Inference for Stochastic Processes

Jean Jacod *

SUMMARY:

We present a review, without proofs, of some of the methods used for estimating an unknown parameter occuring in the coefficients of a diffusion process, under various observation schemes. These schemes are mostly doscrete in time, along a regular grid, with say n observations, and we are mainly interested in asymptotically good (or even optimal) statistical procedures, in various limiting situations: the mesh of the grid is constant, or it goes to 0 at some rate in function of the number n. We make an attempt to compare the pro and con of those methods. We also consider a situation not so commonly studied so far by statisticians namely the case where each obsevation is made with a measurement error, in two cases: the error is an additive error, or it is a round-off error.

SUBJECT CLASSIFICATIONS: 60J60, 62M05, 62M09

KEY WORDS: Diffusion processes; asymptotic statistical theory; LAN property; estimating functions.

*Laboratoire de Probabilités et Modèles aléatoires, Université Paris 6, 4 place Jussieu, 75252 Paris, France (CNRS UMR 7599)

1 Introduction

The title of this paper is clearly overoptimistic and a bit misleading: another appropriate title, perhaps more to the point, would have been "discretely sampled diffusions", but this was already used by other contributors to this volume... and somehow our title stresses our ambition, which is to write a sort of commented review of the topic of inference for diffusion processes.

The reader should ask himself right away whether it is really possible to describe such a topic in any sort of depth within a score of pages ? For example Prakasa Rao devoted two thick books [32] and [33] to this, without exhausting the subject. But here our aim is much more modest and also slightly different in spirit: we essentially consider parametric inference for diffusion processes observed at discrete times, and for each setting we usually describe a single estimation method, sometimes very quickly, and of course no proof is given; on the other hand we describe a variety of observation schemes and we try to compare these various schemes when it comes to applying the methods to concrete data.

Considering only diffusion processes is motivated by their wide use in finance and also by the fact that essentially nothing is known about statistics for other continuous-time processes, apart from point processes which occur in quite other contexts, very far from finance. Considering only parametric inference is motivated mainly by the fact that so far most models used in finance are parametric models, but also by the facts that little is known about non-parametric inference for one-dimensional diffusions, while for multi-dimensional diffusions non-parametric inference is perhaps even meaningless as soon as one wants to infer the diffusion coefficient (the volatility).

As in most statistical contexts, the first concern of the statistician is the structure of the available data. This is all the more important for continuous-time processes, since many different observation schemes might be thought of: one may observe the whole path of the process over some time interval (a very rare occurence indeed), or the exact values taken by the process at some "discrete" times, either regularly spaced or not, or even at random times; one may also observe a "regularization" of the path (this is certainly the case in physical applications, probably much less the case in finance), or the values taken by the process at some times, but blurred by some kind of error, and so on... Here we mainly consider the case where n observations are given, regularly spaced on a grid with mesh Δ_n. The number n is usually very large, this is why we are interested in asymptotic properties as n goes to infinity; as for the mesh Δ_n it might be "small" (relatively to the characteristics of the diffusion process), or not: so we study both the case where $\Delta_n \to 0$ and the case where $\Delta_n = \Delta$ is fixed; in practice, though, n and Δ_n are given, and we have to decide whether one can consider Δ_n as small or not...

Another question will be closely looked at in the paper: what happens when the data are blurred with measurement errors ? Surprisingly enough, very few papers have been devoted so far to this topic, which we feel to be of much importance. Two kinds of errors will be considered, both of them quite likely to occur in finance as well as in other situations: first when each value of the process is measured with an additive independent error, next when each value is measured with a round-off error. We also put a lot of emphasis on the asymptotic "optimality", or lack of optimality, of the procedures we

describe.

The structure of the paper is as follows: we start with a very brief account on diffusion processes (Section 2), and another short reminder about asymptotic optimality in statistics (Section 3). In Section 4 we give some general facts about statistics of diffusions. Sections 5 and 6 are devoted to regularly spaced observations, with a mesh Δ_n going to 0 or being fixed, respectively. In Section 7 we study the situation when the process is observed with measurement errors. Finally Section 8 is devoted to some concluding remarks and some hints about discontinuous processes.

2 About diffusion processes

There is a large litterature on diffusion processes, and one can for example refer to Øksendal [29] for an introductory account, or to "classical" books like Stroock and Varadhan [36], Liptser and Shiryayev [28], Ikeda and Watanabe [15] or Revuz and Yor [35]. Of special interest for finance is of course the book of Karatzas and Shreve [20]. For likelihood ratios we refer to [28] or [16].

In most of this section, we consider only a single given diffusion process: this is in contrast with the rest of the paper, where a whole family of diffusion processes, depending on some parameter θ, is given.

2.1 The basic setting

By a "diffusion process", we mean the solution $X = (X_t)$ of the following stochastic differential equation (SDE in short):

$$dX_t = a(t, X_t)dt + \sigma(t, X_t)dW_t, \qquad X_0 = U. \tag{2.1}$$

Here, time t typically ranges through the real half-line $\mathbb{R}_+ = [0, \infty)$; the process X takes its values in $\mathbb{R}^d$ for some integer d, so X_t has d components $(X_t^i)_{1 \le i \le d}$; next, $W = (W_t)_{t \ge 0}$ stands for a d'-dimensional standard Wiener process. The other ingredients of Equation (2.1) are:

(i) The *initial condition* $U = (U^i)_{1 \le i \le d}$, which is a random vector with values in $\mathbb{R}^d$, and independent from the Wiener process W.

