_{ij} : \mathbb{R} \times \mathbb{R}^d \to R,\ 1 \le i \le d,\ 1 \le j \le r$ are Borel-measurable. $z = (z_1, \dots, z_n)$ is a $\mathcal{F}_0$-measurable vector of initial conditions. Equivalently it can be written as:

$$dZ_t = A(t, Z_t)dt + \sum_{j=1}^{r} B_j(t, Z_t)dW_t^{(j)}, \tag{2.2}$$

where $A(Z_t)$ is a vector-function with elements a_i (a **drift** term), and $B_j(Z_t)$ is a vector-function with elements b_{ij} (**variance** term).

A continuous, adapted d-dimensional process $Z = \{Z_t, \mathcal{F}_t : 0 \le t < \infty\}$ is called a **solution** of (2.1) if it satisfies (2.1) for every $z \in \mathbb{R}^n$. In the literature it is often referred to as an **Itô process**. We say that it is **uniformly elliptic** if $\exists c > 0 \quad BB^T \ge cI_n$ (that corresponds to "ideal" system, where each coordinate is perturbed by exactly one source of independent noise).

Now we can introduce a change-of-variables formula, also known as **Itô formula**. Let Z_t be a solution of a process (2.1), and $g(t, x) \in C^2 : [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$. Then the process $\tilde{Z}(t, \omega) = g(t, Z_t)$ is again an Itô process, with an element $\tilde{Z}_t^{(i)}$ given by:

$$d\tilde{Z}_t^{(i)} = \frac{\partial g_i}{\partial t}(t, Z_t)dt + \sum_{j=1}^{r} \frac{\partial g_j}{\partial Z_t^{(j)}}(t, Z_t)dZ_t^{(j)} + \frac{1}{2}\sum_{j,k=1}^{r} \frac{\partial^2 g_i}{\partial Z_t^{(j)} \partial Z_t^{(k)}}(t, Z_t)dZ_t^{(j)} Z_t^{(k)}.$$

This result, however, does not hold for the SDEs in Stratonovich form. In that case the change-of-variables formula is more closely related to that in deterministic integrals.

Important note is that both forms of SDEs are mutually interchangeable. By that we mean that if some process Z_t satisfies equation (2.2), then it always satisfies an equation in Stratonovich form:

$$dZ_t = \tilde{A}(t, Z_t)dt + \sum_{j=1}^{r} B_j(t, Z_t) \circ dW_t^{(j)},$$

where each element of $\tilde{A}(t, Z_t)$ equals to

$$\tilde{a}_i(t, Z_t) = a_i(t, Z_t) - \frac{1}{2}\sum_{j=1}^{r}\sum_{k=1}^{d} \left(\frac{\partial}{\partial Z^{(k)}} b_{ik}(t, Z_t) \right) b_{ij}(t, Z_t).$$

This result can be found, in particular, in [Kloeden et al., 2003]. Note that when $B(t, Z_t) \equiv const$, both forms coincide. Also note that when we have a system of form (1.1), then both forms coincide except for the drift term of the second variable.

We do not aim to explain the advantages and the drawbacks of using each form, as it is not particularly important for our problem.

Solution of (2.2) can be characterised in terms of it infinitesimal generator, defined as:

$$Lf(z) = \lim_{\Delta \to 0} \frac{\mathbb{E}\big[f(Z_{t+\Delta})|Z_t = x\big] - f(z)}{\Delta},$$

which is equivalent to

$$Lf(z) = \sum_i a_i(z) \frac{\partial f}{\partial z_i}(z) + \frac{1}{2} \sum_{j=1}^{r} \sum_{i,k} \big(B_j(z) B_j(z)^\top\big)_{i,k} \frac{\partial^2 f}{\partial z_i \, \partial z_k}(z). \tag{2.3}$$

On of the most widely-studied cases of the SDE is a linear stochastic differential equation. Solution for this equation can be given in explicit form, which is rather a rare property. Here we consider a linear homogeneous equation and will also refer to this example in Chapter 4, when proposing a numerical scheme for approximating more complex SDEs. Following example is based on Section 5.6 in [Karatzas and Shreve, 1987].

Example 2.1 (**Linear homogeneous equation**). *Consider a d-dimensional equation of the form:*

$$dZ_t = K(t) Z_t dt + \sigma(t) dW_t, \; 0 \le t < \infty, \; Z_0 = const, \tag{2.4}$$

where W_t is a $r-$dimensional Brownian motion, and matrices $K(t)$, $\sigma(t)$ of size $d \times d$ and $d \times r$ respectively are non-random, measurable and locally bounded. We denote by $\Phi(t)$ a matrix function which is a fundamental solution of a corresponding deterministic equation $dz_t = K(t) z_t dt$ with an initial condition $z_0 = Z_0$. Then, by Itô formula we can express the solution of (2.4) in terms of $\Phi(t)$ as:

$$Z_t = \Phi(t) \left[Z_0 + \int_0^t \Phi^{-1}(s) \sigma(s) dW_s \right], \; 0 \le t < \infty.$$

Interesting fact is that when both matrices $K(t)$ and $\sigma(t)$ are constant, $\Phi(t)$ has the following form:

$$\Phi(t) = e^{tK} \triangleq \sum_{i=0}^{\infty} \frac{t^i}{i!} K^i$$

2.1.2 Hypoellipticity

Here our first aim is to give an intuition of the hypoellipticity and not to focus much on the theory. Main reference for this part is the work [Malliavin and Thalmaier, 2006] (especially Chapter 5), which provides a more accurate bridge between the notion of hypoellipticity for vector fields and that for stochastic diffusions.

By **elliptic** diffusions we understand solutions of (2.1) with a non-singular covariance matrix. By **hypoelliptic** diffusions we mean that the covariance matrix is singular, but the transition density still exists.

Definition 2.1. *Given two vector fields A_1 and A_2 their **Lie bracket** $[A_1, A_2]$ is a vector field C with the components:*

$$C^{\xi} \triangleq \sum_{\eta \in \{1,\dots,N\}} \left(A_1^{\eta} \frac{\partial A_1}{\partial \eta^n} - A_2^{\eta} \frac{\partial A_2}{\partial \eta^n} \right), \quad \xi \in \{1, \dots, N\} \tag{2.5}$$

Definition 2.2. ***Lie algebra** $\mathcal{A}$ generated by n vector fields $A_1, \dots, A_n$ is defined as a vector space of all fields obtained as linear combinations with constant coefficients of the*

$$A_k,\ [A_k, A_l],\ [[A_k, A_l], A_s], \quad \textit{etc.}$$

Given $r \in \mathbb{R}^n$, let $\mathcal{A}(r) \triangleq \{\zeta \in \mathbb{R}^n | \zeta = Z(r) \text{ for some } Z \in \mathcal{A}\}$

Definition 2.3. *We say that vector fields $A_1, \dots, A_n$ satisfy the Hörmander criterion for hypoellipticity if $A_1, \dots, A_n$ are infinitely often differentiable and if*

$$\mathcal{A}(r) = \mathbb{R}^N \quad \forall r \in \mathbb{R}^N$$

How the Hörmander criterion can be applied to a specific two-dimensional stochastic differential equation in Itô form will be shown in Chapter 3.

2.1.3 Ergodicity

Informally speaking, if the process is ergodic, then we can deduce its statistical properties from sufficiently long random sample of this process. In terms of dynamical systems that means that almost all points in any subset of a state space eventually revisit this set, and there exists a time average along each trajectory. Exploring an ergodic property of particular type of SDE is out of scope of this work, thus we just restrict ourselves to the most important results and work with the models which are known to be ergodic. For more formal reference reader is refered to [Khasminskii, 1969], where ergodic theory in application to SDEs is nicely elaborated.

Theorem 2.1 (Continuous ergodic theorem)**.** *$\forall f$ with polynomial growth at infinity the following holds with probability 1:*

$$\frac{1}{T} \int_0^T f(Z_t) dt \to \int f(z) \nu_0(dz) = \mathbb{E}[f(\zeta)],$$

*where ν_0 is a **stationary density** and ζ is some random variable with density ν_0.*

However, we need to be careful when moving from considering a continuous time model to its approximated discrete-time version. As will be later explained, approximation does not always preserve an ergodic property of a true process.

2.2 Discrete time models

2.2.1 Approximation methods

Except for some very simple cases, the solution of a SDE cannot be given in an explicit form. In various applications it is enough to approximate the solution numerically in order to understand the behaviour of the stochastic dynamical system. In this work we will approximate a solution of (1.1) as we need to generate data to work with. Discretization schemes are also essential when it comes to inferring the properties of the underlying system from discrete time observations. The main reference for this section is [Kloeden et al., 2003].

Let Z_t be the continuous-time solution of the equation (1.1) and Z^Δ be its discrete time approximation with a fixed step size Δ. Accuracy of the approximation can be measured in terms of *mean* and *absolute* error. Naturally, it is a desirable that the accuracy of the scheme increases as the time step Δ decreases. The **mean error** examines the approximation of the first moment of an Itô process Z_t:

$$\mu_\Delta = \left\| \mathbb{E} Z_T - \mathbb{E} Z_T^\Delta \right\|.$$

The **absolute error** ϵ_Δ is an expectation of the norm of the difference between the approximation and the Itô process at time T:

$$\epsilon_\Delta = \mathbb{E} \left\| Z_T - Z_T^\Delta \right\|$$

When $\epsilon_\Delta \underset{\Delta \to 0}{\longrightarrow} 0$, then we say that the scheme **converges strongly** to an Itô process Z_t.

If we have different strongly convergent schemes, we want to know how to compare them properly. For that purpose we have the notion of the **order of strong convergence** (note that the bound does not depend on the time interval):

Definition 2.4. *We say that a discrete time approximation Z^Δ **converges strongly** to Z at time T **with order** γ, if*

$$\mathbb{E} \left\| Z_T - Z_T^\Delta \right\| \leq \Delta^\gamma.$$

Note that the order of convergence, in general, depends on the type of the equation we consider.

The classical and the most widely used approximation scheme is a first-order Euler-Maruyama method with a strong order of convergence 0.5 [Kloeden et al., 2003]. It works in a similar way as an Euler method for deterministic differential equations: it approximates the solution by a piecewise constant function. When the time step Δ is small enough, it allows to obtain reasonably good approximation. However it is not stable when the system is highly non-linear (as it is often the case with the models of neuronal activity).

With the following example we want to illustrate a major weakness of this scheme when applied to SDE of type (1.1).

Example 2.2 (**Euler-Maruyama scheme**). *Given system* (1.1)*, its solution is approximated by:*

$$Z_{i+1} = \bar{A}_{1.0}(Z_i; \theta) + \bar{B}_{1.0}(Z_i, \sigma)\Xi_i \tag{2.6}$$

where $\Xi = (\xi_1, \xi_2)^T$ *is a two-dimensional standard Gaussian noise, drift term is approximated by*

$$\bar{A}_{1.0}(Z_i; \theta) = \begin{pmatrix} X_i + \Delta a_1(X_i, Y_i; \theta) \\ Y_i + \Delta a_2(X_i, Y_i; \theta) \end{pmatrix}$$

and the variance, respectively, by $\bar{B}_{1.0}$ *given by:*

$$\bar{B}_{1.0}(Z_i, \sigma) = \begin{pmatrix} 0 & 0 \\ 0 & b(X_i, Y_i, \sigma)\sqrt{\Delta} \end{pmatrix}$$

First note that the approximation (2.6) leads to a degenerate Gaussian diffusion, indeed:

$$\mathbb{E}[Z_{i+1}|Z_i] \sim \mathcal{N}(\bar{A}_{1.0}, \Sigma_{1.0}),$$

where $\Sigma_{1.0} = \bar{B}_{1.0}\bar{B}_{1.0}^T$ is singular.

Second important issue is that the discretized process does not preserve the ergodic property of the original process [Mattingly et al., 2002]. The same happens with a slightly more complex Milstein scheme (see [Kloeden et al., 2003] for details). In order to overcome this problem more sophisticated schemes are introduced, which will be discussed further.

The main advantage of Euler-Maruyama scheme is that it is widely studied and there exists a number of important results which allow to accurately examine the properties approximated process. In particular, we recall the result from [Bally and Talay, 1996], which shows the difference between the transition density of an elliptic diffusion and its approximation:

Proposition 2.1. *Let the drift functions and the diffusion coefficients be 2 times differentiable with bounded derivatives of all orders up to 2. Let* $p(z|z_0)$ *and* $p_\Delta(z|z_0)$ *be the exact transition density of* z *at time* Δ *under an elliptic SDE and the transition density of the corresponding Euler approximation. Then:*

$$\begin{aligned} \forall z, \quad & p(z|z_0) + p_\Delta(z|z_0) \leq Ce^{-C'|z-z_0|^2} \\ \forall z, \quad & |p(z|z_0) - p_\Delta(z|z_0)| \leq C\Delta e^{-C'|z-z_0|^2}, \end{aligned}$$

where C *and* C' *are constants not depending on* Δ

To the best of our knowledge, for more complex higher-order schemes similar bounds have not been found yet, but they can be conjectured. However, for a limited number of SDEs — such as Ornstein-Uhlenbeck process, where the transition density can be found in the explicit form, it can be computed directly, giving us an insight about the properties of the discretization scheme.

2.2.2 Parametric inference

In many applications we have to deal with the problem of finding a mathematical model that fits discretely observed data. Moreover, sometimes continuous-time model serves only as an abstraction for real process which exists exclusively in a discrete setting. In this work we focus on finding the parameters of an already known model from discretely observed data.

Let us suppose we observe some process Z_t within equal intervals of time Δ up to moment T, so that we have a vector of observations $\{Z_i\}, i \in 1, \ldots, N$, that is $\Delta N = T$.

The desired properties for an estimator are the **consistency** and **asymptotic normality**. Roughly speaking, we say that the estimator is consistent if it converges to the real value of the parameter, given enough data. How much data is "enough" to estimate the parameters of some particular model is also a question to be answered. In some cases, in order to obtain a consistent scheme it is enough to observe a process up to some finite time and let the time step decrease to zero (*classical* scheme). In more complicated models standard requirement is to let $\Delta_N \to 0$ as $N \to \infty$ (*high frequency* scheme).

Asymptotic normality implies that the distribution of a given estimator around the true parameter θ_0 approaches a normal distribution with a standard deviation shrinking to $\frac{1}{N}$ as the sample size grows.

Vast majority of existing methods of parametric inference is based on maximizing the likelihood of the model. But as the transition density of the continuous time solution is unknown in general case, "classical" MLE cannot by applied. The most common way to deal with this problem is to approximate an unknown joint probability distribution by a sequence of locally Gaussian processes, for which density is known, and then build an estimation with the help of so-called **contrast function**.

This approach is widely studied for one-dimensional SDE (for references see, in particular, [Florens-Zmirou, 1989] and [Kessler, 1997]). The idea is the following: for one-dimensional process $\tilde{X}$ with mean $\mathbb{E}(\tilde{X})$ and variance $Var(\tilde{X})$, the contrast is defined as:

$$\sum_{i=1}^{N} \left[\frac{\tilde{X} - \mathbb{E}(\tilde{X})}{Var(\tilde{X})} + \log(Var(\tilde{X})) \right], \tag{2.7}$$

which is just an absolute value of the log-likelihood of a Gaussian density. Then true values of the parameters can be found by minimizing (2.7).

When the process $\tilde{X}$ is given as a solution of equation (2.2), this contrast is used by approximating drift and the variance term on each interval expanding the infinitesimal generator of the process (2.3), and then replacing $\mathbb{E}(\tilde{X})$ and $Var(\tilde{X})$ by the mean and variance of the approximated process. Parametric inference for multidimensional systems is also often based on the same approach, but with (2.7) being replaced by the log-likelihood of a multivariate Gaussian distribution.

3

Hypoelliptic SDEs

3.1 Model and notations

Let us restrict our attention to the case of (1.1), when $b(X_t, Y_t; \theta) \equiv \sigma = const$, that is:

$$\begin{cases} dX_t = a_1(X_t, Y_t; \theta)dt \\ dY_t = a_2(X_t, Y_t; \theta)dt + \sigma dW_t. \end{cases} \tag{3.1}$$

We assume that both variables are discretely observed at equally spaced periods of time on some finite time interval $[0, T]$, with a vector of observations being $Z_i = (X_i, Y_i)^T$, where Z_i is a value of the process at the time $i\Delta$, $i \in 0 \dots N$. That means that the interval is splitted into N equal parts of size Δ. We further assume that it is possible to draw a sufficiently large and accurate sample of data, by that we mean that T may be infinitely large, and the partition size Δ — infinitely small.

To avoid cumbersome notations, throughout the paper we will use the following abbreviations: $\partial_x f \equiv \frac{\partial f}{\partial x}(x, y; \theta)$ and $\partial_y f \equiv \frac{\partial f}{\partial y}(x, y; \theta)$. Note that we suppress the dependency on the parameter θ, as we do not use it most of the time. In the proofs, where the value of the parameter is important, we will introduce additional indices. Real values of the parameters are denoted by θ_0, σ_0.

In addition, as we will sometimes refer to the system in a vector form, let us also denote a drift term by $A(Z_t; \theta) = (a_1(X_t, Y_t; \theta),\ a_2(X_t, Y_t; \theta))^T$.

We also adopt the notations from work [Pokern et al., 2007] and refer to the variable Y_t which is directly driven by Gaussian noise as "rough", and to the X_t as "smooth".