(ii) The *drift coefficient* $a = (a^i)_{1 \le i \le d}$ which is a measurable function from $\mathbb{R}_+ \times \mathbb{R}^d$ into $\mathbb{R}^d$.

(iii) The *diffusion coefficient* $\sigma = (\sigma^{i,j})_{1 \le i \le d, 1 \le j \le d'}$ which is a $d \times d'$-matrix-valued measurable function on $\mathbb{R}_+ \times \mathbb{R}^d$. We also associate with σ the $d \times d$ symmetrical non-negative matrix $c(t, x) = \sigma(t, x)\sigma(t, x)^\star$ (where "$\star$" stands for the transpose): sometimes c is also called "diffusion coefficient".

Now that the basic terms are defined, we can introduce the notion(s) of a "solution". The simplest notion is called a solution-process: in addition to the data a and σ, we

start with a given initial condition U and a given Wiener process W, all defined on some probability space $(\Omega, \mathcal{F}, I\!P)$ endowed with a *filtration* $(\mathcal{F}_t)$. This space supports a d-dimensional random vector U (the initial condition) which is $\mathcal{F}_0$-measurable. It also supports a d'-dimensional Wiener process W which is in fact an $(\mathcal{F}_t)$-Wiener process: it is *adapted* to the filtration $(\mathcal{F}_t)$ (each variable W_t is $\mathcal{F}_t$-measurable), and for any $0 \leq s \leq t$ the variable $W_t - W_s$ is independent of $\mathcal{F}_s$. The term $(\Omega, \mathcal{F}, (\mathcal{F}_t), U, W, I\!P)$ will be called an *SDE basis.* Then a *solution-process* of (2.1) on this SDE basis is any $I\!R^d$-valued process X which is continuous in time, adapted to the filtration $(\mathcal{F}_t)$, and which satisfies the following (written component-wise):

$$X_t^i \;=\; U^i + \int_0^t a(s, X_s)^i ds + \sum_{j=1}^{d'} \int_0^t \sigma(s, X_s)^{i,j} dW_s^j, \qquad i = 1, \ldots, d. \tag{2.2}$$

The second integrals above are stochastic integrals with respect to the 1-dimensional Wiener processes W^j, uniquely defined up to a null set only; of course, writing (2.2) supposes that all the integrals make sense. When the filtration $(\mathcal{F}_t)$ is the one "generated by" the Wiener process W and the initial condition U, then a solution-process is called a *strong solution.*

Let us now recall a set of hypotheses which yields *existence and uniqueness* of a solution process for our SDE (the uniqueness is to be understood up to null sets, that is if X and X' are two solutions, then the set of all ω for which the paths $t \mapsto X_t(\omega)$ and $t \mapsto X'_t(\omega)$ do not agree is of probability 0). These hypotheses are:

(L) Local Lipschitz condition: For all $T > 0$, $K > 0$ there is a constant $C(T, K)$ such that (with $|.|$ being the Euclidian norm on any relevant space)

$$t \in [0, T], \;\; |x|, |y| \leq K \quad \Rightarrow \quad |a(t,x) - a(t,y)| + |\sigma(t,x) - \sigma(t,y)| \leq C(T,K)|x - y| \tag{2.3}$$

(G) Linear growth condition: For all $T > 0$ there is a constant $C(T)$ such that

$$t \in [0, T] \quad \Rightarrow \quad |a(t,x)| + |\sigma(t,x)| \leq C(T)(1 + |x|). \tag{2.4}$$

If (L) holds but (G) fails, then a (unique) solution exists up to the "explosion" time: there is a stopping time T such that (2.2) holds for $t < T$, and $\limsup_{t \uparrow T} |X_t| = \infty$ on the set $\{T < \infty\}$.

So far, the Wiener process W and the initial condition U were given. In financial applications, though, the actual Wiener process does not really matter. Similarly the actual variable U has no importance, only its law μ matters. Hence it is meaningful to speak about X when the coefficients a and c and the law μ of the initial condition are given, without reference to any pre-defined Wiener process W and random variable U. Mathematically, this means that we are interested in the *law* of X. And, in statistics also, one is usually interested in the laws of the data and not in the actual probability space on which these data are defined.

A *weak solution* to Equation (2.1) is the law of any solution-process of this equation. As the law of any random variable, the law of X will be a probability measure on the space in which X takes its values. So, let us denote by Ω the space of all continuous functions from $\mathbb{R}_+$ into $\mathbb{R}^d$ (here d is fixed and does not appear in our notation). We endow this space with the so-called "canonical process" Y, defined by $Y_t(\omega) = \omega(t)$ when $\omega \in \Omega$, and with the Kolmogorov σ-field $\mathcal{Y} = \sigma(Y_t : t \geq 0)$, and with the canonical filtration $\mathcal{Y}_t = \cap_{s>t}\sigma(Y_r : r \leq s)$. Then if X is a solution-process on any given SDE basis, its law is a **weak solution**. The law μ of the initial condition U (which is also the law of Y_0 under the weak solution) is called the "initial condition" of the weak solution.

Studying weak solutions seems to be a rather difficult task, since a priori different solution-processes, possibly defined on different spaces, lead to different weak solutions. However, due to two remarquable results, weak solutions are indeed quite tractable:

Theorem 2.1 (Yamada and Watanabe) *Let μ a probability measure on $\mathcal{R}^d$, and suppose that, on any SDE basis such that $\mathcal{L}(U) = \mu$, Equation (2.1) admits a unique solution-process. Then the weak solution with initial condition μ (which of course exists !) is unique.* [This holds in particular under (L) and (G).]

Theorem 2.2 (Stroock and Varadhan) *Suppose that the coefficients a and σ satisfy the linear growth condition (G) and are continuous in x, and also that the matrix $c(t,x) = \sigma(t,x)\sigma(t,x)^*$ is everywhere invertible. Then for any probability measure μ on $\mathbb{R}^d$ there is one and only one weak solution with initial condition μ.*

2.2 The Markov property and the infinitesimal generator

In the sequel we assume that our SDE admits, for every initial measure μ, a unique weak solution. Denote by $\mathbb{P}_\mu$ the weak solution associated with the initial measure μ. A crucial property of our diffusion is that the process Y is, under each $\mathbb{P}_\mu$, a Markov process, Of course it is in general non-homogeneous, and it becomes homogeneous when the coefficients a and c do not depend on time.

We will denote by $(P_{s,t})_{0\leq s\leq t}$ the (non-homogeneous) transition semi-group, that is $P_{s,t}(x,.)$ is the law of Y_t, under each $\mathbb{P}_\mu$, conditionally on the fact that $Y_s = x$. In the homogeneous case we get a one-parameter semi-group, denoted by $(P_t)_{t\geq 0}$.

Another characteristic of our diffusion is its infinitesimal generator, which is useful mainly in the homogeneous case. So, assuming that the coefficients a and c do not depend on time, we introduce the following second order elliptic operator:

$$Af(x) = \sum_{i=1}^{d} a(x)^i \frac{\partial}{\partial x_i} f(x) + \frac{1}{2}\sum_{i,j=1}^{d} c(x)^{i,j} \frac{\partial^2}{\partial x_i \partial x_j} f(x). \tag{2.5}$$

Then one can show that a probability measure $\mathbb{P}$ on the canonical space is a weak solution to our SDE if and only if, for any twice continuously differentiable function f on $\mathbb{R}^d$, the following processes

$$M_t^f = f(Y_t) - f(Y_0) - \int_0^t Af(Y_s)ds \tag{2.6}$$

are local martingales under $I\!P$.

This "martingale characterization" of weak solutions is most useful (a similar statement holds in the non-homogeneous case). We will call the operator A the **infinitesimal generator** of the diffusion process, although this is a slight abuse of terminology (our operator A is more like the so-called "extended generator" of Kunita, except that we do not bother here about the actual domain of this unbounded linear operator, using only the fact that C^2 functions are in the domain).

2.3 Examples - Diffusions on a domain

Below we list a number of examples, some of them having an explicit solution in terms of the driving Wiener process: this is rather rare, but such situations are worth mentionning because they provide some of the most commonly used diffusion processes in finance, and also they provide simple case studies in which various statistical procedures can be tested. All these simple examples below concern the 1-dimensional case, $d = d' = 1$. The reader will observe that in all examples there are parameters coming naturally within the coefficients.

Example 1) Wiener process with drift: This is Equation (2.1) with the coefficients $a(x) = \mu$ and $\sigma(x) = \sigma$, where $\mu \in I\!R$ and $\sigma > 0$ are given constants. The "solution" is of course $X_t = X_0 + \mu t + \sigma W_t$.

Example 2) Geometric Brownian motion, or Black-Scholes model: This is Equation (2.1) with the coefficients $a(x) = \mu x$ and $\sigma(x) = \sigma x$, where $\mu \in I\!R$ and $\sigma > 0$:

$$dX_t = \mu X_t dt + \sigma X_t dW_t, \qquad X_0 = U. \tag{2.7}$$

This equation is a "linear" equation, which admits an "explicit" solution given by

$$X_t = U \exp\left(\left(\mu - \frac{\sigma^2}{2}\right) t + \sigma W_t\right). \tag{2.8}$$

Example 3) Ornstein-Uhlenbeck process: This is the equation

$$dX_t = -\mu X_t dt + \sigma dW_t, \qquad X_0 = U, \tag{2.9}$$

where $\mu \in I\!R$ and $\sigma > 0$. This is again a linear equation with an explicit solution given by

$$X_t = U e^{-\mu t} + \sigma \int_0^t e^{-\mu(t-s)} dW_s. \tag{2.10}$$

The above stochastic integral is a Wiener integral (the integrand is a deterministic function), so it gives a centered normal variable with variance $\int_0^t e^{-2\mu(t-s)} ds = (1 - e^{-2\mu t})/2\mu$ if $\mu \neq 0$, and t if $\mu = 0$.