3.2 Assumptions

3.2.1 Hypoellipticity

First, we want to show that a sufficient condition for (3.1) to be hypoelliptic is the following:

A1 Functions $a_1(x, y; \theta)$ and $a_2(x, y; \theta)$ have bounded derivatives of every order up to 3 with respect to every coordinate, uniformly in θ, furthermore:

$$\forall (x, y) \in \mathbb{R}^2 : \quad \partial_y a_1 \neq 0$$

Note that the Stratonovich form of (3.1) coincides with the Itô representation, as σ is constant:

$$\begin{cases} dX_t = a_1(X_t, Y_t; \theta)dt \\ dY_t = a_2(X_t, Y_t; \theta)dt + \sigma \circ dW_t, \end{cases}$$

Now we write the coefficients of this system as two vector fields:

$$A_0(x, y) = \begin{pmatrix} a_1(x, y; \theta) \\ a_2(x, y; \theta) \end{pmatrix} \qquad A_1(x, y) = \begin{pmatrix} 0 \\ \sigma \end{pmatrix}$$

and compute their Lie bracket:

$$[A_0, A_1] = \begin{pmatrix} -\sigma \partial_y a_1 \\ -\sigma \partial_x a_2 \end{pmatrix}.$$

By (A1) the first element of this vector is not equal to 0, we conclude that A_1 and $[A_0, A_1]$ generate $\mathbb{R}^2$. Thus the Hörmander condition is satisfied and the system is indeed hypoelliptic. As a consequence we may state that the transition density for system (3.1) exists, though not necessarily has an explicit form. It will be shown later that this property allows us to propagate the noise to all the elements of the system.

Strictly speaking, the derivative may be equal to 0 on finite set of points, without breaking the hypoellipticity assumption. Nevertheless, we want it to be non-zero in every point, because as it will be shown later, otherwise at the discretization step some of the functions we use may not be defined.

3.2.2 Existence of the solution and ergodicity

Before considering discrete data, we need to know if the process governed by (3.1) is well conditioned. In particular, we want to be sure that the solution exists and is unique, and that the data sample can be used to investigate this solution.

A2 *Lipschitz and linear growth conditions.* There exists a constant K_θ such that:

$$\begin{aligned} \|A(Z_t; \theta) - A(Z_s; \theta\| &\leq K_\theta \|Z_t - Z_s\| \\ \|A(Z_t; \theta)\| &\leq K_\theta^2 (1 + \|Z_t\|^2) \end{aligned}$$

for every $t, s \in [0, \infty)$. Further, let ξ_0 be the random variable which denotes the initial value of the process Z_t, such that $\mathbb{E}\|\xi_0\|^2 < \infty$.

A3 Process Z_t is ergodic and there exists a unique invariant probability measure ν_0 with finite moments of any order.

A4 Both functions $a_1(Z_t;\theta)$ and $a_2(Z_t;\theta)$ are identifiable, that is $a_i(Z_t;\theta) = a_i(Z_t;\theta_0) \Leftrightarrow \theta = \theta_0$.

(A2) implies existence and uniqueness in law of the weak solution of the system (3.1) [Karatzas and Shreve, 1987] and (A4) is a standard condition which is needed to prove the consistency of the estimator.

(A3) ensures that we can apply the weak ergodic theorem, that is, for any continuous function f with polynomial growth at infinity:

$$\frac{1}{T}\int_0^T f(Z_s)ds \underset{T\to\infty}{\longrightarrow} \nu_0(f) \quad \text{a.s.}$$

We do not investigate the conditions on the system (3.1) under which process Z_t is ergodic, as it is not the main focus of this work. Ergodicity of stochastic damping Hamiltonian system is studied in [Wu, 2001] — thus we refer to this work for necessary conditions for this particular case of hypoelliptic systems. Conditions for a wider class of hypoelliptic SDEs can be found in [Mattingly et al., 2002].

For us it is also important to know that if the process Z_t is ergodic, then its sampling $\{Z_i\}$, $i \in [0, N]$ is also ergodic [Genon-Catalot et al., 2000].

Further, under our assumptions holds an important Lemma from [Kessler, 1997] which is a consequence of the continuous ergodic theorem 2.1. It states

Lemma 3.1. *Let $\Delta \to 0$ and $N\Delta \to \infty$, let $f \in \mathbb{R}\times\Theta \to \mathbb{R}$ be such that f is differentiable with respect to z and θ, with derivatives of polynomial growth in z uniformly in θ. Then:*

$$\frac{1}{N}\sum_{i=1}^{N} f(Z_i,\theta) \xrightarrow{\mathbb{P}_{\theta_0}} \int f(z,\theta)\nu_0(dz) \text{ as } N\to\infty \text{ uniformly in } \theta.$$

Lemma is proven in [Kessler, 1997] for one-dimensional case, however, as its proof is based only on ergodicity of the process and the assumptions we set in the beginning of the chapter, and not on the discretization scheme or dimensionality, we take it for granted without giving a formal generalization for a multi-dimensional case.

4

Discretization scheme

The main problem of the parametric inference in stochastic diffusions is that in most cases, explicit expression for the transition density governed by (3.1) cannot be found. Thus, we need to find a way to approximate it. Further, as we are working with a hypoelliptic diffusions, the "true" covariance matrix of the process is degenerate. In order to overcome this problem we want to develop a scheme which allow to propagate the noise to the first coordinate. For that purpose, we will use a modified local linearization scheme which was originally proposed by Ozaki for system (1.2) (see [Ozaki, 1989, Ozaki, 2012]).

4.1 Local linearization scheme

4.1.1 Derivation of the scheme

We start by introducing a discretization scheme which allows to approximate a transition density. Our derivation somewhat differs from that given in [Ozaki, 2012] and is more closely related to Chapter 5.6 in [Karatzas and Shreve, 1987]. We start with considering a particular case of system (3.1), when the variance is constant:

$$dZ_t = A(Z_t)dt + \tilde{\sigma}dW_t, \quad Z_0 = \omega_0,\ t \in [0, T] \tag{4.1}$$

where $Z_t = (X_t, Y_t)^T$, W_t is a standard one-dimensional Brownian motion, $\tilde{\sigma} = (0, \sigma)^T$ and ω_0 is $\mathcal{F}_0$-measurable $2-$dimensional random vector. Now we assume that on some sufficiently small interval $(\tau, \tau + \Delta]$ we can approximate the solution of the original system by the solution of a linear stochastic differential equation. It is the case when the Jacobian $J(Z_t; \theta) \triangleq J_t$ of the drift coefficient $A(Z_t; \theta)$ is constant on each small interval of length Δ.

In other words, instead of working with (4.1), we study system of the form:

$$d\mathcal{Z}_s = J_\tau \mathcal{Z}_s ds + \sigma d\eta_s, \quad \mathcal{Z}_0 = Z_\tau,\ s \in (\tau, \tau + \Delta] \tag{4.2}$$

where J_τ depends solely on the value of the process Z_t at the moment τ, and $\eta_t = (0, W_t)$ is an original vector of noise.

We may give the solution of (4.2) in the explicit form (recall Example 2.1):

$$\mathcal{Z}_s = Z_\tau e^{J_\tau \Delta} + \sigma \int_\tau^{\tau+\Delta} e^{J_\tau(\tau+\Delta-\delta)} d\eta_\delta \tag{4.3}$$

Then the first moment of the process $\mathcal{Z}_s$ on each interval $(\tau, \tau+\Delta]$ is:

$$\mathbb{E}[\mathcal{Z}_s|\mathcal{F}_\tau] = Z_\tau e^{J_s \Delta}. \tag{4.4}$$

Respectively, the covariance matrix of process $\mathcal{Z}_s$ is defined in the following way:

$$\Sigma_{\mathcal{Z}_s} = \sigma^2 \mathbb{E}\left[\left(\int_\tau^{\tau+\Delta} e^{J_\tau(\tau+\Delta-\delta)} d\eta_\delta\right)\left(\int_\tau^{\tau+\Delta} e^{J_\tau(\tau+\Delta-\delta)} d\eta_\delta\right)^T\right]. \tag{4.5}$$

To keep this representation more comprehensible, we propose to consider its approximated form. Thus, we have the following proposition, proof of which we postpone to appendix:

Proposition 4.1. *Second-order Taylor approximation of matrix $\Sigma_{\mathcal{Z}_s}$ defined in* (4.5) *has the following form:*

$$\Sigma_{\mathcal{Z}_s} = \sigma^2 \begin{pmatrix} (\partial_y a_1)^2 \frac{\Delta^3}{3} & (\partial_y a_1)^2 \frac{\Delta^2}{2} + (\partial_y a_1)(\partial_y a_2)\frac{\Delta^3}{3} \\ (\partial_y a_1)^2 \frac{\Delta^2}{2} + (\partial_y a_1)(\partial_y a_2)\frac{\Delta^3}{3} & \Delta + (\partial_y a_2)\frac{\Delta^2}{2} + (\partial_y a_2)^2 \frac{\Delta^3}{3} \end{pmatrix} + \mathcal{O}(\Delta^4),$$

where $\partial_y a_1 \equiv \frac{\partial}{\partial y} a_1(Z_\tau;\theta)$ and $\partial_y a_2 \equiv \frac{\partial}{\partial y} a_1(Z_\tau;\theta)$.

Now it is easy to see that noise in the first coordinate appears only through $\partial_y a_1$, thus it is indeed necessary to keep this term not equal to zero. Under the hypoellipticity assumption (4.5) has rank 2, while the covariance matrix of original process is of rank 1. Thus, now we can find an inverse of the covariance matrix. Also note that the variance of the first and the second coordinate is of different order (Δ^3 and Δ respectively).

At this point we give up the continuous time setting and move to the discrete process, as we are only interested in the sampling of the process $\{Z_i\}$, $i \in 1, \dots N$. Considering the derivations above, and knowing the value of the process Z_i, we can approximate Z_{i+1} in the following way:

$$Z_{i+1} = \bar{A}(Z_i;\theta) + \Upsilon(Z_i;\theta), \tag{4.6}$$

where $\Upsilon = (\upsilon_1, \upsilon_2)^T$ is a bivariate normally distributed random vector with zero mean and the covariance matrix approximated by (4.5), and $\bar{A}$ being a conditional expectation of Z_{i+1} given Z_i which is defined in (4.4).

We can consider υ_1 as an integrated "original" Brownian motion $\upsilon_2 \triangleq W_t$, bringing this representation in accordance with [Kloeden et al., 2003] (Chapter 4, Strong Approximations). In other words, υ_1 can be defined as:

$$\upsilon_1 = \int_\tau^{\tau+\Delta} \left(\int_\tau^s d\upsilon_2\right) ds$$

Remark. *Note that when the model is already linear (as in Example 2.1), then the proposed approximation corresponds to the true transition density of the process. Also note that throughout the derivation we did not use much the dimensionality of the process Z_t, thus this approach can be easily extended to higher dimensions.*

Numerical implementation

Note that the expressions for drift and variance term that we have introduced in (4.6) are not straightforwardly implementable in some programming languages, as they involve calculation of the matrix exponent and integrals. Thus we expand each term with Taylor series and further throughout this work will consider only an approximated version of (4.4) and (4.5).

Expansion of (4.4) immediately follows from the definition of a matrix exponent. Also, taking into account our initial assumption that $A(Z_t;\theta) \approx J(Z_t;\theta)Z_t$ (and denoting $J(Z_i;\theta) \triangleq J_i$), we get the following:

$$\bar{A}(Z_i;\theta) = Z_i + \Delta A(Z_i;\theta) + \frac{\Delta^2}{2} J_i A(Z_i;\theta) + \cdots + \frac{\Delta^k}{k!} J_i^k A(Z_i;\theta) + O(\Delta^{k+1}) \tag{4.7}$$

What about the variance, it is sufficient to use only the higher-order terms present in (4.5). Thus, we define a matrix Σ_Δ which will be used most of the time instead of the full form of (4.5):

$$\Sigma_\Delta(Z_i;\theta,\sigma) \triangleq \sigma^2 \begin{pmatrix} (\partial_{Y_i} a_1)^2 \frac{\Delta^3}{3} & \partial_{Y_i} a_1 \frac{\Delta^2}{2} \\ \partial_{Y_i} a_1 \frac{\Delta^2}{2} & \Delta \end{pmatrix} \tag{4.8}$$

Remark. *Note that if we neglect the terms of order Δ^2 and higher in* (4.7) *and* (4.8)*, approximation with the linearization scheme completely coincides with Euler-Maruyama approximation* (2.6)*. If we leave out only the terms of order Δ^3 and higher, drift approximation is equivalent to including first two terms of an infinitesimal generator defined in* (2.3)*.*

Further note that when $a_1(X_t, Y_t;\theta) \equiv Y_t$ (and, respectively, $\partial_y a_1 \equiv 1$) representation in form (4.8) *corresponds to the covariance matrix for stochastic damping Hamiltonian system obtained independently for different schemes in [Ditlevsen and Samson, 2017, Pokern et al., 2007]. More details are provided in Chapter 5.*

Now we face an important problem — the matrix (4.8) now depends on the parameters from the first equation. Also we should take into account that the matrix (4.5) is still highly ill-conditioned for $0 < \Delta \ll 1$, as $\det \Sigma_\Delta \approx \sigma^4 (\partial_y a_1)^2 \frac{\Delta^4}{12} \approx \mathcal{O}(\Delta^4)$. As will be explained further in Chapter 6, it causes difficulties in estimating the parameters.

Remark. *It is possible to obtain an explicit expression for each element of matrix Σ_Δ in terms of eigenvalues of matrix J_t, but we omit it here as we are more interested in statistical properties of this representation. However, it can be found in [Ozaki, 2012].*

Also note that from a computational point of view, representation (4.6) is not really convenient. Indeed, in order to simulate data with this scheme, we would have to generate correlated random variables, while it is more natural to use uncorrelated ones. We can easily express $\Upsilon(Z_i;\theta)$ in terms of uncorrelated random variables as $\bar{B}(Z_i;\theta)\Xi_i$, where $\Xi_i = (\xi_1(i), \xi_2(i))^T$ — bivariate standard Gaussian noise with covariance matrix $\sigma^2 I$ and

$\bar{B}(Z_i;\theta)$ is some real-valued matrix such that $\bar{B}\bar{B}^T = \frac{1}{\sigma^2}\Sigma_\Delta \triangleq \bar{\Sigma}$. Then the approximation for the solution (3.1) becomes:

$$Z_{i+1} = \bar{A}(Z_i;\theta) + \bar{B}(Z_i;\theta)\Xi_i \tag{4.9}$$

Then we note that the discretized process is Gaussian on each small interval:

$$\mathbb{E}\left[Z_{i+1}|Z_i\right] \sim \mathcal{N}\left(\bar{A}(Z_i;\theta), \Sigma_\Delta(Z_i;\theta)\right)$$

and its transition density is given by:

$$p_\Delta(Z_{i+1}|Z_i;\theta,\sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}}|\det\Sigma_\Delta|}\times \\ \times \exp\left[-\frac{1}{2\sigma^2}[Z_{i+1}-\bar{A}(Z_i;\theta)]^T{\Sigma_\Delta}^{-1}[Z_{i+1}-\bar{A}(Z_i;\theta)]\right], \tag{4.10}$$

or, equivalently, in terms of $\bar{B}$:

$$p_\Delta(Z_{i+1}|Z_i;\theta,\sigma^2) = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma^2|\det\bar{B}|^2}\times\exp\left[-\frac{1}{2\sigma^2}\left\|\bar{B}^{-1}[Z_{i+1}-\bar{A}(Z_i;\theta)]\right\|^2\right], \tag{4.11}$$

Remark. *The equivalence of both expressions may not be instantly recognizable, however it straightforwardly follows from matrix properties. Indeed, given matrices C and D such that product CD exists, it is easy to note that:*

$$C^T(DD^T)^{-1}C = C^T(D^T)^{-1}D^{-1}C = (D^{-1}C)^TD^{-1}C = \|D^{-1}C\|^2$$

Now let us discuss the choice of $\bar{B}(Z_i;\theta)$. Though an estimation based on this scheme does not depend of the representation we choose, we still want to make this part explicit, as it will be easier to explain the similarity between different schemes in Chapter 5.

First note that Σ_Δ is positive definite. Then we can use one of the following approaches:

1. **Cholesky decomposition.** This method is the most computationally straightforward, and is already pre-implemented in the most popular programming languages (`chol` in **R**, `cholesky` from library `numpy` in **Python**, etc). In this case $\bar{B}(Z_i;\theta)$ is a lower-triangular matrix. Such matrix always exists and is unique.

2. **Spectral decomposition.** In this case we use the matrix $\bar{B} = U\Lambda^{\frac{1}{2}}$ obtained from a spectral decomposition $\Sigma_\Delta = U\Lambda U^T$ of Σ_Δ, where Λ is a diagonal matrix with eigenvalues of Σ_Δ as entries, and each column of U is a corresponding eigenvector. Spectral decomposition is used in original work devoted to local linearization scheme for Hamiltonian system [Ozaki, 1989].

3. **"Inverse" Cholesky decomposition.** Though we do not refer to a "classical" method of linear algebra here, we still note that it is also possible to find a matrix similar to that in Cholesky decomposition with that difference that $\bar{B}(Z_i;\theta)$ is upper-triangular. In particular, this approach is used in [Pokern et al., 2007]. Further advantage of using this scheme is that after decomposing (4.8), we do not introduce any additional sources of noise nor change the coefficients in the second equation, but add Gaussian noise of order $\Delta^{\frac{3}{2}}$ to the first coordinate only.

4.1.2 Properties of the scheme

Before we proceed, we would like to verify whether the scheme is convergent. Throughout this section we assume that condition (A2) holds. First we need the following result (which is an almost straightforward consequence of Problem 3.15 in [Karatzas and Shreve, 1987]):

Proposition 4.2. *Let Z_t be a solution of (3.1) with an initial value ξ such that $\mathbb{E}\|\xi_0\|^2 < \infty$, then $\forall \alpha \in (0,1)$, $\exists C$ such that the following holds:*

$$\mathbb{E}\left[\int_0^{\Delta} \|Z_s - Z_{\alpha\Delta}\|^2 ds\right] \leq C(1+\mathbb{E}\|\xi_0\|^2)\frac{\Delta^2}{2}$$

Then the following proposition holds:

Proposition 4.3. *Denote by Z_t continuous solution of the system (3.1), and by $Z_i, \quad i \in 1,\dots,N, \quad N\Delta = T$ its discrete time approximation defined in (4.6). Then:*

$$\mathbb{E}\,\|Z_T - Z_N\| \leq C\Delta^{\frac{3}{2}}$$

Now we are interested in statistical properties of the suggested discretization scheme. First we note that the behaviour of each coordinate differs because of the variance of noise. Thus we should study each coordinate of (4.7) separately. Let us denote by $\bar{A}_1(Z_i;\theta)$ the first element of the vector $\bar{A}(Z_i;\theta)$, and by $\bar{A}_2(Z_i;\theta)$ — the second. Let us further assume that the drift term is approximated up to k-th term of (4.7), and that $k \geq 3$, as otherwise the performance of the method will be poor (see Chapter 6).

Proposition 4.4 (Weak convergence of the LL scheme)**.** *Following holds:*

$$\begin{aligned}
\mathbb{E}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right) &= \mathcal{O}(\Delta^k)\\
\mathbb{E}\left(Y_{i+1} - \bar{A}_2(Z_i;\theta)\right) &= \mathcal{O}(\Delta^k)\\
\mathbb{E}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right)^2 &= \left(\frac{\partial a_1}{\partial y}\right)^2 \frac{\Delta^3}{3}\sigma^2 + \mathcal{O}(\Delta^4)\\
\mathbb{E}\left(Y_{i+1} - \bar{A}_2(Z_i;\theta)\right)^2 &= \Delta\sigma^2 + \mathcal{O}(\Delta^2)
\end{aligned}$$

These bounds in combination with the continuous ergodic theorem and Lemma 3.1 [Kessler, 1997] allow us to establish the following important result:

Lemma 4.1. *Let $f : \mathbb{R}^2 : \Theta \to \mathbb{R}$ be a function with derivatives of polynomial growth in x, uniformly in θ. Then:*

1. *Assume $\Delta_n \to 0$ and $n\Delta_n \to \infty$.*

$$\frac{1}{n\Delta_n^3}\sum_{i=0}^{n-1} \frac{f(Z_i;\theta)}{(\partial_{Y_i} a_1)^2}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right)^2 \xrightarrow{P_\Theta} \frac{\sigma_0^2}{3}\int f(z;\theta)\nu_0(dz)$$

2. Assume $\Delta_n \to 0$ *and* $n\Delta_n \to \infty$.

$$\frac{1}{n\Delta_n}\sum_{i=0}^{n-1} f(Z_i;\theta)\left(Y_{i+1} - \bar{A}_2(Z_i;\theta)\right)^2 \xrightarrow{P_\Theta} \sigma_0^2 \int f(z;\theta)\nu_0(dz)$$

3. Assume $\Delta_n \to 0$ *and* $n\Delta_n \to \infty$.

$$\frac{1}{n\Delta_n^2}\sum_{i=0}^{n-1} \frac{f(Z_i;\theta)}{\partial_{Y_i} a_1}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right)\left(Y_{i+1} - \bar{A}_2(Z_i;\theta)\right) \xrightarrow{P_\Theta} \frac{\sigma_0^2}{2}\int f(z;\theta)\nu_0(dz)$$

Proof is postponed to appendix as well.

4.2 Parametric inference

Now we can finally introduce a method to estimate the parameters. We adopt a classical approach, where the first step consists in computing a likelihood of the discretized process (4.9) (or pseudo-likelihood of a continuous model (3.1)). Then we maximize it with respect to the unknown parameters.

As we have reduced our problem down to considering the sequence of normally distributed random variables, discrete version of a complete likelihood or the process $\{Z_i\}$ has the following form:

$$p_\Delta(Z_{0:N};\theta,\sigma^2) = p(Z_0;\theta,\sigma^2)\prod_{i=0}^{N-1} p_\Delta(Z_{i+1}|Z_i;\theta,\sigma^2), \tag{4.12}$$

where $p_\Delta(Z_{i+1}|Z_i;\theta,\sigma^2)$ is defined in (4.10).

Then we introduce a contrast function, which we can intuitively define as -2 times pseudo-likelihood:

$$\mathcal{L}_\Delta(\theta,\sigma^2;Z_{0:N}) = \sum_{i=0}^{N-1}(Z_{i+1} - \bar{A}(Z_i;\theta))^T\Sigma_\Delta^{-1}(Z_i;\theta,\sigma^2)(Z_{i+1} - \bar{A}(Z_i;\theta)) + \sum_{i=0}^{N-1}\log\det(\Sigma_\Delta(Z_i;\theta,\sigma^2)). \tag{4.13}$$

However, we need to be careful with this form, as it leads to the biased estimation of the variance coefficient. And let us explain why.

First, consider ordinary elliptic system with two sources of not necessarily independent Gaussian noise and the same constant variance coefficient in both equations. Then assume we have approximated it with the sequence of normally distributed variables (using, for example, Euler scheme) with mean vector $(\mu_1,\mu_2)^T$ and the covariance matrix given by:

$$\Sigma_{elliptic} = \sigma^2\begin{pmatrix}\Delta & C_W\\ C_W & \Delta\end{pmatrix},$$

where C_W is a correlation coefficient, which may also depend on the time step. For us it is only important to know that $\det \Sigma_{elliptic} = \Delta^2 - C_W^2 > 0$. Then the inverse matrix $\Sigma^{-1}_{elliptic}$ is given as:

$$\Sigma^{-1}_{elliptic} = \frac{1}{\sigma^2 \det \Sigma_{elliptic}} \begin{pmatrix} \Delta & -C_W \\ -C_W & \Delta \end{pmatrix},$$

Then, if we compute the expectation of the first part of the contrast by formula 4.13 for an elliptic system, we will get the following:

$$\mathbb{E}\left[\frac{1}{\sigma^2 \det \Sigma_{elliptic}} \sum_{i=0}^{N-1} \begin{pmatrix} \mu_1 & \mu_2 \end{pmatrix} \begin{pmatrix} \Delta & -C_W \\ -C_W & \Delta \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right] =$$
$$= \frac{1}{\sigma^2 \det \Sigma_{elliptic}} \sum_{i=0}^{N-1} \mathbb{E}\left[\Delta\mu_1^2 - 2C_W\mu_1\mu_2 + \Delta\mu_2^2\right] = \frac{N}{\sigma^2 \det \Sigma_{elliptic}} \left[2\Delta^2 - 2C_W^2\right] = \frac{2N}{\sigma^2}$$

However, if we do the same with a covariance matrix defined in 4.8, we end up with a bias. Indeed, note that the matrix Σ_Δ^{-1} is of the following form:

$$\Sigma_\Delta^{-1} = \frac{1}{\sigma^2} \begin{pmatrix} 12 \left(\partial_{Y_i} a_1\right)^{-2} \Delta^{-3} & -6 \left(\partial_{Y_i} a_1\right)^{-1} \Delta^{-2} \\ -6 \left(\partial_{Y_i} a_1\right)^{-1} \Delta^{-2} & 4\Delta^{-1} \end{pmatrix} \tag{4.14}$$

Then, recalling Proposition 4.3 we have:

$$\mathbb{E}\left[\frac{1}{\sigma^2} \sum_{i=0}^{N-1} \left(\frac{12}{(\partial_{Y_i} a_1)^2 \Delta^3}\mu_1^2 - \frac{12}{(\partial_{Y_i} a_1)\Delta^2}\mu_1\mu_2 + \frac{4}{\Delta}\mu_2^2\right)\right] = \frac{4N}{\sigma^2}$$

Thus, our contrast must be corrected with respect to this fact by dividing the first part by 2. Note that we do not introduce any additional constants in the second part because of the logarithm:

$$\mathcal{L}(\theta, \sigma^2; Z_{0:N}) = \frac{1}{2} \sum_{i=0}^{N-1} (Z_{i+1} - \bar{A}(Z_i;\theta))^T \Sigma_\Delta^{-1}(Z_i;\theta,\sigma^2)(Z_{i+1} - \bar{A}(Z_i;\theta))$$
$$+ \sum_{i=0}^{N-1} \log\det(\Sigma_\Delta(Z_i;\theta,\sigma^2)). \tag{4.15}$$

Finally, the estimator is defined as:

$$(\hat{\theta}, \hat{\sigma}^2) = \arg\min_{\theta,\sigma^2} \mathcal{L}(\theta, \sigma^2; Z_{0:N}) \tag{4.16}$$

This criteria is difficult to study because of the different order of variance for the first and the second coordinate. Another problem consists in the approximated matrix (4.8), which now also depends on the parameters present in the first coordinate. Finally, we have certain numerical difficulties, which will be discussed in Chapter 6. But though we will

verify its performance experimentally in Chapter 6, we also want to learn the properties of the proposed estimation from mathematical point of view.

To begin with, note that the approximation of the likelihood coincides up to order $\mathcal{O}(\Delta^4)$ with an approximation introduced in [Ditlevsen and Samson, 2017] (more details can be found in Chapter 5: 1.5 strong order scheme). When smooth and rough variables are described by equations which depends on different sets of parameters (which is often the case — see, for example, FitzHugh-Nagumo model in Chapter 6), instead of considering the "complete" discretized log-likelihood we can work with separate contrasts for each coordinate and it is proven that these estimators are consistent.

On the contrary, we will stick to a 2-dimensional criterion, but following the idea from the above mentioned work we will consider each variable separately. Let us denote the parameters in the first coordinate by φ, and in the second — by ψ.

Theorem 4.1. *Under assumptions (A1)-(A4) and $\Delta_N \to 0$ and $N\Delta_N \to \infty$ the following holds:*

$$\hat{\sigma}^2_{N,\Delta_N} \xrightarrow{\mathbb{P}_\theta} \sigma_0^2$$
$$\hat{\varphi}_{N,\Delta_N} \xrightarrow{\mathbb{P}_\theta} \varphi_0$$
$$\hat{\psi}_{N,\Delta_N} \xrightarrow{\mathbb{P}_\theta} \psi_0$$

Statement of the theorem essentially follows from three lemmas, which are postponed to appendix, as well as the proof of the theorem itself.

Also note that in a particular case of the system (3.1), that is (1.2), we can split estimation for the drift and the variance parameters. We present them here in order to visualize the link between our approach and some results for system (1.2), which were available before. In original work [Ozaki, 1989] separate estimators for the drift and the variance term are introduced, as expression for $\hat{\sigma}^2$ can be found in explicit form. Discrete likelihood of the model is expressed in terms as in (4.11). More precisely, estimators are defined as follows:

$$\hat{\theta}_{N,\Delta} = \arg\min_\theta \frac{1}{2(N-1)} \sum_{i=0}^{N-1} \|\bar{B}^{-1}(Z_i;\theta)[Z_{i+1} - \bar{A}(Z_i;\theta)]\|^2 \tag{4.17}$$

$$\hat{\sigma}^2_{N,\Delta} = \frac{1}{2(N-1)} \sum_{i=0}^{N-1} \|\bar{B}^{-1}(Z_i;\theta)[Z_{i+1} - \bar{A}(Z_i;\theta)]\|^2, \tag{4.18}$$

where $\bar{B}(Z_i;\theta)$ is a spectral decomposition of $\bar{\Sigma}$.

Remark. *Explicit expression for the variance term is obtained by taking the derivative of* (4.16) *with respect to σ^2 and finding a point in which it is equal to zero. We can do it, as the expression for discretized likelihood is convex with respect to σ^2.*

5

Link to the other schemes

As it was already mentioned in the introduction, till recently no estimation techniques for general models of type (3.1) were available. Thus, we begin with considering systems of type (1.2). In particular, we want to compare our approach with the estimation technique suggested by Pokern [Pokern et al., 2007] and with the Euler contrast which has been proposed in [Samson and Thieullen, 2012]. The first method was developed to work with multidimensional hypoelliptic equations with constant variance, though it was numerically tested exclusively on the models of type (1.2). Second method was initially designed for models (1.2), but also allowed a non-constant variance term.

But as the explicit analysis of the properties and limitations of each scheme would not suit in this thesis, we restrict ourselves to the two-dimensional case with the constant variance, as it would still give us a plausible overview of the existing techniques.

Here we consider only the schemes which are based on maximizing the pseudo log-likelihood of the model, though there also exist few works devoted to a non-parametric estimation [Cattiaux et al., 2014], [Cattiaux et al., 2016].

5.1 Euler contrast for stochastic damping Hamiltonian system

We start the comparison with the most simple scheme, which is based on the Euler-Maruyama scheme for models of type (1.2). Main reference is [Samson and Thieullen, 2012]. Auxiliary model for this scheme completely coincides with the standard Euler-Maruyama approximation of form (2.6).

Parameter estimation is based on the second equation of the discretized model (2.6). The following contrast function is introduced:

$$\mathcal{L}_E(\theta, \sigma^2; Z_i) = \sum_{i=0}^{N-1} \left(\frac{\left(Y_{i+1} - Y_i - \Delta a_2(X_i, Y_i; \theta)\right)^2}{\Delta \sigma^2} + \log \sigma^2 \right).$$

Then the optimal value of parameters can be found as:

$$(\hat{\theta}_E, \hat{\sigma}^2_E) = \arg\min_{\theta,\sigma^2} \mathcal{L}_E(\theta, \sigma^2; Z_i). \tag{5.1}$$

Note that the value of σ^2 may be found explicitely by taking the derivative (as $\mathcal{L}_E(\theta, \sigma)$ is convex):

$$\hat{\sigma}^2_E = \frac{1}{\Delta(N-1)} \sum_{i=0}^{N-1} \left(Y_{i+1} - Y_i - \Delta a_2(X_i, Y_i; \theta)\right)^2.$$

The main difference of this approach with the estimator of form (4.16) is that rather than evaluating the norm of the full vector of observations it minimizes the difference between all successive terms presented in the discretized model of the second coordinate only. It is proven in [Samson and Thieullen, 2012] that the estimator is consistent both in the cases of fully and partially observed data.

5.2 Modified 1.0 scheme (Pokern et al.)

Now we consider a more advanced scheme, suggested in [Pokern et al., 2007]. To begin with, let us introduce an auxiliary model, which serves as a link between the observations sampled from the continuous solution of system (1.2) and the discrete-time model.

The drift term is approximated in exactly the same way as in (2.6), that is:

$$\bar{A}_P(Z_i; \theta) = \begin{bmatrix} X_i + \Delta Y_i \\ Y_i + \Delta a_2(X_i, Y_i; \theta) \end{bmatrix}$$

Note that the approximation of the drift is the same as in (2.6), thus $\bar{A}_P \equiv \bar{A}_{1.0}$. However, approximation of the variance term $\bar{B}_P \xi$ differs from that in the scheme (2.6), because it includes terms of order $\Delta^{\frac{3}{2}}$ in the first coordinate. More precisely, the matrix $\bar{B}_P$ has the following form:

$$\bar{B}_P = \sigma \begin{pmatrix} \frac{1}{12}\Delta^{\frac{3}{2}} & \frac{1}{2}\Delta^{\frac{3}{2}} \\ 0 & \sqrt{\Delta} \end{pmatrix}$$

Note that the covariance matrix $\bar{B}_P \bar{B}_P^T$ has exactly the same form as (4.8), and the matrix $\bar{B}_P$ corresponds to "inverse" Cholesky decomposition of Σ_Δ defined in Chapter 4.

Estimation for the parameters of the drift term and the variance is conducted separately by the following estimators:

$$\hat{\theta}_P = \arg\min_{\theta} \frac{1}{2(N-1)} \sum_{i=0}^{N-1} \|\bar{B}_P^{-1}(Z_i - \bar{A}_P(Z_i; \theta))\|^2,$$

$$\hat{\sigma}^2_P = \frac{1}{2(N-1)} \sum_{i=0}^{N-1} \|\bar{B}_P^{-1}(Z_i - \bar{A}_P(Z_i; \theta))\|^2.$$

Note that both expressions are equivalent to the estimation proposed in [Ozaki, 1989] for the system of type (1.2) with a constant variance (here: (4.17) and (4.18) respectively), with that difference that $\bar{A}_P$ includes terms only up to Δ, while $\bar{A}$ can be expanded up to any term we consider plausible.

Taking into consideration the numerical performance of the method even for the most simple models (Section 5 in [Pokern et al., 2007]), it becomes obvious that suggested scheme does not ensure an accurate estimation of the parameters. Main problem is that while the drift is approximated only up to the order Δ, the variance term includes terms up to Δ^3.

Further, it is noticed that in the case of multidimensional systems with n smooth and m rough variables, only the estimation for the rough variables works good, as the errors accumulated in the estimation of the parameters in the smooth variables are not neglectable. In order to estimate the parameters of smooth equations, it is better to choose another approach.

5.3 1.5 strong order Euler scheme

Parameter estimation proposed in [Ditlevsen and Samson, 2017] is based on 1.5 strong order Euler scheme (see Chapter 4 in [Kloeden et al., 2003] for reference). It is designed for multidimensional hypoelliptic systems, when one of the variables is smooth, and all the others — rough. However, here we only consider two-dimensional systems with a constant variance.

We want to show that the drift approximation in 1.5 strong order scheme corresponds to that in the local linearization scheme up to the terms of order $\mathcal{O}(\Delta^3)$. In 1.5 order strong scheme drift term is approximated in the following way:

$$\bar{A}_{1.5}(Z_i;\theta) = \begin{pmatrix} X_i + \Delta a_1(X_i,Y_i;\theta) + \frac{\Delta^2}{2}\left[\frac{\partial a_1}{\partial X_i}a_1 + \frac{\partial a_1}{\partial Y_i}a_2\right] \\ Y_i + \Delta a_2(X_i,Y_i;\theta) + \frac{\Delta^2}{2}\left[\frac{\partial a_2}{\partial X_i}a_1 + \frac{\partial a_2}{\partial Y_i}a_2\right] \end{pmatrix} + \mathcal{O}(\Delta^3). \tag{5.2}$$

Note that (5.2) coincides with (4.4) up to $\mathcal{O}(\Delta^3)$. Approximation of the covariance matrix $\bar{\Sigma}_{1.5}$ also coincides with (4.8) up to $\mathcal{O}(\Delta^3)$ (see [Ditlevsen and Samson, 2017]). $\bar{B}_{1.5}(Z_i;\theta)$ was not specified in terms of uncorrelated random variables.