Example 4) The Vasicek model: This is the equation

$$dX_t = \mu(\nu - X_t) dt + \sigma dW_t, \qquad X_0 = U, \tag{2.11}$$

where $\mu, \nu \in \mathbb{R}$ and $\sigma > 0$. This generalizes the Ornstein-Uhlenbeck Equation and is again a linear equation with an explicit solution given by

$$X_t = Ue^{-\mu t} + \sigma \int_0^t e^{-\mu(t-s)} dW_s + \nu(1 - e^{-\mu t}). \tag{2.12}$$

Example 5) The Cox-Ingersoll-Ross model: This is the equation

$$dX_t = \mu(\nu - X_t)dt + \sigma\sqrt{X_t}dW_t, \qquad X_0 = U, \tag{2.13}$$

where $\mu, \nu \in \mathbb{R}$ and $\sigma > 0$ and U is a positive random variable. In this case (G) is satisfied but not (L); however a "version" of (L) holds on $(0, \infty)$, namely (2.3) holds for x and y in any compact subset K of $(0, \infty)$, and we have a unique solution-process up to the first time when this solution hits 0. In other words, there is a unique solution-process on the whole half-line as soon as we are sure that this solution never hits 0: this is the case iff $2\mu\nu > \sigma^2$ and $\mu > 0$.

In the last example it is crucial that the solution remains positive. In the Black-Scholes model, we see from (2.8) that, if the initial condition U is positive, then X remains always positive: in both cases we can consider that the state space is $(0, \infty)$ instead of $\mathbb{R}$.

These are examples where the solution takes its values in a domain D of $\mathbb{R}^d$. Considering for example Equation (2.1) on such a domain D means that the functions a and σ are defined on $\mathbb{R}_+ \times D$ and that the solution takes its values in D. Two extreme cases are possible:

1 The domain D is open: one extends a and σ over the whole set $\mathbb{R}_+ \times \mathbb{R}^d$ in an arbitrary fashion, and consider initial conditions U taking values in D only. One solves the equation in $\mathbb{R}^d$ and with some luck the solution will not leave D. However, it is not always easy (nor even possible) to extend a and σ in such a way that (L) and (G) or the conditions for Theorem 2.2 hold, and further there is no general criterion yielding that X will stay in D; only ad-hoc arguments for each special case, as for the Cox-Ingersoll-Ross model, will (sometimes) do the job.

2 The domain D is closed. It is then more difficult because one has to specify what happens when the solution X hits the boundary ∂D: it can reflect instantaneously towards the interior of D, or stick for a while on the boundary ∂D at the hitting point, or diffuse over the boundary itself for a while before bouncing back inside the interior of D. All these behaviours necessitate additional requirements.

And, of course, there are "mixed" cases, where D is neither open nor closed, and the specifications of the process are even harder in these cases... However, in Case 1 above, all what we have said for diffusions over $\mathbb{R}^d$ remains true for diffusions over an open domain as soon as we know that the process never exits the domain. As a matter of fact, in this paper we will always assume that we are in Case 1, whenever the state space is a domain D of $\mathbb{R}^d$: that is, we can and will *always* do as if the state space were the whole of $\mathbb{R}^d$ (this is an important remark, because when a diffusion hits its boundary the statistical properties

might be radically modified, and in particular the rates of convergence of estimators might be greatly improved, or on the opposite the asymptotic variance of the estimators can be increased...). That means that for the Cox-Ingersoll-Ross process for example, we restrict our attention to the set of parameters $\{(\mu, \nu, \sigma) : \mu > 0,\ 2\mu\nu > \sigma^2\}$.

2.4 Likelihood ratio

One of the main tools in statistics is the likelihood ratios for the solutions of Equation (2.1) associated with various coefficients.

Let us consider two sets of coefficients (a, σ) and (a', σ') with the same dimensionality, and $c = \sigma\sigma^*$ and $c' = \sigma'\sigma'^*$, and consider solutions X and X' corresponding to these coefficients, and starting at the same point x_0 for simplicity. The likelihood ratio "of X' w.r.t. X " is the likelihood ration (or Radon-Nikodym derivative) of the law $I\!P'$ of X' w.r.t. the law $I\!P$ of X: that is, we consider the weak solutions $I\!P$ and $I\!P'$ of our equations, and we want to compute the likelihood ratio in terms of the coefficients.

Two preliminary remarks are in order: first, as seen for example in Theorem 2.2, the weak solution $I\!P$ depend on a and c, but not on σ itself (we may have different functions σ with the same "square" $\sigma\sigma^*$): so the likelihood ration can at the best be expressed in terms of a, a', c, c': for example the two "equations" $dX_t = dW_t$ and $dX'_t = -dW_t$ have the same unique weak solution, and thus we cannot discriminate between the two equations upon observing even the whole processes X and X'. Second, we obviously need that $I\!P$ and $I\!P'$ be completely characterized by the pairs (a, c) and (a', c'), that is we need existence and uniqueness of the weak solutions to our two equations. And, the likelihood ratio will be computed *on the canonical space* $(\Omega, \mathcal{Y})$.