The approximated log-likelihood for this method coincides with (4.15), and an estimation contrast is defined in exactly the same way, as in (4.16). Thus both linearization scheme and 1.5 strong order scheme are expected to give the same performance.

Now we want to point out some statistical properties of this estimator. As it is difficult to study the general criterion, it is suggested to split the estimation for the parameters included in smooth and rough coordinates. Main idea of this approach is that we can treat each variable X and Y as a sequence of normally distributed random variables with means $A_1(Z_i;\theta)$ and $A_2(Z_i;\theta)$ and variances $\sigma^2\frac{\Delta^3}{3}(\partial_{Y_i}a_1)^2$ and $\sigma^2\Delta$ respectively. Further, let us assume that each variable depends on the different set of parameters θ_1 and θ_2.

Then, recalling (2.7), the estimator for each group of parameters is defined in the following way:

$$\hat{\theta}_{1,N} = \arg\min_{\theta_1} \left(\frac{3}{\Delta^3} \sum_{i=0}^{N-1} \frac{\left(X_{i+1} - \bar{A}_{(1),1.5}(Z_i;\theta_1,\theta_2)\right)^2}{\left(\partial_{Y_i} a_1\right)^2 \sigma^2} + \sum_{i=0}^{N-1} \log((\partial_{Y_i} a_1)^2 \sigma^4) \right)$$

$$(\hat{\theta}_{2,N}, \hat{\sigma}^2_N) = \arg\min_{\theta_2,\sigma^2} \left(\sum_{i=0}^{N-1} \frac{\left(Y_{i+1} - \bar{A}_{(2),1.5}(Z_i;\theta_1,\theta_2)\right)^2}{\Delta\sigma^2} + \sum_{i=0}^{N-1} \log \sigma^4 \right),$$

where $\bar{A}_{(1),1.5}$ and $\bar{A}_{(2),1.5}$ denote first and the second coordinate in the vector $\bar{A}_{1.5}$ respectively. For theoretical properties of this method see [Ditlevsen and Samson, 2017]. Proofs in general follow the work of [Kessler, 1997].

Remark. *Note, that if we consider system* (1.2) *and leave out the terms of order* Δ^2 *in the approximation of drift* (5.2)*, then the estimator* $(\hat{\theta}_2, \hat{\sigma})^2$ *coincides with* (5.1)*.*

It was proven that the estimator is consistent under the following conditions:

1. If $\Delta \to 0,\ N\Delta \to \infty$, then
$$\hat{\theta}_{1,N} \xrightarrow{P} \theta_{1,0}$$
2. If $\Delta \to 0,\ N\Delta \to \infty$, then
$$(\hat{\theta}_{2,N}, \hat{\sigma}^2_N) \xrightarrow{P} (\theta_{2,0},\ \sigma^2_0),$$
where $\theta_{1,0}$ and $\theta_{2,0}$ are real values of the parameters.

It was also proven that the estimator for θ_2 and σ^2 is asymptotically normal under the condition that $N\Delta^2 \to 0$ and $N\Delta \to \infty$.

5.4 Summary

As a conclusion we want to highlight an important moment about the considered schemes. Note that in the case with Euler contrast and 1.5 strong order scheme estimator it is shown that the estimation can be conducted for smooth and rough variables separately when the variables of different type are driven by different sets of parameters. It also allows to vastly reduce the computational time, but it can also result in decreasing accuracy.

Number of early works (including, in particular, [Ozaki, 1989, Pokern et al., 2007]) did not investigated conditions under which the estimator is consistent and asymptotically normal, as the main focus of the study was on the numerical performance of the scheme. For us, however, this question is of great importance as it can give us some insight into restrictions of the scheme. It is shown that the Euler contrast is consistent and asymptotically normal under conditions $N\Delta \to \infty$ and $N\Delta^2 \to 0$. The same holds for the estimator

for the rough coordinate in 1.5 strong order scheme. What about the smooth coordinate — despite the fact that consistency holds for the introduced estimator, asymptotic normality remains under question.

In prospective, we can build more stable schemes by approximating the solution by the infinitesimal generator of the transition density with higher-order terms (see [Kloeden et al., 2003]). However, it will not necessarily result in a better performance when working with real data, as all the models are not accurate. In other words, adding more terms to the approximation will give better results when we know for sure that the underlying model is the one we consider in approximation.

6

Simulation study

In this part we test the numerical performance of the proposed method. Throughout the chapter we stick to using **R**. Good overview of methods of stochastic modelling with implementations in **R** can be found in [Iacus, 2009].

Data is generated with the Local Linearization scheme, that is:

$$Z_{i+1} = \bar{A}(Z_i; \theta_0) + B(Z_i; \theta_0)\Xi_i,$$

where Ξ is a bivariate Gaussian noise with covariance matrix $\sigma^2 I$ and zero mean, and matrices $A(Z_i; \theta_0)$ and $B(Z_i; \theta_0)$ defined in Chapter 4 (we have used Cholesky decomposition of (4.8)). We generated increments of the Brownian motion using standard function `rnorm` in **R**.

The trials are organized as follows: for each trial we randomly generate 100 trajectories of length 1000000 and $\Delta = 0.001$ using the LL scheme of order $O(\Delta^3)$, unless other specified. Then we draw diffent number of samples with different step size, which will be specified further. We use subsampling keeping in mind that we are not working with an original process, but with its discrete-time approximation. If we estimate the parameters with exactly the same time step with which the data was generated, we may obtain significantly better results, as in that case the appoximated log-likelihood corresponds, in fact, to the real log-likelihood of the generated discrete model. In order to minimize this effect, we use subsamples.

6.1 FitzHugh-Nagumo model

We consider stochastic hypoelliptic FitzHugh-Nagumo model [Fitzhugh, 1961]. It is a simplified version of the Hodgkin-Huxley model [Hodgkin and Huxley, 1952], which models in a detailed manner activation and deactivation dynamics of a spiking neuron. First it was studied in the deterministic case (that is without a diffusion coefficient), then it was improved by adding two sources of noise to the both coordinates resulting in elliptic SDE. Parametric inference for elliptic FitzHugh-Nagumo model both in fully- and partially

observed case is investigated in [Jensen, 2014]. Here we consider a hypoelliptic version with noise only in the second coordinate.

The behaviour of the neuron is defined through the solution of the system

$$\begin{cases} dV_t = \frac{1}{\varepsilon}(V_t - V_t^3 - U_t - s)dt \\ dU_t = (\gamma V_t - U_t + \beta)dt + \sigma dW_t, \end{cases} \tag{6.1}$$

where the variable V_t represents the membrane potential of the neuron at time t, and U_t is a recovery variable, which could represent channel kinetic. Parameter s is the magnitude of the stimulus current and is known in experiment, ε is a time scale parameter, which is typically significantly smaller than 1, since V_t moves "faster" than U_t. Parameters to be estimated are $\theta = (\gamma, \beta, \varepsilon, \sigma)$.

The properties of this system (in particular, hypoellipticity and ergodicity) are studied in [Leon and Samson, 2017]. In [Jensen et al., 2012] it was proven that s is unidentifiable, when only one coordinate is observed, and it was also mentioned that the accuracy of the estimation technique heavily depends on the parameter ε. We should also note that ε causes the most problems in estimation, as it is included in higher-order derivative of $\partial_y a_1$. Moreover, in the simulation study devoted to the parametric estimation in a partially observed FitzHugh-Nagumo model [Ditlevsen and Samson, 2017], this parameter was fixed, as it breaks down their algorithm.

Depending on the coefficients of the model we may observe either oscillatory or excitatory behaviour. In particular, when β is relatively large, we may not observe spikes at all (in that case we say that the system has an *oscillatory* behaviour). When there is no clear pattern of spiking activity, data becomes too noisy and it is much harder to obtain a good estimation [Leon and Samson, 2017]. It is also noted in [Jensen et al., 2012] that the accuracy of the estimation especially depends on the parameter ε, which is a time scale (since both coordinates are moving in different time scales). In order to make our experiments more representative, we consider two different sets of parameters: with clear excitatory and oscillatory behaviour.

First, we calculate the Jacobian of the system:

$$J_t = \begin{pmatrix} \frac{1}{\varepsilon}(1 - 3V_t^2) & -\frac{1}{\varepsilon} \\ \gamma & -1 \end{pmatrix},$$

then the vector $\bar{A}(Z_i; \theta)$ is of the following form:

$$\bar{A}(Z_i;\theta) = Z_i + \begin{bmatrix} \frac{\Delta}{\varepsilon}(V_i - V_i^3 - U_i - s) + \frac{\Delta^2}{2\varepsilon}(\frac{1}{\varepsilon}(1 - 3V_i^2)(V_i - V_i^3 - U_i - s) - (\gamma V_i - U_i + \beta)) \\ \Delta(\gamma V_i - U_i + \beta) + \frac{\Delta^2}{2}(\frac{\gamma}{\varepsilon}(V_i - V_i^3 - U_i - s) - (\gamma V_i - U_i + \beta)) \end{bmatrix}.$$

And, finally, matrix Σ_Δ is

$$\Sigma_\Delta = \frac{1}{\sigma^2} \begin{pmatrix} \frac{\Delta^3}{3\varepsilon^2} & \frac{\Delta^2}{2\varepsilon} \\ \frac{\Delta^2}{2\varepsilon} & \Delta \end{pmatrix}.$$

Note that now parameter ε is included in the variance, and it directly "regulates" the amount of noise which propagates into the first equation.

In order to use the Euler contrast, we need to bring this system to a Hamiltonian form. For this purpose, we use an Itô formula and transform the system, setting $X_t := V_t$ and $Y_t := \frac{1}{\varepsilon}(V_t - V_t^3 - U_t - s)$ (here we change notations for both variable in order not to mix them with the original system).

Then (6.1) becomes:

$$\begin{cases} dX_t = Y_t dt \\ dY_t = \frac{1}{\varepsilon}\left(Y_t(1-\varepsilon-3X_t^2) - X_t(\gamma-1) - X_t^3 - (s+\beta)\right)dt - \frac{\sigma}{\varepsilon}dB_t, \end{cases} \tag{6.2}$$

However, we should note that in this transformation we have fixed the true value of the unknown parameter ε. Otherwise it would be impossible to estimate the real parameters of the system using the Euler contrast. Fixing the parameter we satisfy the assumption about the existence of an explicit transform of the model to Hamiltonian form, as it is required in [Samson and Thieullen, 2012].

We conduct our experiments with two different sets of parameters, representing the model with clearly excitatory and oscillatory behaviour. Behaviour of each variable and a phase portrait can be found on Figure 6.7.

Before we discuss the results, let us mention some problems that have arisen on the very first step of the experimental design. Initial idea was to construct a two-dimensional contrast using matrix-vector multiplication, as it would make for us easier to generalize the code for higher-dimensional systems. However, this approach includes computing of an inverse covariance matrix on each run, and it appeared to be problematic. Indeed, notice that Δ has to be sufficiently small in order to capture the behaviour of the complex non-linear system. On the other hand, when Δ is too small, estimation may break down because the matrix Σ_Δ is computationally singular. Indeed, in our case if $\Delta = 10^{-3}$ and $\sigma = 10^{-1}$, then $\det \Sigma_\Delta \approx 10^{-16}$. It already causes difficulties when we compute an inverse matrix (at least with standard methods like `solve` in **R**).

This problem can be avoided if we compute the expression of the contrast explicitly and then minimize the function without having to compute an inverse matrix on every iteration of the algorithm (even better approach, of course, consists in using separate estimators for the parameters in each variable). It would not be a particularly good solution for a system of higher dimension, but in our case it is helpful. In all experiments we used standard Conjugate Gradient method in **R**, as it proved to be more stable than, for example, Nelder-Mead, and minimize the following expression:

$$\mathcal{L}(\theta, \sigma^2; Z_{0:N}) = \sum_{i=0}^{N-1}\left[\frac{1}{\sigma^2}\left(\frac{3\varepsilon^2}{\Delta^3}(X_{i+1} - \bar{A}_1(Z_i;\theta))^2 - \frac{3\varepsilon}{\Delta^2}(X_{i+1} - \bar{A}_1(Z_i;\theta))(Y_{i+1} - \bar{A}_2(Z_i;\theta)) + \right.\right.$$
$$\left.\left. + \frac{1}{\Delta}(Y_{i+1} - \bar{A}_2(Z_i;\theta))^2\right) + \log\frac{\sigma^2}{\varepsilon}\right]. \tag{6.3}$$

6.1.1 ε is not fixed.

We start with ε not being fixed. Each table presented in the results corresponds to a different set of the real parameters: for excitatory see Table 6.1 and for oscillatory — 6.2 respectively. We see that estimation of ε has failed, and let us explain why.

Consider the very first term present in (6.3), that is:

$$\frac{12\varepsilon^2}{\Delta^3}(X_{i+1} - \bar{A}_1(Z_i;\theta))^2. \tag{6.4}$$

Recall that $\bar{A}_1(Z_i;\theta) = X_i + \frac{\Delta}{\varepsilon}(X_i - X_i^3 - Y_i - s) + \mathcal{O}(\Delta^2)$. Then when we try to calculate the expression (6.4), we end up with:

$$\frac{12}{\Delta^3}(\varepsilon(X_{i+1} - X_i) - \Delta(X_i - X_i^3 - Y_i - s))^2$$

Similar thing happens when we consider the second term. Notice that ε is no more present in the term which is responsible for the approximation of the drift. By intuition we can say that this parameter should be treated as an additional parameter of the variance term — but this is not completely accurate, as it is not present in the third term at all. In addition, it was noticed during the experiments that when ε was not fixed, σ became much more sensible to the initial value of $\hat{\sigma}^2$ with which the optimization routine was launched. There was no issues of that kind with parameters γ and β.

Brief explanation of why it is hard to estimate the parameters of the smooth variables correctly may also be found in [Pokern et al., 2007]. It was mentioned that usage of the standard likelihood leads to non-negligible errors in the first term and it should be taken into account. The problem disappears if the likelihood is based on the scheme which does not include the variance term. We will propose a possible solution of this problem in the next section, when we attempt an estimation for a stochastic approximation of the Hawkes process.

However, we also should notice that the estimation was not stable in general. Say, for 100 trials with different trajectories with the same properties, we have obtained more or less accurate results in 70% of cases, and got obviously non-plausible values in the other 30 %. Thus, in the table we included the median of the sample with the estimated parameters, as the mean value was not particularly representable.

Also notice that the estimation in the case of clearly excitatory behaviour (Table 6.1) is much more accurate that for the second set of parameters (Table 6.2). It is logical, as if we consider the plots of the same system driven by different sets of parameters (see Figure 6.7), we notice that in the oscillatory system it is much more difficult to recognize any clear behaviouristic patterns, despite the fact that the variance coefficient itself is smaller than for the excitatory system. However, it affects only the estimation of the drift parameters, as the variance is estimated without any problems

In general, as it was not possible before to estimate the true values of the parameters without fixing ε, we consider it to be a good result. We may just add that in the case of incomplete observations this task becomes even more difficult. Perhaps, it will not be possible to estimate this parameter at all, when the data is incomplete.

6.1.2 ε is fixed: comparison between two schemes.

Now, in order to compare the performance of Euler contrast and LL scheme, we fix the value of ε (see Tables 6.3 and 6.4 for the system with excitatory behaviour and 6.5 and 6.6 for an oscillatory). Note that we do not see any significant difference in the accuracy of estimation for drift parameters for both schemes in the case of excitatory behaviour when the discretization step Δ is small enough. But the estimation for variance is not so accurate with the Euler scheme. Its accuracy was highly dependent on the initial value of parameter with which the optimization routine was initialized. Probably, it is explained by the fact that we miss part of information considering only one coordinate. However we noticed that when we estimated the value of the drift parameters by minimizing the variance of the second coordinate, and then computed σ^2 by exact formula, it gave us much more accurate results. Again, it proves that splitting the estimation for the drift and the variance term is a good idea.

Also note that both schemes face certain difficulties in estimating the parameters, when the data comes from an oscillatory system.

What is also of the great importance in the experiments is that we have to observe the process on the sufficiently long interval of time. Indeed, we can see that for $\Delta = 0.001$ and $N = 100000$ we have the same performance, as for $\Delta = 0.01$ and $N = 10000$, both of them correspond to the time interval $T = 100$. However, for $\Delta = 0.001$ and $N = 10000$ (which corresponds to $T = 10$), accuracy drops down. Thus, we may conclude that it is not always necessary to have an extremely accurate sampling, as it vastly increases the computation time without giving a big impact on the precision.

It is also essential to give good initial conditions before running an optimization, otherwise the algorithm may fail to find a global minimum. Thus we should learn how to analyse the system having the plots of the data. When both variables are available, this task is much more simple that in incomplete observations case. First we note that the spike occurrences are heavily influenced by parameter β — thus, if we see a lot of spikes, it is logical to assume that β is small. Value of σ directly affects the "noisiness" of the plot for the second coordinate. However, it is hard to infer anything about the value of σ, when we see only the first coordinate, as it regulates the noise in the smooth variable only in conjunction with the parameter ε.

	N	1000	1000	10000	10000	100000
	Δ	0.1	0.01	0.01	0.001	0.001
$\beta_0 = 0.3$	(median)	0.21692	0.36126	0.30550	0.46211	0.31903
	(sd)	0.00690	0.03353	0.00153	0.07126	0.00422
$\gamma_0 = 1.5$	(median)	1.19638	1.61290	1.52588	1.60936	1.56972
	(sd)	0.09219	0.03082	0.00255	0.05027	0.00737
$\sigma_0 = 0.6$	(median)	0.62848	0.66630	0.60602	0.62272	0.61189
	(sd)	0.00081	0.00567	0.00004	0.00052	0.00014
$\varepsilon_0 = 0.1$	(median)	603.778	1334.449	1530.759	1705.799	1650.777

Table 6.1: LL scheme: excitatory behaviour

	N	1000	1000	10000	10000	100000
	Δ	0.1	0.01	0.01	0.001	0.001
$\beta_0 = 1.3$	(median)	1.68728	2.01509	1.76156	1.85551	1.63670
	(sd)	0.16811	0.54309	0.21314	0.30883	0.11344
$\gamma_0 = 1.2$	(median)	1.58142	1.88196	1.64123	1.65170	1.55890
	(sd)	0.14807	0.48968	0.19471	0.20405	0.12882
$\sigma_0 = 0.4$	(median)	0.40676	0.41530	0.40472	0.40712	0.40422
	(sd)	0.00007	0.00023	0.00002	0.00005	0.00002
$\varepsilon_0 = 0.1$	(median)	424.734	867.489	862.827	905.706	1019.538

Table 6.2: LL scheme: oscillatory behaviour

	N	1000	1000	10000	10000	100000
	Δ	0.1	0.01	0.01	0.001	0.001
$\beta_0 = 0.3$	(mean)	0.22980	0.32906	0.30758	0.45238	0.33390
	(sd)	0.00805	0.05126	0.00428	0.08918	0.00826
$\gamma_0 = 1.5$	(mean)	1.21619	1.56947	1.51413	1.70697	1.54032
	(sd)	0.08406	0.05714	0.00558	0.09762	0.01499
$\sigma_0 = 0.6$	(mean)	0.61450	0.59847	0.60028	0.60063	0.59995
	(sd)	0.00036	0.00021	0.00001	0.00002	0.00000

Table 6.3: LL scheme: excitatory behaviour

	N	1000	1000	10000	10000	100000
	Δ	0.1	0.01	0.01	0.001	0.001
$\beta_0 = 0.3$	(mean)	0.28682	0.33352	0.34188	0.37761	0.34457
	(sd)	0.00494	0.04763	0.00487	0.04679	0.00555
$\gamma_0 = 1.5$	(mean)	1.22477	1.52921	1.48263	1.58663	1.50122
	(sd)	0.08359	0.04725	0.00477	0.07955	0.00541
$\sigma_0 = 0.6$	(mean)	0.75024	0.78708	0.79755	0.75172	0.78341
	(sd)	0.10986	0.12105	0.10258	0.10965	0.12248

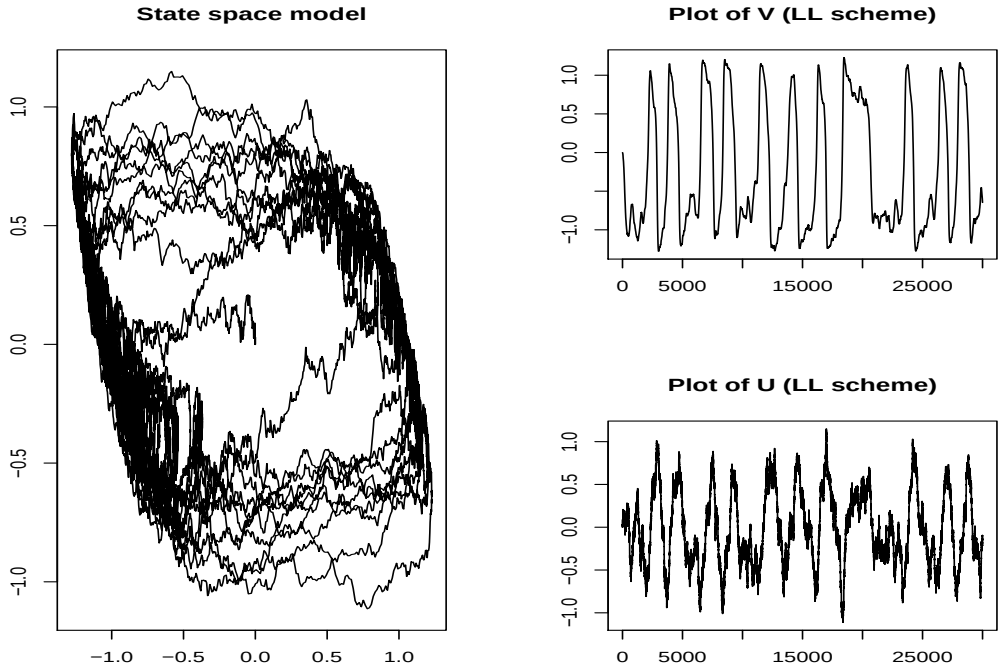
Table 6.4: Euler contrast: excitatory behaviour

	N	1000	1000	10000	10000	100000
	Δ	0.1	0.01	0.01	0.001	0.001
$\beta_0 = 1.3$	(mean)	1.48063	1.67449	1.63759	1.65503	1.69071
	(sd)	0.09724	0.18868	0.16729	0.19375	0.19702
$\gamma_0 = 1.2$	(mean)	1.38544	1.56795	1.52661	1.56570	1.57978
	(sd)	0.09372	0.17997	0.15791	0.20132	0.18444
$\sigma_0 = 0.4$	(mean)	0.40027	0.40088	0.40015	0.40037	0.40011
	(sd)	0.00006	0.00007	0.00001	0.00001	0.00000

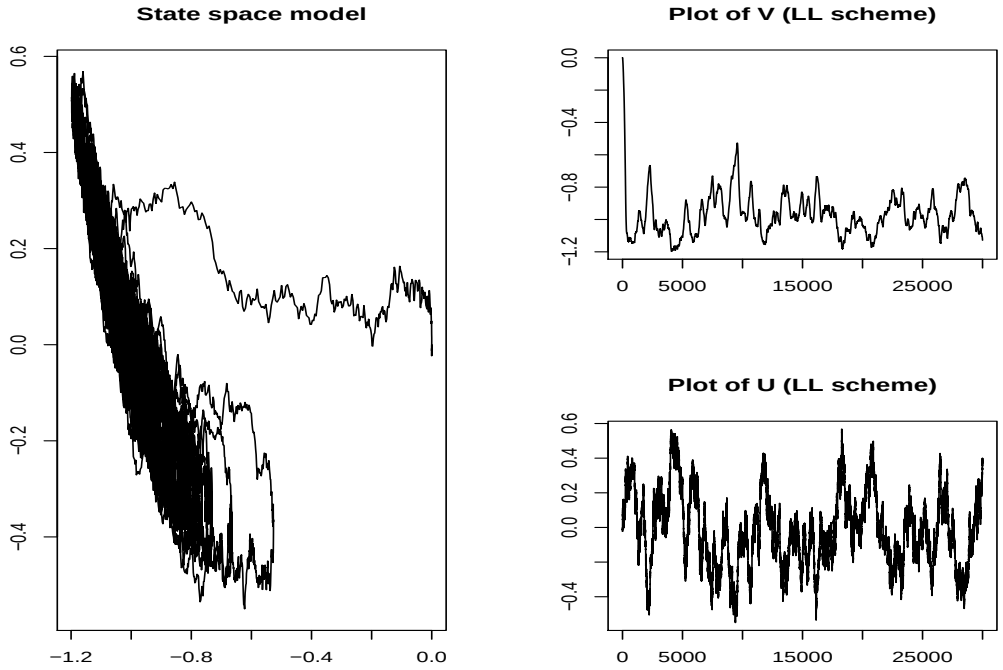
Table 6.5: LL scheme: oscillatory behaviour

	N	1000	1000	10000	10000	10000
	Δ	0.1	0.01	0.01	0.001	0.001
$\beta_0 = 1.3$	(mean)	0.32371	1.51067	1.23086	1.67320	1.40919
	(sd)	1.04718	0.50259	0.04219	0.43565	0.10454
$\gamma_0 = 1.2$	(mean)	0.20556	1.34888	1.09825	1.51356	1.27518
	(sd)	1.08518	0.45730	0.04366	0.37199	0.09307
$\sigma_0 = 0.4$	(mean)	0.68092	0.69061	0.63665	0.67606	0.65343
	(sd)	0.15353	0.15685	0.12509	0.14640	0.12872

Table 6.6: Euler contrast: oscillatory behaviour



(a) Excitatory behaviour: $\sigma = 0.6$, $\gamma = 1.5$, $\beta = 0.3$, $\varepsilon = 0.1$, $s = 0.01$



(b) Oscillatory behaviour: $\sigma = 0.4$, $\gamma = 1.2$, $\beta = 1.3$, $\varepsilon = 0.1$, $s = 0.01$

Figure 6.7: FitzHugh-Nagumo model

6.2 Stochastic approximation of the Hawkes process

Here we consider a system which consists of several populations of neurons, each of them representing a different functional group of neurons (layers in the visual cortex, pools of excitatory and inhibitory neurons in a network, etc.). This system is described by multivariate counting process, which counts the spikes occurrences [Ditlevsen and Löcherbach, 2015]. Let $Z^{\eta}_{j,i}$ represents the number of spikes of the i-th neuron belonging to the j-th population during observed time interval $[0, t]$. Number of classes η as well as the amount of neurons η_j, $k \in 1, \dots, \eta$ in each class remains fixed.

The sequence of counting processes $(Z^{\eta}_{j,i}) \quad \forall 1 \leq j \leq \eta$, $1 \leq i \leq \eta_j$ is characterized by its intensity process $(\lambda^{\eta}_{j,i}(t))$ which is defined through the following relation:

$$\mathbb{P}(Z^{\eta}_{j,i} \text{ has a jump in } (t, t+\Delta]|\mathcal{F}_t) = \lambda^{\eta}_{j,i}(t)\Delta,$$

where $\mathcal{F}_t$ contains all the information about the process $Z^{\eta}_{j,i}$ up to time t. We consider a mean-field framework where $\lambda^{\eta}_{j,i}(t)$ is given by:

$$\lambda^{\eta}_{j,i}(t) = f_j \left(\sum_{l=1}^{\eta} \frac{1}{\eta_l} \sum_{1 \leq j \leq \eta_l} \int_{(0,t)} h_{jl}(t-s) dZ^{\eta}_{l,j}(s) \right),$$

where $\{h_{jl} : \mathbb{R}_+ \to \mathbb{R}\}$ is a family of *synaptic weight functions*, which model the influence of population k on population j, and $f_j : \mathbb{R} \to \mathbb{R}_+$ is the *spiking rate function* of population j.

In [Ditlevsen and Löcherbach, 2015] it was proven that the limit behaviour of this system can be approximated by a stochastic diffusion. We also refer to this work for more explicit description and properties of the system considered in this section. In particular, we know that this process is ergodic and strong Feller and thus we assume that techniques described for the 2-dimensional hypoelliptic diffusion can be applied here.

We consider two interacting populations of neurons ($\eta = 2$), each consists of η_1 and η_2 neurons respectively. One of the populations is inhibitory, and the other one — excitatory. Then the oscillatory behaviour of this system is approximated by the following diffusion process:

$$dZ_t = A(Z_t; \theta)dt + \frac{1}{\sqrt{\eta}} B(Z_t; \theta) dW_t, \tag{6.5}$$

where the observation vector is denoted by $Z_t = (X_1, \dots, X_K)^T$, where $X_i, i \in [1, \dots, \eta_1, \eta_1 + 2, \dots, \eta_2]$ is a membrane potential of a neuron in population 1 or 2, X_{η_j+1} are the recovery variables. $W_t = (W_{1,t}, W_{2,t})^T$ is a 2-dimensional Brownian motion, the drift vector is given by

$$A(Z_t;\theta) = \begin{pmatrix} -\nu_1 X_1 + X_2 \\ -\nu_1 X_2 + X_3 \\ \vdots \\ -\nu_1 X_{\eta_1+1} + c_1 f_2(X_{\eta_1+2}) \\ -\nu_2 X_{\eta_1+2} + X_{\eta_1+3} \\ \vdots \\ -\nu_2 X_k + c_2 f_2(X_1) \end{pmatrix},$$

the diffusion $(k \times 2)$ matrix by

$$B(Z_t;\theta) = \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ 0 & \frac{c_1}{\sqrt{p_2}}\sqrt{f_2(X_{\eta_1+2})} \\ 0 & 0 \\ \vdots & \vdots \\ \frac{c_2}{\sqrt{p_1}}\sqrt{f_1(X_1)} & 0 \end{pmatrix}.$$

Here parameters c_1 and c_2 are responsible for the behaviour of the system: if $c_i > 0$, the population exhibits exhibitory behaviour, otherwise — inhibitory. For simplicity we assume that we deal with only 1 neuron in each population. That is, we have 2 equations with and 2 without noise — they are also called "memory variables". We also assume that we know the form of functions $f_1(x)$ and $f_2(x)$ explicitly, otherwise we had to face a problem of non-parametric estimation, which is out of scope of this thesis.

Explicit condition on the functions f_1 and f_2 can be found in [Ditlevsen and Löcherbach, 2015], we will just mention that these functions are Lipschitz, positive, non-decreasing and differentiable. In the simulation study we use the following functions:

$$f_1(x) = \begin{cases} 10e^x & \text{for } x < \log(20) \\ \frac{400}{1+400e^{-2x}} & \text{for } x \geq \log(20) \end{cases}; \quad f_2(x) = \begin{cases} e^x & \text{for } x < \log(20) \\ \frac{40}{1+400e^{-2x}} & \text{for } x \geq \log(20) \end{cases}.$$

We see that this system includes just 2 equations driven by Brownian motion, all the others are deterministic. This differs from the type of equations that we have considered. Systems of similar type, but with the constant matrix of variance, were already briefly studied in the paper [Pokern et al., 2007], but no computational results for the systems of order higher than 2 were given. It also differs from the model which was investigated in [Ditlevsen and Samson, 2017]. We still want to attempt an estimation for this particular type of systems in order to illustrate that our approach can be generalized to higher-order models and how the problem of numerical instability of the covariance matrix may be possible avoided, at least for certain systems.

Before we proceed, note that the parameters we want to estimate are included in every equation of the system, and that the drift and the variance term depend on the same set

of parameters. Thus, it is natural to try to estimate them using only the drift. The idea is the following: though the variance depends on the process and the parameters, let us assume that for some small enough time step we can consider this term as some noise with certain constant coefficient. Then, in order to estimate the parameters of the drift we have to minimize the quadratic variance on each small interval. In other words, we define the pseudo-contrast in the following way:

$$\mathcal{L}_p(\theta; Z_{0:N}) = \sum_{i=1}^{N-1} (Z_{i+1} - \bar{A}(Z_i;\theta))(Z_{i+1} - \bar{A}(Z_i;\theta))^T,$$

where $\bar{A}(Z_i;\theta)$ is an approximation of the drift which will be defined below.

Here we should also mention work [Le-Breton and Musiela, 1985], where in the similar way was constructed a consistent maximal likelihood estimator for the drift parameters of the linear homogeneous equation.

In fact, this approach must be particularly useful when we have multiple memory variables. Indeed, in that case the order of noise propagated to the smooth equations increases, and the full criterion becomes much more difficult to handle. Though we did not compute the extended covariance matrix for this case, it is natural to guess that if we have only one rough equation in the system (6.5), then the order of noise for smooth equations will be equal to $\Delta^{k+\frac{1}{2}}$, where k is the number of memory variables, thus we can expect that the computational routine become even less stable than in the case with FitzHugh-Nagumo system.

We start from computing the Jacobian of the given system. It is given by following matrix:

$$J_i = \begin{bmatrix} -\nu_1 & 1 & 0 & 0 \\ 0 & -\nu_1 & c_1 f_2'(X_3) & 0 \\ 0 & 0 & -\nu_2 & 1 \\ c_2 f_1'(X_1) & 0 & 0 & \nu_2 \end{bmatrix}.$$

Then the drift term is approximated by $\bar{A}(Z_i;\theta)$ defined as:

$$\begin{pmatrix} X_{1,i} + \Delta(-\nu_1 X_{1,i} + X_{2,i}) + \frac{\Delta^2}{2}\left(-\nu(-\nu_1 X_{1,i} + X_{2,i}) + (-\nu_1 X_{2,i} + c_1 f_2(X_{3,i}))\right) \\ X_{2,i} + \Delta(-\nu_1 X_{2,i} + c_1 f_2(X_{3,i})) + \frac{\Delta^2}{2}\left(-\nu_1(-\nu_1 X_{2,i} + c_1 f_2(X_{3,i})) + c_1 f_2'(X_{3,i})(-\nu_2 X_{3,i} + X_{4,i})\right) \\ X_{3,i} + \Delta(-\nu_2 X_{3,i} + X_{4,i}) + \frac{\Delta^2}{2}\left(-\nu(-\nu_2 X_{3,i} + X_{4,i}) + (-\nu_2 X_{3,i} + c_2 f_1(X_{1,i}))\right) \\ X_{4,i} + \Delta(-\nu_2 X_{3,i} + c_2 f_1(X_{1,i})) + \frac{\Delta^2}{2}\left(-\nu_2(-\nu_2 X_{3,i} + c_2 f_1(X_{1,i})) + c_2 f_1'(X_{1,i})(-\nu_1 X_{1,i} + X_{2,i})\right) \end{pmatrix}$$

First, we generate 30 trajectories approximating the variance term with the Euler scheme in order to illustrate the behaviour of the system (see Figure 6.9) and then we proceed with estimation. We took smaller amount of trajectories, as in 4-dimensional case the estimation works extremely slow. However we believe that the results can still give an idea of the performance of the scheme. Considering Table 6.8, we see that the estimation of the parameters is quite accurate. As in the previous case, it depends on the length of the observed integral more than on the size of the discretization step.