Next, an extremely important observation: before computing any likelihood ratio, we need to specify on which σ-field this ratio will be computed (in statistical terms: what is the form of the actual observations). This is because on the largest σ-field $\mathcal{Y}$ the measures $I\!P$ and $I\!P'$ are – typically – mutually singular: in statistical terms, if the whole process is observed (up to infinity!) then one can discriminate for sure between (a, c) and (a', c').

Although one may think of several other possibilities (some of them considered later on), we mainly consider two main schemes:

(1) The σ-field is $\mathcal{G} = \mathcal{Y}_T$ for some given $T < \infty$; this corresponds to observing the path of the solution over the interval $[0, T]$.

(2) The σ-field is $\mathcal{G} = \sigma(Y_{t_i} : i = 0, 1, \ldots, n)$ for some times $0 \leq t_0 < t_1 < \ldots < t_n$: this corresponds to observing the solution at discrete times t_i.

Let us consider first (1). If we do not have $c' = c$, or at least if the two processes $c(t, Y_t)$ and $c'(t, Y_t)$ do not coincide $I\!P$-almost surely on the interval $[0, T]$, then $I\!P'$ is not absolutely continuous w.r.t. $I\!P$ on the σ-field $\mathcal{G}$: so for simplicity we assume that $c' = c$. We also need that for each (t, x) the vector $a'(t, x) - a(x, t)$ be of the form

$$a'(t, x) - a(x, t) = c(t, x)b(t, x) \tag{2.14}$$

for some measurable vector-valued function b which is such that the integrals in the next formula below make sense. Then the likelihood ratio of $\mathbb{P}'$ w.r.t. $\mathbb{P}$, in restriction to $\mathcal{G} = \mathcal{Y}_T$, takes the form

$$Z_T = \exp\left\{\int_0^T b(t, Y_t)^* dY_t - \frac{1}{2}\int_0^T b(t, Y_t)^\star (a(t, Y_t) + a'(t, Y_t))dt\right\}. \tag{2.15}$$

Next we consider (2). The σ-field $\mathcal{G}$ is much smaller now, so it is much easier for $\mathbb{P}'$ to be absolutely continuous w.r.t. $\mathbb{P}$ in restriction to $\mathcal{G}$: in particular we no longer need something like $c' = c$. In fact, as soon as for instance c is invertible on the domain D where the diffusion process lives (an hypothesis usually satisfied by financial models), the transition semi-group admits positive densities w.r.t. Lebesgue measure on D: that means that the measures $P_{s,t}(x, .)$ admit positive probability densities $y \mapsto p_{s,t}(x, y)$ for all $x \in D$ and all $s < t$.

Suppose that our two diffusion processes live on the same domain D, with transition semi-groups admitting positive densities $p_{s,t}$ and $p'_{s,t}$ respectively. Then, due to the Markov structure, the likelihood ratio in restriction to $\mathcal{G} = \sigma(Y_{t_i} : i = 0, 1, \ldots, n)$ is

$$Z = \prod_{i=1}^{n} \frac{p'_{t_{i-1},t_i}(Y_{t_{i-1}}, Y_{t_i})}{p_{t_{i-1},t_i}(Y_{t_{i-1}}, Y_{t_i})}. \tag{2.16}$$

3 Parametric estimation: asymptotic optimality criteria

Let us now come to statistical estimation, from the asymptotic point of view. We have a parameter set $\Theta \subset \mathbb{R}^q$, and for each $\theta \in \Theta$ a probability measure $\mathbb{P}_\theta$ on our basic space $(\Omega, \mathcal{Y})$ (for diffusions, each $\mathbb{P}_\theta$ is the weak solution of an equation (2.1) with coefficients depending on the value θ).

Typically (at least in the diffusion setting) one does not observe the whole σ-field $\mathcal{Y}$, but some sub-σ-field $\mathcal{G}_n$: here n stands for the "number" of available data, it is large, and we are looking at what happens when $n \to \infty$. More precisely, we want to construct for each n an estimator $\widehat{\theta}_n$ for θ, in such a way that the sequence $(\widehat{\theta}_n)_n$ behaves as well as possible when n grows. This question of asymptotic optimality in estimation for general statistical models was taken on essentially by LeCam (see [26] or LeCam and Yang [27]; see also the book of Ibragimov and Khashminski [13]).

We suppose that all measures $\mathbb{P}_\theta$ are equivalent on the σ-field $\mathcal{G}_n$, and we use the following notation for the likelihood ratios:

$$Z_n(\zeta/\theta) = \frac{d\mathbb{P}_\zeta}{d\mathbb{P}_\theta\,|\mathcal{G}_n}.$$

Let us assume that the "true" value of the parameter is θ, some point in the interior of $\Theta \subset \mathbb{R}^q$. For each ζ the sequence of random variables $(Z_n(\zeta/\theta))_n$ is tight, so it is not a drastic assumption to assume that these sequences converge in law (under $\mathbb{P}_\theta$). Two extreme phenomena can arise:

1) For all ζ in Θ the limit of $Z_n(\zeta/\theta)$ under $I\!P_\theta$ is a strictly positive variable; then "in the limit" we still have a statistical model where all measures are equivalent. For diffusions this arises for example if at stage n one observes the values $Y_{i/n}$ for $i = 0, \ldots, n$, and when all measures $I\!P_\zeta$ are equivalent on the σ-field $\mathcal{Y}_1$ (typically when the diffusion coefficient c does not depend on the parameter): in the limit we have the full observation of the diffusion over the time interval $[0,1]$, and the likelihood is then given by (2.15).