	N	1000	1000	10000	10000
	Δ	0.1	0.01	0.01	0.001
$c_{1,0} = -1$	(mean)	-1.04630	-0.98202	-0.99832	-1.04953
	(sd)	0.00803	0.00417	0.00472	0.00563
$c_{2,0} = 1$	(mean)	1.05643	1.03604	0.98295	1.10056
	(sd)	0.01843	0.03183	0.00714	0.04015
$\nu_{1,0} = 0.8$	(mean)	0.80139	0.78769	0.80772	0.80524
	(sd)	0.00055	0.00037	0.00037	0.00016
$\nu_{2,0} = 1.2$	(mean)	1.29863	1.28867	1.27981	1.38266
	(sd)	0.00820	0.00566	0.00749	0.01439

Table 6.8: Hawkes process

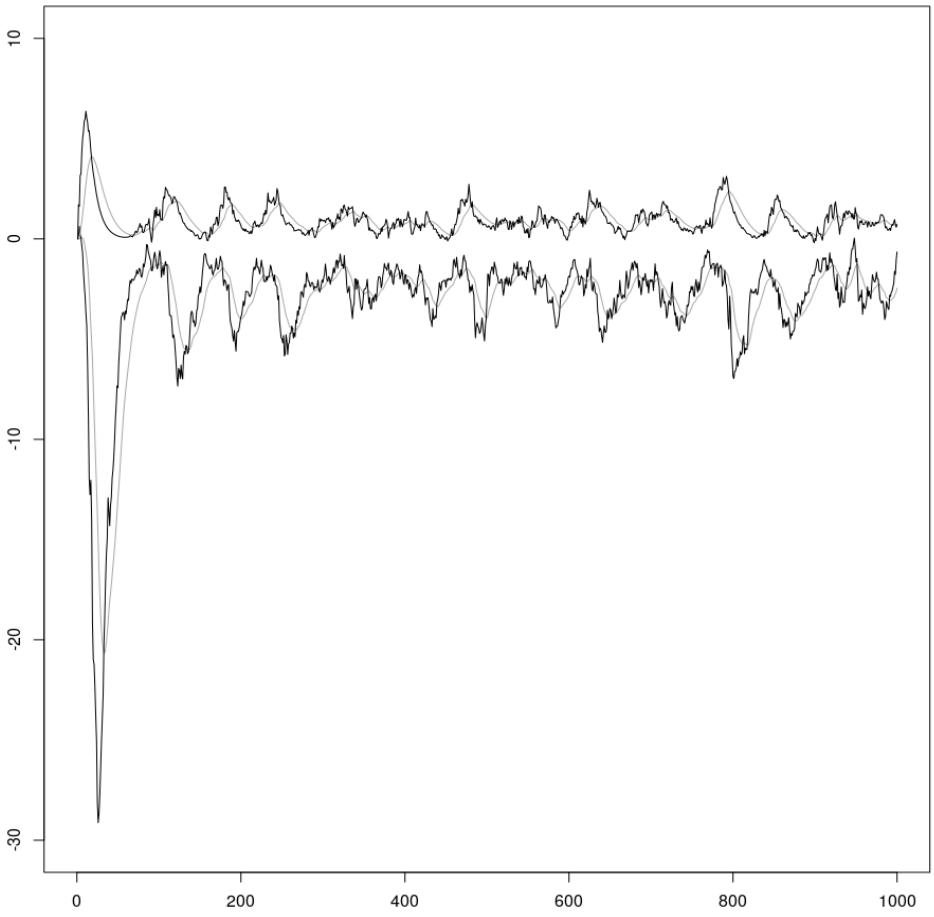


Figure 6.9: Activity of two populations of neurons (excitatory and exhibitory)

6.3 Summary

As a conclusion we may state that it is possible to obtain good estimation of the parameters of the rough equations with a proposed scheme in the case when the data is fully observed. It also allows us to conduct an estimation without having to transform a system. Higher order terms in the approximation of drift make our scheme stable even when the discretization time step is not very small. One of the drawbacks of our approach is that we have to deal with almost singular covariance matrix, which is not numerically stable. The other important issue is that minimization of the complex function like (4.15) leads to a vast increase of a computation time in comparison with the schemes which are based on the approximation of only one coordinate. Finally, proposed scheme does not work good for drift parameters which drive the smooth variables.

On the step which follows the implementation of the contrast itself, we have to deal with the optimization problem. Some general-purpose optimization methods proved their inefficiency in our case. They often fail as we have to find a global minimum for a complex function by few variables simultaneously and it is better to use more stable — though more computationally complex methods.

In addition, during the experiments it was observed that the 2-dimensional contrast is highly sensitive to the initial conditions, and it often breaks down. If we have a possibility to run the estimation several times, choosing different sample sizes and time steps, then it is good idea to consider not only a mean value, but the median as well, as with the first approach we may see inaccurate results due to the outliers. Also note that it is very difficult to estimate the parameters, when the data is very noisy — either because its noisy by its structure, or because the observations are not accurate enough.

In general, it is good idea to split the estimation for the parameters of the first and the second variable whenever it is possible. On the other hand, if we want to develop a general framework to deal with an arbitrary system — we need to find a way to overcome the described problems.

7

Discussion

As we have shown, even though it is possible to adapt the techniques developed for the elliptic diffusions, they are not nearly as effective in hypoelliptic case. Main reason for that is that though we can overcome the problem of non-invertibility of the covariance matrix, it remains highly ill-conditioned and causes numerical problems. While we successfully estimated the parameters included in the equations directly driven by noise, estimation for the parameters included in smooth coordinate is not plausible due to the high-order variance. We suspect that this problem can be solved if we choose a diffusion-independent estimation.

For example, in work [Le-Breton and Musiela, 1985] was proposed a Maximum Likelihood Estimator for the drift parameters of linear homogeneous equations (recall Example 2.1), which is based on the Girsanov formula and does not take into account variance. It is constructed as:

$$\hat{\mathcal{L}}_{MLE} = \left[\int_0^T dZ_t Z_t^T\right]\left[\int_0^T Z_t Z_t^T dt\right]^{-1}$$

It was proven that this estimator is consistent. We expect that it is possible to generalize this approach to the models of general type if we recall our initial assumption that we can represent complex SDEs as piece-wise linear systems. Though we have shown in Chapter 6 that excluding variance from the estimation procedure exhibits good results computationally, due to the lack of time we did not give neither theoretical justifications nor statistical properties of this scheme.

Also the question of the asymptotic normality of our estimator remains open. In particular, we suspect that it does not hold for the parameters of the smooth equation even if we conduct an estimation for every equation separately. This issue has been already discussed in [Ditlevsen and Samson, 2017].

But despite the fact that our mission is far from being done even for toy examples, where the data we work with is simulated and thus for sure obeys the "true" equation, let us highlight few important moments about prospective work. While we already have certain tools that help us to analyze specific models, it is still hard to deal with the real data obtained from intracellular recordings. First of all, all mathematical models of a neuronal activity are accurate only up to some extent, and we need to know how to

measure this accuracy. That means, before we proceed with parametric inference, we need to know whether the data comes from elliptic or hypoelliptic model, as they have different properties. To this moment there are no goodness-of-fit tests which could help us to answer this question given discrete-time observations.

Second obvious problem is that more accurate from neurobiological point of view models cannot be described in terms of two-dimensional SDE with a constant variance. Thus, we have to extend our approach to a general case, adjusting the contrast to multidimensional systems with an arbitrary variance term. And while we may state that it is not hard to apply a described approximation scheme to a multidimensional system, we can expect that numerical performance will not be plausible even for fully-observed case, especially when most of the variables are not driven by noise directly. Indeed, we have shown in Chapter 6 that the covariance matrix is highly ill-conditioned even for the 2-dimensional system.

Also, in more complex cases — for example, for a system of interacting neurons described by a Hawkes process, we cannot apply estimation contrast out of the box. We must keep in mind that the real data which corresponds to this model consists only in the number of spikes occurred on a given interval time, while we worked with a whole trajectory from which spike occurrences can be inferred. Second problem is that we do not know exactly which functions f_1 and f_2 represent the process in the most accurate way. Thus, our methods are to be adjusted.

Apart from statistical tests and problems of generalization, we need to focus on the **parameter inference in the case of incomplete observations** As it was already mentioned, neurophysiologists can observe only some of the processes occurring in a neuron (i.e. we can measure only a membrane potential in FitzHugh-Nagumo model). It is extremely difficult to work with incomplete data, as while the full process is Markovian, process described by only one of the variables is no longer Markovian, and we cannot apply methods from Markov theory.

Several methods have been already developed for *elliptic models*. For example, parametric inference for partially observed data in elliptic FitzHugh-Nagumo model was treated in [Jensen, 2014]. In [Ditlevsen and Samson, 2012] a parameter estimation method for partially observed Morris-Lecar model [Morris and Lecar, 1981] is proposed. It is also notable that numerical experiments in this work were conducted not only for simulated data, but also for real intracellular recordings of neuronal activity in a turtle brain.

When it comes to a *hypoelliptic* case, works are scarce. For SDEs in Hamiltonian form (1.2) it is possible to replace missing variable Y_i by the difference $\frac{X_{i+1}-X_i}{\Delta}$ and adjust an estimator with respect to the bias (see [Samson and Thieullen, 2012], and also [Gloter, 2000, Gloter, 2006] for details). Works [Cattiaux et al., 2014], [Cattiaux et al., 2016] challenge the problem of incomplete observations in systems (1.2) by introducing non-parametric estimators for drift and the variance terms.

In general, main difficulty of statistical inference is that in order to apply the same approach as in the case with complete observation, instead of complete likelihood for the process Z_i defined in (4.12), the estimation has to be build on the marginal likelihood with

respect to an unobserved coordinate:

$$p_\Delta(X_{0:N};\theta,\sigma^2) = \int p(Z_0;\theta,\sigma^2)\prod_{i=0}^{N-1} p_\Delta(Z_{i+1}|Z_i;\theta,\sigma^2)dY.$$

This integral, however, is difficult to process. Instead of trying to build an estimation with the help of a discretized likelihood of the observed process, more popular way to deal with missing data is to input hidden coordinate using filtering (see, for instance, Kalman filter, Paninski filter, Sequential Monte-Carlo method [Kalman et al., 1960], [Doucet et al., 2001], [Del Moral et al., 2001] etc.), and then use stochastic adaptations of an Expectation-Maximization algorithm [Dempster et al., 1977]. First attempt to apply filtering algorithms to the hypoelliptic systems of general type is a preprint [Ditlevsen and Samson, 2017]. In this work was used Sequential Monte-Carlo method coupled with an Expectation-Maximization algorithm.

We should note that despite it is possible to generalize the algorithm to work with multidimensional model, numerical performance reveals certain limitations — in particular, in the case with FitzHugh-Nagumo model, authors pointed out that estimation of the parameter ε is an extremely challenging task and the algorithm breaks down if the initial value of the parameter is not accurate.

We also have to be aware of the fact that not all the parameters can be found when the data is incomplete. The question of which parameters are unidentifiable must be answered with respect to each particular model. Obviously, if we have a model which consists of two completely independent variables, and we do not observe one of them, then all the parameters which are included in the "hidden" equation are unidentifiable. When the variables are correlated, identifiability of the parameters is no more straightforward.

To sum it all up, parametric inference in hypoelliptic diffusions is quite a challenging task, which requires a solid amount of knowledge in all mathematical fields. This thesis is accomplished with a strong feeling that more questions have arisen that the author was able to answer within given time — but also with a hope that to learn how to pose questions in a correct way can also be considered an achievement.

Bibliography

[Bally and Talay, 1996] Bally, V. and Talay, D. (1996). The law of the euler scheme for stochastic differential equations (ii): convergence rate of the density.

[Brette et al., 2007] Brette, R., Rudolph, M., Carnevale, T., Hines, M., Beeman, D., Bower, J. M., Diesmann, M., Morrison, A., Goodman, P. H., Harris, F. C., et al. (2007). Simulation of networks of spiking neurons: a review of tools and strategies. *Journal of computational neuroscience*, 23(3):349–398.

[Cattiaux et al., 2014] Cattiaux, P., León, J. R., and Prieur, C. (2014). Estimation for stochastic damping hamiltonian systems under partial observation. ii drift term. *ALEA (Latin American Journal of Probability and Statistics)*, 11(1):p–359.

[Cattiaux et al., 2016] Cattiaux, P., León, J. R., Prieur, C., et al. (2016). Estimation for stochastic damping hamiltonian systems under partial observation. iii. diffusion term. *The Annals of Applied Probability*, 26(3):1581–1619.

[Del Moral et al., 2001] Del Moral, P., Jacod, J., and Protter, P. (2001). The monte-carlo method for filtering with discrete-time observations. *Probability Theory and Related Fields*, 120(3):346–368.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

[Ditlevsen and Löcherbach, 2015] Ditlevsen, S. and Löcherbach, E. (2015). Multi-class oscillating systems of interacting neurons.

[Ditlevsen and Samson, 2012] Ditlevsen, S. and Samson, A. (2012). Estimation in the partially observed stochastic morris-lecar neuronal model with particle filter and stochastic approximation methods.

[Ditlevsen and Samson, 2017] Ditlevsen, S. and Samson, A. (2017). Approximate likelihood inference in partially observed hypoelliptic diffusions. *(Working paper, preprint).*

[Doucet et al., 2001] Doucet, A., De Freitas, N., and Gordon, N. (2001). An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*. Springer.

[Fitzhugh, 1961] Fitzhugh, R. (1961). Impulses and physiological states in theoretical models of nerve membrane.

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

[Gardiner and Collett, 1985] Gardiner, C. W. and Collett, M. J. (1985). Input and output in damped quantum systems: Quantum stochastic differential equations and the master equation. *Phys. Rev. A*, 31:3761–3774.

[Genon-Catalot and Jacod, 1993] Genon-Catalot, V. and Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. In *Annales de l'IHP Probabilités et statistiques*, volume 29, pages 119–151.

[Genon-Catalot et al., 2000] Genon-Catalot, V., Jeantheau, T., and Larédo, C. (2000). Stochastic volatility models as hidden markov models and statistical applications. bernoulli 6 1051–1079. *Mathematical Reviews (MathSciNet): MR1809735 Digital Object Identifier: doi*, 10:3318471.

[Gerstein and Mandelbrot, 1964] Gerstein, G. L. and Mandelbrot, B. (1964). Random walk models for the spike activity of a single neuron. *Biophysical journal*, 4(1):41–68.

[Gloter, 2000] Gloter, A. (2000). Discrete sampling of an integrated diffusion process and parameter estimation of the diffusion coefficient. *ESAIM: Probability and Statistics*, 4:205–227.

[Gloter, 2006] Gloter, A. (2006). Parameter estimation for a discretely observed integrated diffusion process. *Scandinavian Journal of Statistics*, 33(1):83–104.

[Goldwyn and Shea-Brown, 2011] Goldwyn, J. H. and Shea-Brown, E. (2011). The what and where of adding channel noise to the hodgkin-huxley equations.

[Hodgkin and Huxley, 1952] Hodgkin, A. L. and Huxley, A. F. (1952). A quantitative description of membrane currents and its application to conduction and excitation in nerve.

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

[Jensen, 2014] Jensen, A. C. (2014). *Statistical Inference for Partially Observed Diffusion Processes*. Phd thesis, University of Copenhagen.

[Jensen et al., 2012] Jensen, A. C., Ditlevsen, S., Kessler, M., and Papaspiliopoulos, O. (2012). Markov chain monte carlo approach to parameter estimation in the fitzhugh-nagumo model. *Physical Review E*, 86(4):041114.

[Kalman et al., 1960] Kalman, R. E. et al. (1960). A new approach to linear filtering and prediction problems. *Journal of basic Engineering*, 82(1):35–45.

[Karatzas and Shreve, 1987] Karatzas, I. and Shreve, S. E. (1987). *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics. Springer, 1 edition.

[Kessler, 1997] Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scandinavian Journal of Statistics*, 24(2):211–229.

[Khasminskii, 1969] Khasminskii, R. (1969). *Stability of stochastic differential equations*. Nauka. (in Russian).

[Kloeden et al., 2003] Kloeden, P. E., Platen, E., and Schurz, H. (2003). *Numerical solution of SDE through computer experiments*. Universitext. Springer.

[Le-Breton and Musiela, 1985] Le-Breton, A. and Musiela, M. (1985). Some parameter estimation problems for hypoelliptic homogeneous gaussian diffusions. *Banach Center Publications*, 16(1):337–356.

[Leon and Samson, 2017] Leon, J. R. and Samson, A. (2017). Hypoelliptic stochastic FitzHugh-Nagumo neuronal model: mixing, up-crossing and estimation of the spike rate. working paper or preprint.

[Malliavin and Thalmaier, 2006] Malliavin, P. and Thalmaier, A. (2006). *Stochastic calculus of variations in mathematical finance*. Springer finance. Springer, 1 edition.

[Mattingly et al., 2002] Mattingly, J. C., Stuart, A. M., and Higham, D. J. (2002). Ergodicity for sdes and approximations: locally lipschitz vector fields and degenerate noise. *Stochastic processes and their applications*, 101(2):185–232.

[Morris and Lecar, 1981] Morris, C. and Lecar, H. (1981). Voltage oscillations in the barnacle giant muscle fiber. *Biophysical journal*, 35(1):193–213.

[Nagumo et al., 1962] Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070.

[Oksendal, 2003] Oksendal, B. (2003). *Stochastic differential equations, Universitext*. Springer,.

[Ozaki, 1989] Ozaki, T. (1989). Statistical identification of Nonlinear Random Vibration Systems. *Journal of Applied Mechanics* .

[Ozaki, 2012] Ozaki, T. (2012). *Time series modeling of neuroscience data*. Interdisciplinary statistics. Taylor & Francis.

[Pokern et al., 2007] Pokern, Y., Stuart, A. M., and Wiberg, P. (2007). Parameter estimation for partially observed hypoelliptic diffusions.

[Reynaud-Bouret et al., 2014] Reynaud-Bouret, P., Rivoirard, V., Grammont, F., and Tuleau-Malot, C. (2014). Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *The Journal of Mathematical Neuroscience*, 4(1):3.

[Samson and Thieullen, 2012] Samson, A. and Thieullen, M. (2012). Contrast estimator for completely or partially observed hypoelliptic diffusion. *Stochastic Processes and their Applications*.

[Tuckwell, 2005] Tuckwell, H. C. (2005). *Introduction to theoretical neurobiology: volume 2, nonlinear and stochastic theories*, volume 8. Cambridge University Press.

[Van der Pol, 1920] Van der Pol, B. (1920). A theory of the amplitude of free and forced triode vibrations. *Radio Review*, 1(1920):701–710.

[Wu, 2001] Wu, L. (2001). Large and moderate deviations and exponential convergence for stochastic damping hamiltonian systems. *Stochastic processes and their applications*, 91(2):205–238.

8

Appendix

8.1 Proofs

8.1.1 Properties of the scheme

Approximation of the covariance matrix

Proof. Let us consider each integral of (4.8) separately. Denote:

$$\mathcal{W}_{t+\Delta} = \int_t^{t+\Delta} e^{J(t+\Delta-s)} d\eta_s,$$

where we suppressed dependency of the Jacobian of the starting point on the interval in order to keep notations simple.

Recalling the jacobian of system (3.1) and the definition of the matrix exponent, we have:

$$\mathcal{W}_{t+\Delta} = \int_t^{t+\Delta} (I + J(t+\Delta-s) + \mathcal{O}(\Delta^2)) d\eta_s = \\ = \begin{bmatrix} \partial_y a_1 \int_t^{t+\Delta} (t+\Delta-s) dW_s + \mathcal{O}(\Delta^2) \\ \int_t^{t+\Delta} dW_s + \partial_y a_2 \int_t^{t+\Delta} (t+\Delta-s) dW_s + \mathcal{O}(\Delta^2) \end{bmatrix}$$

Then we can calculate $\mathbb{E}[\mathcal{W}\mathcal{W}']$ (we consider only the terms up to Δ^3:

$$\mathbb{E}[\mathcal{W}\mathcal{W}'] = \mathbb{E} \begin{pmatrix} \Sigma_\Delta^{(1)} & \Sigma_\Delta^{(12)} \\ \Sigma_\Delta^{(12)} & \Sigma_\Delta^{(2)} \end{pmatrix},$$

where entries are given by:

$$\Sigma_\Delta^{(1)} = (\partial_y a_1)^2 \left[\int_t^{t+\Delta} (t+\Delta-s)dW_s\right]^2$$
$$\Sigma_\Delta^{(12)} = \left(\partial_y a_1 \int_t^{t+\Delta} (t+\Delta-s)dW_s\right)\left(\int_t^{t+\Delta} dW_s + \partial_y a_2 \int_t^{t+\Delta} (t+\Delta-s)dW_s\right)$$
$$\Sigma_\Delta^{(2)} = \left(\int_t^{t+\Delta} dW_s + \partial_y a_2 \int_t^{t+\Delta} (t+\Delta-s)dW_s\right)^2$$

First entry can be easily calculated by Itô isometry:

$$\mathbb{E}[\Sigma_\Delta^{(1)}] = (\partial_y a_1)^2 \,\mathbb{E}\left[\int_t^{t+\Delta} (t+\Delta-s)dW_s\right]^2 = (\partial_y a_1)^2 \int_t^{t+\Delta} (t+\Delta-s)ds = (\partial_y a_1)^2 \,\frac{\Delta^3}{3}$$

Now we recall the definition of Itô integral and consider the product of two stochastic integrals involved in the remaining terms (assume for simplicity that $t=0$):

$$\mathcal{I}_{12} := \int_0^\Delta (\Delta-s)dW_s \int_0^\Delta dW_s = $$
$$= \lim_{n\to\infty} \sum_{t_i,t_{i-1}\in[0,\Delta]} (\Delta-s)(W_{t_i}-W_{t_{i-1}})\cdot(W_\Delta - W_0) =$$
$$\lim_{n\to\infty} \sum_{t_i,t_{i-1}\in[0,\Delta]} (\Delta-s)(W_{t_j}-W_{t_{j-1}}) \sum_{t_i,t_{i-1}\in[0,\Delta]} (W_{t_i}-W_{t_{i-1}})$$

Then, after taking an expectation:

$$\mathbb{E}[\mathcal{I}_{12}] = \mathbb{E}\left[\lim_{n\to\infty} \sum_{t_i,t_{i-1}\in[0,\Delta]} (\Delta-s)(W_{t_i}-W_{t_{i-1}}) \sum_{t_i,t_{i-1}\in[0,\Delta]} (W_{t_i}-W_{t_{i-1}})\right] =$$
$$= \lim_{n\to\infty} \sum_{t_i,t_{i-1}\in[0,\Delta]} (\Delta-s)\mathbb{E}\left[(W_{t_i}-W_{t_{i-1}})^2\right] = \lim_{n\to\infty} \sum_{t_i,t_{i-1}\in[0,\Delta]} (\Delta-s)(t_i-t_{i-1}) =$$
$$= \int_0^\Delta (\Delta-s)ds = \frac{\Delta^2}{2}$$

Then, opening the brackets we indeed have:

$$\mathbb{E}[\Sigma_\Delta^{(1)}] = (\partial_y a_1)^2 \,\frac{\Delta^3}{3}$$
$$\mathbb{E}[\Sigma_\Delta^{(12)}] = (\partial_y a_1)^2 \,\frac{\Delta^2}{2} + (\partial_y a_1)(\partial_y a_2)\frac{\Delta^3}{3}$$
$$\mathbb{E}[\Sigma_\Delta^{(2)}] = \Delta + (\partial_y a_2)\frac{\Delta^2}{2} + (\partial_y a_2)^2\frac{\Delta^3}{3}$$

□

Bounds (Proposition 4.4)

Proof. We recall (4.6), that means that each coordinate can be written in the following way:

$$X_{i+1} = \bar{A}_1(Z_i;\theta) + \upsilon_1$$
$$Y_{i+1} = \bar{A}_2(Z_i;\theta) + \upsilon_2$$

First two equalities hold by the definition of the scheme (as they depend just on the chosen accuracy of the representation):

$$\mathbb{E}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right) = \mathcal{O}(\Delta^k) + \mathbb{E}[\upsilon_1] = \mathcal{O}(\Delta^k)$$

The same holds for the second term. Then, obviously, we have:

$$\mathbb{E}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right)^2 = \mathcal{O}(\Delta^{2k}) + \mathbb{E}[\upsilon_1^2] = \left(\partial_{Y_i} a_1\right)^2 \frac{\Delta^3}{3}\sigma^2 + \mathcal{O}(\Delta^4)$$

Analogously for the fourth term. Then, finally:

$$\mathbb{E}\left[\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right)\left(Y_{i+1} - \bar{A}_2(Z_i;\theta)\right)\right] = \mathbb{E}\left[(\mathcal{O}(\Delta^k) + \upsilon_1)(\mathcal{O}(\Delta^k) + \upsilon_2)\right] =$$
$$= \mathcal{O}(\Delta^k) + \mathbb{E}[\upsilon_1\upsilon_2] = \partial_{Y_i} a_1 \frac{\Delta^2}{2}\sigma^2 + \mathcal{O}(\Delta^3)$$

□

Convergence in probability (Lemma 4.1)

Proof. Let us denote:

$$\zeta_i^{(1)} = \frac{1}{n\Delta_n^3}\frac{f(Z_i;\theta)}{(\partial_{Y_i} a_1)^2}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right)^2$$
$$\zeta_i^{(2)} = \frac{1}{n\Delta_n} f(Z_i;\theta)\left(Y_{i+1} - \bar{A}_2(Z_i;\theta)\right)^2$$
$$\zeta_i^{(1,2)} = \frac{1}{n\Delta_n^2}\frac{f(Z_i;\theta)}{\partial_{Y_i} a_1}\left(X_{i+1} - \bar{A}_1(Z_i;\theta)\right)\left(Y_{i+1} - \bar{A}_2(Z_i;\theta)\right)$$

Thanks to Proposition 4.4 we know that:

$$\mathbb{E}_{\theta_0}\left[\zeta_i^{(1)}|\mathcal{F}_i\right] = \frac{\sigma_0^2}{3N}\sum_{i=0}^{N-1} f(Z_i;\theta) + \mathcal{O}(\Delta_N)$$
$$\mathbb{E}_{\theta_0}\left[\zeta_i^{(2)}|\mathcal{F}_i\right] = \frac{\sigma_0^2}{N}\sum_{i=0}^{N-1} f(Z_i;\theta) + \mathcal{O}(\Delta_N)$$
$$\mathbb{E}_{\theta_0}\left[\zeta_i^{(1,2)}|\mathcal{F}_i\right] = \frac{\sigma_0^2}{2N}\sum_{i=0}^{N-1} f(Z_i;\theta) + \mathcal{O}(\Delta_N)$$

Then from Lemma 3.1 it follows that for $N \to \infty$:

$$\mathbb{E}_{\theta_0}\left[\zeta_i^{(1)}|\mathcal{F}_i\right] \xrightarrow{P_\Theta} \frac{\sigma_0^2}{3}\int f(z;\theta)\nu_0(dz)$$
$$\mathbb{E}_{\theta_0}\left[\zeta_i^{(2)}|\mathcal{F}_i\right] \xrightarrow{P_\Theta} \sigma_0^2\int f(z;\theta)\nu_0(dz)$$
$$\mathbb{E}_{\theta_0}\left[\zeta_i^{(1,2)}|\mathcal{F}_i\right] \xrightarrow{P_\Theta} \frac{\sigma_0^2}{2}\int f(z;\theta)\nu_0(dz)$$

□

Proof of Proposition 4.2. According to [Karatzas and Shreve, 1987], we may state that:

$$\mathbb{E}\left[\int_0^\Delta \|Z_s - Z_{\alpha\Delta}\|^2 ds\right] = \int_0^\Delta \mathbb{E}\|Z_s - Z_{\alpha\Delta}\|^2 ds \leq$$
$$\leq \int_0^\Delta C(1+\mathbb{E}\|\xi_0\|^2)|s-\alpha\Delta|ds = C(1+\mathbb{E}\|\xi_0\|^2)\left[\int_0^{\alpha\Delta}(\alpha\Delta - s)ds + \int_{\alpha\Delta}^\Delta (s-\alpha\Delta)ds\right] =$$
$$= C(1+\mathbb{E}\|\xi_0\|^2)\left[\alpha\Delta - \frac{(\alpha\Delta)^2}{2} + \frac{\Delta^2}{2} - \frac{(\alpha\Delta)^2}{2} - \alpha\Delta\right] =$$
$$= C(1+\mathbb{E}\|\xi_0\|^2)(1-\alpha^2)\frac{\Delta^2}{2} \leq C(1+\mathbb{E}\|\xi_0\|^2)\frac{\Delta^2}{2}$$

□

Proof of Proposition 4.3. We start with considering expressions for continuous and discrete time model (which is given by (4.6). For simplicity, assume that the starting point is given and equals to 0 (that is $\xi_0 \equiv 0$). Note that:

$$Z_T = \int_0^T A(Z_t;\theta)dt + \tilde{\sigma}\int_0^T dW_t,$$

where $\tilde{\sigma} = (0,\sigma)^T$ and

$$Z_N = \sum_{i=0}^{N-1}\left[\bar{A}(Z_i;\theta) + \sigma\Upsilon_i\right]$$

To begin with, we split the integrals in the first expression, such that they coincide with the partition of the time interval, and then evaluate the norm, using Proposition 4.2

$$\mathbb{E}\,\|Z_T - Z_N\| = \mathbb{E}\left\|\sum_i^{N-1}\left[\int_{i\Delta}^{(i+1)\Delta} A(Z_t;\theta)dt + \tilde{\sigma}\int_{i\Delta}^{(i+1)\Delta} dW_t - \bar{A}(Z_i;\theta) + \sigma\Upsilon_i\right]\right\| \leq$$
$$\leq \sum_i^{N-1}\mathbb{E}\left\|\int_{i\Delta}^{(i+1)\Delta} A(Z_t;\theta)dt - \bar{A}(Z_i;\theta)\right\| + \sum_i^{N-1}\mathbb{E}\left\|\int_{i\Delta}^{(i+1)\Delta} dW_t - \sigma\Upsilon_i\right\|$$

Now we consider each part separately:

$$\mathbb{E}\left\|\int_{i\Delta}^{(i+1)\Delta} A(Z_t;\theta)dt - \bar{A}(Z_i;\theta)\right\| \approx \mathbb{E}\left\|\int_{i\Delta}^{(i+1)\Delta} \left(A(Z_t;\theta) - A(Z_i;\theta)\right) dt\right\| \leq$$

$$\leq \mathbb{E}\left[K\int_{i\Delta}^{(i+1)\Delta} \|Z_t - Z_i\|\, dt\right] \leq \tilde{C}\Delta^2$$

Recall that $\upsilon_2 = W_{(i+1)\Delta} - W_{i\Delta}$ and υ_1 is integrated Brownian motion, and that $\mathbb{E}[\upsilon_1^2] = \frac{\Delta^3}{3}$, $\mathbb{E}[\upsilon_1\upsilon_2] = \frac{\Delta^2}{2}$

$$\mathbb{E}\left\|\tilde{\sigma}\int_{i\Delta}^{(i+1)\Delta} dW_t - \sigma\Upsilon_i\right\| = \sigma\mathbb{E}\left\|\begin{matrix}\upsilon_1\\ \upsilon_2 - \upsilon_2\end{matrix}\right\| = \sigma\mathbb{E}[\upsilon_1^2]^{\frac{1}{2}} = \frac{\sigma}{\sqrt{3}}\Delta^{\frac{3}{2}}$$

Thus, we may state that for the system (3.1) Local Linearization scheme has an order of strong convergence 1.5. □

8.1.2 Consistency of the estimator

Recall that the second-order Taylor approximation of matrix Σ_Δ has the form (4.8), consequently $\det(\Sigma_\Delta) = \sigma^4\left(\partial_{Y_i}a_1\right)^2\frac{\Delta^4}{12}$, and the inverse is given by (4.14). Note that dependency on σ^2 is, in fact, linear, thus it is reasonable to replace the original matrix by $\sigma^2\bar{\Sigma}(Z_i;\theta)$ and split the estimation for drift and variance parameters. We also denote the parameters in the first coordinate by φ, and in the second — by ψ.

Throughout the section we will use notations T_i in order to split cumbersome expressions to more easily analysable parts, and they are redefined each time independently of what was used before (we do it in order to avoid numerous unnecessary variables). We also introduce index in the notation for the time step in lemmas in order to highlight that it depends on the experimental design (thus we will use Δ_N instead of Δ). However, in calculations, where this dependency is not important, we will use the old notations.

Before we proceed, let us introduce an auxiliary lemma which just repeats Lemma 3 in [Ditlevsen and Samson, 2017] and give the idea of the proof. The proof is based on Lemma 9 from [Genon-Catalot and Jacod, 1993] and Lemma 10 [Kessler, 1997].

Lemma 8.1. *Let $f : \mathbb{R}^2 \times \Theta \to \mathbb{R}$ be a function with derivatives of polynomial growth in x, uniformly in θ.*

1. *Assume $\Delta_N \to 0$ and $N\Delta_N \to \infty$. Then*

$$\frac{1}{N\Delta_N^2}\sum_{i=0}^{N-1} f(Z_i;\theta)(X_{i+1} - \bar{A}_1(Z_i;\theta_0)) \xrightarrow{\mathbb{P}_\theta} 0$$

uniformly in θ.

2. *Assume* $\Delta_N \to 0$ *and* $N\Delta_N \to \infty$. *Then*

$$\frac{1}{N\Delta_N}\sum_{i=0}^{N-1} f(Z_i;\theta)(Y_{i+1} - \bar{A}_2(Z_i;\theta_0)) \xrightarrow{\mathbb{P}_\theta} 0$$

uniformly in θ.

3. $\Delta_N \to 0$ *and* $N \to \infty$. *Then*

$$\frac{1}{N}\sum_{i=0}^{N-1} f(Z_i;\theta)(Y_{i+1} - \bar{A}_2(Z_i;\theta_0)) \xrightarrow{\mathbb{P}_\theta} 0$$

uniformly in θ.

Proof. Expectation of the first sum tends to zero for $\Delta_N \to 0$ and $N\Delta_N \to \infty$ due to Proposition 4.4. Convergence for θ is due Lemma 9 in [Genon-Catalot and Jacod, 1993] and uniformity in θ follows the proof of Lemma 10 [Kessler, 1997] Second and the third assertions are proved in the same way, with the proper scaling. □

Consistency of $\hat{\sigma}^2$

Start with considering the expression:

$$\mathcal{L}_N(\theta,\sigma^2;Z_{0:N}) = \frac{1}{2\sigma^2}\sum_{i=0}^{N-1}(Z_{i+1} - \bar{A}(Z_i;\theta))^T\tilde{\Sigma}^{-1}(Z_{i+1} - \bar{A}(Z_i;\theta)) + \sum_{i=0}^{N-1}\log\sigma^4$$

Define

$$\hat{\sigma}^2 = \arg\min_{\sigma^2} \mathcal{L}_N(\theta,\sigma^2;Z_{0:N})$$

Consistency follows from the following Lemma:

Lemma 8.2. *Assume* $\Delta_N \to 0$ *and* $N \to \infty$ *and* $\varphi \equiv \varphi_0$. *Then*

$$\frac{1}{N}\mathcal{L}_N(\theta,\sigma^2;Z_{0:N}) \xrightarrow{\mathbb{P}_\theta} 2\int\left(\frac{\sigma_0^2}{\sigma^2} + \log\sigma^2\right)\nu_0(dz)$$

uniformly in θ.