2) For any ζ the sequence $Z_n(\zeta/\theta)$ goes to 0 in $I\!P_\theta$-measure: that means that "in the limit" the measures become mutually singular and a "perfect" estimation becomes possible. For diffusions this arises for example if at stage n one observes the values $Y_{i/n}$ for $i = 0, \ldots, n$, and when all measures $I\!P_\zeta$ are mutually singular on the σ-field $\mathcal{Y}_1$ (typically when the diffusion coefficient $c(\theta, .)$ are distinct for different values of θ): in the limit we have the full observation of the diffusion over the time interval $[0,1]$, which gives us the function c, hence the value θ.

In case (1) above there is nothing more to say, except to wish good luck to the statistician (observe that there is then no consistent sequences of estimators). In case (2), contrarywise, we can go much further since it is possible to find weakly consistent sequences of estimators $(\widehat{\theta}_n)_n$: this means that $\widehat{\theta}_n \to \theta$ in $I\!P_\theta$-probability for any θ. Then one can look for rates of convergence, and this is what the so-called "local behaviour" around the "true" value θ is all about. More precisely consider a sequence u_n going to 0 (and which may depend of course on the value θ). Here again several situations are possible:

(i) $u_n \to$ "slowly": then $Z_n(\theta + u_n h/\theta)$ goes to 0 in $I\!P_\theta$-measure and, exactly as in (2) above one can "asymptotically" estimate perfectly the parameter at the scale u_n.

(ii) $u_n \to 0$ "fast": then $Z_n(\theta + u_n h/\theta)$ goes to 1 in $I\!P_\theta$-measure. This means that the measures $I\!P_\theta$ and $I\!P_{\theta+u_n h}$ become more and more indistinguishable as n increases and we cannot do any sensible estimation at the scale u_n.

(iii) In between, there is hopefully a choice of u_n such that for any choice $h_1, \ldots, h_r$ of vectors in $I\!R^q$ the sequence $(Z_n(\theta + u_n h_i/\theta)_n)_{1 \le i \le r}$ converges in law under $I\!P_\theta$ to a limit whose components take their values in $(0, \infty)$ and have expectation equal to 1.

Suppose now that we can find a sequence u_n such that (iii) above holds. The value θ is fixed here. One can find a statistical model $\mathcal{B}' = (\Omega', \mathcal{G}', (I\!P'_h)_{h \in I\!R^q})$ (everything depends on θ) such that the measures $I\!P'_h$ are all equivalent, and the likelihood ratios $Z'(h/0) = dI\!P'_h/dI\!P'_0$ are limits in law of the variables $Z(\theta + u_n h/\theta)_n$ under $I\!P_\theta$. Then one says that the *local models* $\mathcal{B}_n^\theta = (\Omega, \mathcal{G}_n, (I\!P_{\theta+u_n h})_{h \in I\!R^q})$ *converge weakly* to $\mathcal{B}'$.

In the limit we can identify $\mathcal{B}_n^\theta$ with $\mathcal{B}'$, and LeCam showed interesting properties, which we state in a rather heuristic way: if $\widehat{h}$ is an estimator of h for $\mathcal{B}'$, there is a sequence $\widehat{\theta}_n$ of estimators for θ such that $\frac{1}{u_n}(\widehat{\theta}_n - (\theta + u_n h))$ converges in law under $I\!P_{\theta+u_n h}$ towards the law of $\widehat{h} - h$ under $I\!P'_h$ for any h; conversely for any sequence of estimators $\widehat{\theta}_n$ such that the sequence $\frac{1}{u_n}(\widehat{\theta}_n - (\theta + u_n h))$ converges in law under $I\!P_{\theta+u_n h}$ to a variable U_h for all h, then there exists an estimator $\widehat{h}$ on $\mathcal{B}'$ such that the law of U_h is the same as the

law of $\widehat{h} - h$ under $I\!P'_h$. Further, at least in a neighbourhood of θ (shrinking to 0 at speed u_n), the "asymptotically best" estimators $\widehat{\theta}_n$ converge to the true value of the parameter with the rate $1/u_n$ and $\frac{1}{u_n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P_\theta$ to the "best" estimator $\widehat{h}$ of h at point 0 for the model $\mathcal{B}'$, if such a best estimator exists.

In particular if the weak convergence of local models holds at any point θ, with a rate $u_n(\theta)$ which in principle may depend on θ, we will say that a sequence $\widehat{\theta}_n$ of estimators is *rate-efficient* if for any value θ the sequence $\frac{1}{u_n(\theta)}(\widehat{\theta}_n - \theta)$ is tight under $I\!P_\theta$.