Let us denote the integral on the right by $\mathcal{I}(\sigma,\sigma_0)$. Then we can find such a subsequence N_k that $\hat{\sigma}^2_{N_k}$ converges to σ^2_∞ and then $\frac{1}{N}\mathcal{L}_N(\theta,\sigma^2;Z_{0:N}) \xrightarrow{\mathbb{P}_\theta} \mathcal{I}(\sigma_\infty,\sigma_0)$ by continuity of $\sigma \to \mathcal{I}(\sigma,\sigma_0)$. By definition of the estimator for σ^2 we know that $\mathcal{I}(\sigma_\infty,\sigma_0) \le \mathcal{I}(\sigma_0,\sigma_0)$. At the same time we know that $\frac{\sigma_0^2}{\sigma^2} + \log\sigma^2 \ge 1 + \log\sigma_0^2$. Then $\mathcal{I}(\sigma_\infty,\sigma_0) \le \mathcal{I}(\sigma_0,\sigma_0)$. That proves the consistency.

Consistency of $\hat{\theta}$

Before proceed with the proof, let us introduce a rescaled version of the original contrast function (4.15), which is defined in the following way:

$$\bar{\mathcal{L}}_{N,\Delta}(\theta, \sigma^2; Z_{0:N}) = \frac{1}{2\sigma^2} \sum_{i=0}^{N-1} \begin{pmatrix} \sqrt{\Delta} V_1(Z_i;\theta) \\ \sqrt{\frac{1}{\Delta}} V_2(Z_i;\theta) \end{pmatrix} \tilde{\Sigma}^{-1} \begin{pmatrix} \sqrt{\Delta} V_1(Z_i;\theta) \\ \sqrt{\frac{1}{\Delta}} V_2(Z_i;\theta) \end{pmatrix} + \sum_{i=0}^{N-1} \log(\partial_{Y_i} a_1)_\theta$$

The idea of rescaling follows from the fact that each coordinate has a variance of different order, and in order to insure the convergence of the criterion under the common assumption, we have to compensate this fact. Thus, we will prove the consistency considering each coordinate separately and taking into account the different scaling of the original contrast function (4.15).

Recall the expression (4.15) and consider the following difference, with the parameter ψ of the second variable being fixed to the true value:

Lemma 8.3. *Assume $\Delta_N \to 0$ and $N\Delta_N \to \infty$. Then*

$$\frac{\Delta}{N} \left[\mathcal{L}_{N,\Delta}(\varphi, \psi_0, \sigma^2; Z_{0:N}) - \mathcal{L}_{N,\Delta}(\varphi_0, \psi_0, \sigma^2; Z_{0:N}) \right] \xrightarrow{\mathbb{P}_\theta} \frac{6}{\sigma_0^2} \int \frac{(a_1(z;\theta_0) - a_1(z;\theta))^2}{(\partial_y a_1)_\theta^2} \nu_0(dz)$$

uniformly in θ.

From this lemma we deduce that there exists a subsequence of $\hat{\varphi}_{N_k}$ which tends to φ_∞. Then the consistency follows from the identifiability of the drift functions (assumption (A4)).

Similar result holds for the parameter ψ, with $\varphi = \varphi_0$, giving us the consistency with the same reasoning:

Lemma 8.4. *Assume $\Delta_N \to 0$ and $N\Delta_N \to \infty$. Then*

$$\frac{1}{N\Delta} \left[\mathcal{L}_{N,\Delta}(\varphi_0, \psi, \sigma^2; Z_{0:N}) - \mathcal{L}_{N,\Delta}(\varphi_0, \psi_0, \sigma^2; Z_{0:N}) \right] \xrightarrow{\mathbb{P}_\theta} \frac{2}{\sigma_0^2} \int (a_2(z;\theta_0) - a_2(z;\theta))^2 \nu_0(dz)$$

uniformly in θ.

It remains to prove the lemmas itself:

Proof of Lemma 8.2. Consider

$$\frac{1}{N}\mathcal{L}_N(\theta,\sigma^2;Z_{0:N}) =$$
$$= \frac{1}{N\sigma^2}\sum_{i=0}^{N-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0)+\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))^T\tilde{\Sigma}^{-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0)+\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))+$$
$$+\frac{1}{N}\sum_{i=0}^{N-1}\log\sigma^4 = \frac{1}{N\sigma^2}\sum_{i=0}^{N-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0))^T\tilde{\Sigma}^{-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0))+$$
$$+\frac{2}{N\sigma^2}\sum_{i=0}^{N-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0))^T\tilde{\Sigma}^{-1}(\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))+$$
$$+\frac{1}{N\sigma^2}\sum_{i=0}^{N-1}(\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))^T\tilde{\Sigma}^{-1}(\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))+\log\sigma^4$$

Now let us denote

$$T_1 = \frac{1}{2N\sigma^2}\sum_{i=0}^{N-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0))^T\tilde{\Sigma}^{-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0))$$
$$T_2 = \frac{1}{N\sigma^2}\sum_{i=0}^{N-1}(Z_{i+1}-\bar{A}(Z_i;\theta_0))^T\tilde{\Sigma}^{-1}(\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))$$
$$T_3 = \frac{1}{2N\sigma^2}\sum_{i=0}^{N-1}(\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))^T\tilde{\Sigma}^{-1}(\bar{A}(Z_i;\theta_0)-\bar{A}(Z_i;\theta))$$
$$T_4 = \log\sigma^4$$

We start with considering the first term, recalling Lemma 4.1 . For readability denote $X_{i+1}-\bar{A}_1(Z_i;\theta) := V_1(Z_i;\theta)$ and $Y_{i+1}-\bar{A}_2(Z_i;\theta) := V_2(Z_i;\theta)$.

$$T_1 = \frac{1}{2N\sigma^2}\sum_{i=0}^{N-1}\left(\frac{12}{(\partial_{Y_i}a_1)^2_{\theta_0}\Delta^3}V_1^2(Z_i;\theta_0) - \frac{12}{(\partial_{Y_i}a_1)_{\theta_0}\Delta^2}V_1(Z_i;\theta_0)V_2(Z_i;\theta_0) + \frac{4}{\Delta}V_2^2(Z_i;\theta_0)\right)$$

For the first, second and the last term Lemma 4.1 is applied directly. Note that the third term converges to zero, as it has the variance of order Δ^3. Then we see that

$$T_1 \xrightarrow{\mathbb{P}_\theta} \frac{1}{2N\sigma^2}\sum_{i=0}^{N-1}\left[12\sigma_0^2 - 12\sigma_0^2 + 4\sigma_0^2\right] = 2\frac{\sigma_0^2}{\sigma^2}$$

Write T_2 as following:

$$T_2 = \frac{1}{2N\sigma^2}\sum_{i=0}^{N-1}(V_1(Z_i;\theta_0) \quad V_2(Z_i;\theta_0))\begin{pmatrix}12(\partial_{Y_i}a_1)^{-2}_{\theta_0}\Delta^{-3} & -6(\partial_{Y_i}a_1)^{-1}_{\theta_0}\Delta^{-2}\\ -6(\partial_{Y_i}a_1)^{-1}_{\theta_0}\Delta^{-2} & 4\Delta^{-1}\end{pmatrix}\begin{pmatrix}U_1(Z_i;\theta,\theta_0)\\ U_2(Z_i;\theta,\theta_0)\end{pmatrix},$$

where $U_1(Z_i;\theta,\theta_0) = \bar{A}_1(Z_i;\theta_0) - \bar{A}_1(Z_i;\theta)$ and $U_2(Z_i;\theta,\theta_0) = \bar{A}_2(Z_i;\theta_0) - \bar{A}_2(Z_i;\theta)$.

Computing this expression explicitly, we get:

$$T_2 = \frac{1}{2N\sigma^2}\sum_{i=0}^{N-1}\left(\frac{12}{(\partial_{Y_i}a_1)^2_{\theta_0}\Delta^3}V_1(Z_i;\theta_0)U_1(Z_i;\theta,\theta_0) - \frac{6}{(\partial_{Y_i}a_1)_{\theta_0}\Delta^2}V_2(Z_i;\theta_0)U_1(Z_i;\theta,\theta_0) - \right.$$
$$\left. - \frac{6}{(\partial_{Y_i}a_1)_{\theta_0}\Delta^2}V_1(Z_i;\theta_0)U_2(Z_i;\theta,\theta_0) + \frac{4}{\Delta}V_2(Z_i;\theta_0)U_2(Z_i;\theta,\theta_0)\right)$$

By Lemma 8.1 and the fact that U_i is of order $\mathcal{O}(\Delta^2)$ (consequently, U_i^2 is of order $\mathcal{O}(\Delta^4)$) when $\varphi \equiv \varphi_0$ we conclude that

$$T_2 \xrightarrow{\mathbb{P}_\theta} 0$$

Then we consider T_3:

$$T_3 = \frac{1}{2N\sigma^2}\sum_{i=0}^{N-1}\left(\frac{12}{(\partial_{Y_i}a_1)^2_{\theta_0}\Delta^3}U_1^2(Z_i;\theta,\theta_0) - \frac{12}{(\partial_{Y_i}a_1)_{\theta_0}\Delta^2}U_1(Z_i;\theta,\theta_0)U_2(Z_i;\theta,\theta_0). + \frac{4}{\Delta}U_2^2(Z_i;\theta,\theta_0)\right)$$

T_3 converges to 0 as $\Delta_N \to 0$ as each term is of order Δ.

Finally, we obtain:

$$T_1 + T_2 + T_3 + T_4 \xrightarrow{\mathbb{P}_\theta} 2\int\left(\frac{\sigma_0^2}{\sigma^2} + \log\sigma^2\right)\nu_0(dz)$$

□

Proof of the Lemma 8.3.

$$\frac{\Delta}{2N}\left[\mathcal{L}_{N,\Delta}(\theta,\sigma^2;Z_{0:N}) - \mathcal{L}_{N,\Delta}(\theta_0,\sigma^2;Z_{0:N})\right] =$$
$$= \frac{\Delta}{2N\sigma_0^2}\sum_{i=0}^{N-1}\left[\left(\frac{12}{(\partial_{Y_i}a_1)^2_{\theta}\Delta^3}V_1^2(Z_i;\theta) - \frac{12}{(\partial_{Y_i}a_1)_{\theta}\Delta^2}V_1(Z_i;\theta)V_2(Z_i;\theta_0) + \frac{4}{\Delta}V_2^2(Z_i;\theta_0)\right) - \right.$$
$$\left. - \left(\frac{12}{(\partial_{Y_i}a_1)^2_{\theta_0}\Delta^3}V_1^2(Z_i;\theta_0) - \frac{12}{(\partial_{Y_i}a_1)_{\theta_0}\Delta^2}V_1(Z_i;\theta_0)V_2(Z_i;\theta_0) + \frac{4}{\Delta}V_2^2(Z_i;\theta_0)\right)\right] +$$
$$+ \Delta\log\left(\frac{(\partial_{Y_i}a_1)_\theta}{(\partial_{Y_i}a_1)_{\theta_0}}\right) = \frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}\left[\frac{1}{(\partial_{Y_i}a_1)^2_{\theta}}V_1^2(Z_i;\theta) - \frac{1}{(\partial_{Y_i}a_1)^2_{\theta_0}}V_1^2(Z_i;\theta_0)\right] +$$
$$+ \frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}V_2(Z_i;\theta_0)\left[\frac{1}{(\partial_{Y_i}a_1)_{\theta_0}}V_1(Z_i;\theta_0) - \frac{1}{(\partial_{Y_i}a_1)_{\theta}}V_1(Z_i;\theta)\right] + \Delta\log\left(\frac{(\partial_{Y_i}a_1)_\theta}{(\partial_{Y_i}a_1)_{\theta_0}}\right)$$

Consider:

$$T_1 = \frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}\left[\frac{1}{(\partial_{Y_i}a_1)_\theta^2}V_1^2(Z_i;\theta) - \frac{1}{(\partial_{Y_i}a_1)_{\theta_0}^2}V_1^2(Z_i;\theta_0)\right]$$
$$T_2 = \frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}V_2(Z_i;\theta_0)\left[\frac{1}{(\partial_{Y_i}a_1)_{\theta_0}}V_1(Z_i;\theta_0) - \frac{1}{(\partial_{Y_i}a_1)_\theta}V_1(Z_i;\theta)\right]$$
$$T_3 = \Delta\log\left(\frac{(\partial_{Y_i}a_1)_\theta}{(\partial_{Y_i}a_1)_{\theta_0}}\right)$$

Start with T_1:

$$T_1 = \frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}\left[\frac{1}{(\partial_{Y_i}a_1)_\theta^2}V_1^2(Z_i;\theta) - \frac{1}{(\partial_{Y_i}a_1)_{\theta_0}^2}V_1^2(Z_i;\theta_0)\right] =$$
$$= \frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}\left[\frac{1}{(\partial_{Y_i}a_1)_\theta^2}(X_{i+1} - \bar{A}_1(Z_i;\theta_0) + \bar{A}_1(Z_i;\theta_0) - \bar{A}_1(Z_i;\theta))^2 - \frac{1}{(\partial_{Y_i}a_1)_{\theta_0}^2}V_1^2(Z_i;\theta_0)\right] =$$
$$= \frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}\left[V_1^2(Z_i;\theta_0)\left(\frac{1}{(\partial_{Y_i}a_1)_\theta^2} - \frac{1}{(\partial_{Y_i}a_1)_{\theta_0}^2}\right) + \frac{(\bar{A}_1(Z_i;\theta_0) - \bar{A}_1(Z_i;\theta))^2}{(\partial_{Y_i}a_1)_\theta^2} + \right.$$
$$\left. + \frac{2V_1(Z_i;\theta_0)(\bar{A}_1(Z_i;\theta_0) - \bar{A}_1(Z_i;\theta))}{(\partial_{Y_i}a_1)_\theta^2}\right]$$

We see that the first and the last term of the expression tend to 0, as the variance of each term is bigger than Δ^2. From Lemma 8.1 we see that the second term converges in probability to the following value:

$$\frac{6\Delta}{N\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}\frac{(\bar{A}_1(Z_i;\theta_0) - \bar{A}_1(Z_i;\theta))^2}{(\partial_{Y_i}a_1)_\theta^2} \xrightarrow{\mathbb{P}_\theta} \frac{6}{\sigma_0^2}\int\frac{(a_1(z;\theta_0) - a_1(z;\theta))^2}{(\partial_y a_1)_\theta^2}\nu_0(dz)$$

Finally, consider T_2:

$$T_2 = \frac{6\Delta}{\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}V_2(Z_i;\theta_0)\left[\frac{1}{(\partial_{Y_i}a_1)_{\theta_0}}V_1(Z_i;\theta_0) - \frac{1}{(\partial_{Y_i}a_1)_{\theta_0}}V_1(Z_i;\theta) + \right.$$
$$\left. + \frac{1}{(\partial_{Y_i}a_1)_{\theta_0}}V_1(Z_i;\theta) - \frac{1}{(\partial_{Y_i}a_1)_\theta}V_1(Z_i;\theta)\right] = \frac{6\Delta}{\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}\left[\frac{V_2(Z_i;\theta_0)}{(\partial_{Y_i}a_1)_{\theta_0}}(\bar{A}_1(Z_i;\theta_0) - \bar{A}_1(Z_i;\theta))\right] +$$
$$+ \frac{6\Delta}{\sigma_0^2\Delta^3}\sum_{i=0}^{N-1}V_2(Z_i;\theta_0)V_1(Z_i;\theta)\left[\frac{1}{(\partial_{Y_i}a_1)_{\theta_0}} - \frac{1}{(\partial_{Y_i}a_1)_\theta}\right]$$

By the Lemma 8.1 it is easy to see that T_2 converges to 0. □

Proof of the Lemma 8.4.

$$\frac{1}{N\Delta}\left[\mathcal{L}_{N,\Delta}(\theta,\sigma^2;Z_{0:N})-\mathcal{L}_{N,\Delta}(\theta_0,\sigma^2;Z_{0:N})\right]=\frac{6V_1(Z_i;\theta_0)}{\sigma_0^2N\Delta^3(\partial_{Y_i}a_1)_{\theta_0}}\sum_{i=0}^{N-1}\left[V_2(Z_i;\theta_0)-V_2(Z_i;\theta)\right]+$$

$$+\frac{2}{\sigma_0^2N\Delta^2}\sum_{i=0}^{N-1}\left[V_2^2(Z_i;\theta)-V_2^2(Z_i;\theta_0)\right]=\frac{6V_1(Z_i;\theta_0)}{\sigma_0^2N\Delta^3(\partial_{Y_i}a_1)_{\theta_0}}\sum_{i=0}^{N-1}\left[\bar{A}_2(Z_i;\theta_0)-\bar{A}_2(Z_i;\theta)\right]+$$

$$+\frac{2}{\sigma_0^2N\Delta^2}\sum_{i=0}^{N-1}\left[(\bar{A}_2(Z_i;\theta_0)-\bar{A}_2(Z_i;\theta))^2-2V_2(Z_i;\theta_0)(\bar{A}_2(Z_i;\theta_0)-\bar{A}_2(Z_i;\theta)\right]$$

By Lemma 8.1, we know that the first and the last term tends to 0. By the analogy with the previous example the term in the middle converges in probability to a finite value:

$$\frac{2}{\sigma_0^2N\Delta^2}\sum_{i=0}^{N-1}(\bar{A}_2(Z_i;\theta_0)-\bar{A}_2(Z_i;\theta))^2\xrightarrow{\mathbb{P}_\theta}\frac{2}{\sigma_0^2}\int(a_2(z;\theta_0)-a_2(z;\theta))^2\nu_0(dz)$$

□