Remark: When the parameter is multi-dimensional, there is also the possibility that the rate differs according to the components of the parameter. So we can do the same analysis with a q-dimensional vector u_n whose components decrease to 0, and the change of parameter becomes, componentwise: $\zeta^i = \theta^i + u_n^i h^i$, for $i = 1, \ldots, q$. We will see such an example for ergodic diffusions later.

3.1 The LAN property

Finding rate-efficient estimators is good, but in some cases we can even go further: when it is possible to derive "best" estimators for the limiting model $\mathcal{B}'$ we can in principle find accordingly "best" estimators, in the asymptotic sense, for the original model.

The simplest of these cases is by far when $\mathcal{B}'$ is the so-called *Gaussian shift*: let I be an invertible symmetric $q \times q$-matrix); the associated Gaussian shift experiment consists in taking $\Omega' = I\!R^q$ and $\mathcal{G}' = \mathcal{R}^q$ (the Borel σ-field) and $I\!P'_h = \mathcal{N}(h, I^{-1})$ (the Gaussian distribution with mean h and covariance matrix I^{-1}. With the notation $X(\omega') = \omega'$, we thus have

$$Z'(h/0) = \exp\left(h^* I X - \frac{1}{2}\, h^* I h\right). \tag{3.1}$$

Observe that the matrix I is also the Fisher information matrix of the model $\mathcal{B}'$, for all values $h \in I\!R^q$.

We will say that *the LAN ("local asymptotic normality") property holds at point θ, with rate u_n,* if the sequence of local models $\mathcal{B}^\theta_n$ around θ converges weakly to the Gaussian shift experiment $\mathcal{B}'$ described above. Of course the matrix I depends on θ, and is usually written $I = I(\theta)$.

Due to the Gaussian property of the limit, we have LAN at point θ as soon as the following convergence in law holds true:

$$Z_n(\theta + u_n h/\theta) \;\to^{\mathcal{L}(I\!P_\theta)}\; \exp\left(h^* X - \frac{1}{2}\, h^* I(\theta) h\right) \tag{3.2}$$

for all h. Equivalently, we have LAN as soon as we can write

$$\log Z_n(\theta + u_n h/\theta) \;=\; h^* U_n - \frac{1}{2}\, h^* \Gamma_n(\theta) h + R_n(h) \tag{3.3}$$

where $\Gamma_n(\theta) \to I(\theta)$ and $R_n(h) \to 0$ in $I\!P_\theta$-probability and U_n converges in law under $I\!P_\theta$ to $\mathcal{N}(0, I(\theta))$.

Now the model $\mathcal{B}'$ is the simplest of all possible models, for which the best estimator for h (recall that $I(\theta)$ is known) is obviously $\widehat{h} = X$, in all possible senses of "best". Moreover under $I\!P'_h$, the variable $\widehat{h} - h$ has the law $\mathcal{N}(0, I(\theta)^{-1})$.

Therefore if the LAN property holds at a point θ, a sequence $\widehat{\theta}_n$ will be asymptotically optimal in a neighbourhood of θ if

$$\left.\begin{array}{ll} \frac{1}{u_n}(\widehat{\theta}_n - \theta) \ \rightarrow^{\mathcal{L}(I\!P_\theta)} \ \mathcal{N}(0, I(\theta)^{-1}), & \text{or equivalently} \\ \frac{\sqrt{\Gamma_n(\theta)}}{u_n}(\widehat{\theta}_n - \theta) \ \rightarrow^{\mathcal{L}(I\!P_\theta)} \ \mathcal{N}(0, I_q) & \end{array}\right\} \tag{3.4}$$

where I_q is the $q \times q$ identity matrix. Observe that these estimators achieve asymptotically the Cramer-Rao bound for the estimation variance. Moreover, such estimators will also satisfy for all h:

$$\frac{1}{u_n}(\widehat{\theta}_n - (\theta + u_n h)) \ \rightarrow^{\mathcal{L}(I\!P_{\theta+u_n h})} \ \mathcal{N}(0, I(\theta)^{-1}) \tag{3.5}$$

as well. Finally, such estimators are, in principle, easy to get: it suffices to set $\widehat{\theta}_n = \Gamma_n(\theta)^{-1} U_n$; however in practice finding $\Gamma_n(\theta)$ and U_n is quite a different matter !

Sequences of stimators having the property (3.4) are called *asymptotically efficient* around θ, and simply "asymptotically efficient" if this holds for each θ (with $u_n = u_n(\theta)$ if it happens that u_n actually depends on θ).

3.2 LAMN, LAQ

Of course there are other limiting models than Gaussian shifts. Two of them are of particular interest, and are sometimes obtained when dealing with diffusion processes:

Suppose that the likelihood ratios of the local model $\mathcal{B}_n^\theta$ satisfy (3.3). Suppose also that the pair $(U_n, \Gamma_n(\theta))$ converge in law, under $I\!P_\theta$, to a limit $(U, I(\theta))$, and that the (random) matrix $I(\theta)$ is everywhere invertible. Then

a) We have *the LAMN ("local asymptotic mixed normality") property at point θ with rate u_n* if further we can write $U = I(\theta)^{1/2} U'$ where U' is independent of $I(\theta)$ and distributed according to $\mathcal{N}(0, I_q)$. The matrix $I(\theta)$ is called the *random Fisher information matrix.*

b) We have *the LAQ ("local asymptotic quadraticity") property at point θ with rate u_n* if further for any $h \in I\!R^q$ we have

$$E\left(e^{h^* U - \frac{1}{2} h^* I(\theta) h}\right) = 1. \tag{3.6}$$

"Quadraticity" means that the log-likelihood is approximately a quadratic form in h (in (3.3)), while mixed normality means that the variable U has a mixed Gaussian distribution. Obviously LAQ $\Rightarrow$ LAMN $\Rightarrow$ LAN.

The LAMN property was introduced by Jeganathan [19], see also [27]. The LAQ property has been introduced by a number of different authors: see LeCam and Yang [27], Shiryaev, Spokoiny, etc...

Exactly as for LAN, if the LAMN property holds at a point θ, a sequence $\widehat{\theta}_n$ will be asymptotically optimal in a neighbourhood of θ if we have the analogue of (3.4) (note that the first line in (3.4) makes no sense here):

$$\frac{\sqrt{\Gamma_n(\theta)}}{u_n}(\widehat{\theta}_n - \theta) \ \to^{\mathcal{L}(I\!P_\theta)} \ \mathcal{N}(0, I_q) \tag{3.7}$$

Equivalently, one says: the sequence $\frac{1}{u_n}(\widehat{\theta}_n - \theta)$ converges in law, under $I\!P_\theta$, towards a centered mixed Gaussian variable, with condional covariance matrix $I(\theta)^{-1}$.

4 Diffusions and Statistics

Let us now come back to our diffusion processes. We have a parameter set $\Theta \subset I\!R^q$ and, for each $\theta \in \Theta$, a pair $(a(\theta,.), \sigma(\theta,.))$ of coefficients with the same dimensionality, and a given starting point x_0. We set $c(\theta,.) = \sigma(\theta,.)\sigma(\theta,.)^\star$. We suppose that the equation

$$dX_t = a(\theta, t, X_t)dt + \sigma(\theta, t, X_t)dW_t, \qquad X_0 = x_0 \tag{4.1}$$

has a unique weak solution $I\!P_\theta$ for every $\theta \in \Theta$. Our statistical model is $(\Omega, \mathcal{Y}, (I\!P_\theta)_{\theta\in\Theta})$, where Ω is the canonical space with the canonical process Y. Recall that according to the end of Subsection 2.3, we assume that the state space of our diffuions is the whole of $I\!R^d$. Taking a deterministic and known starting point x_0 is just for convenience, most of what follows accomodates random initial conditions (in most observation schemes anyway the value of the process at time 0 is observed and may thus be considered as known).

We also have a set of data, which generates a sub-σ-field $\mathcal{G}_n$ of $\mathcal{Y}$, where n here stands for the size of the data set. On the basis of these data we want to construct an estimate $\widehat{\theta}_n$ for θ, and we are particularly interested in the asymptotic optimality as $n \to \infty$. As a matter of fact, estimation procedures and even rates of convergence may greatly differ for the drift coefficient and for the diffusion coefficient. So it might be useful to label the parameters on which a and σ depend with different symbols. This leads to write the equation as

$$dX_t \ = \ a(\theta_1, X_t)dt + \sigma(\theta_2, X_t)dW_t, \qquad X_0 = x_0. \tag{4.2}$$

The full parameter is then $\theta = (\theta_1, \theta_2)$. The two equations (4.1) and (4.2) are two ways of writing the same thing, and the most convenient one depends on the problem at hand, and especially on the structure of the set Θ: in the extreme case where $\theta_1 = \theta_2$ for all $(\theta_1, \theta_2) \in \Theta$ one prefers (4.1), while (4.2) is handier in the other extreme case where $\Theta = \Theta_1 \times \Theta_2$.

Since we want to have at least consistent sequences of estimators, we obviously need a minimal *identifiability assumption*, which can be expressed as follows: the measures $I\!P_\theta$ should be mutually singular for different values of θ, in restriction to the σ-field $\mathcal{G}_\infty = \bigvee_n \mathcal{G}_n$ which represents the biggest possible observed σ-field (note that the sequence $(\mathcal{G}_n)$ is not necessarily increasing, though). Of course this is not fully satisfactory, since checking this property on the coefficients a and c themselves is not always a trivial matter. So in practice we sometimes impose more restrictive identifiability assumptions, which are not necessary but much handier than the minimal one.

Let us now quickly review below a number of more or less commonly encountered observation schemes, which amounts to specify the observed σ-field $\mathcal{G}_n$. Some of these schemes are studied more leisurely in the forthcoming sections. But, all throughout we apply the methods to a special case in order to make comparisons, namely to the Ornstein-Uhlenbeck process because, due to its Gaussian properties, it is particularly amenable to explicit computations. In accordance with the notation of (4.2), we write it as

$$dX_t = -\theta_1 X_t dt + \sqrt{\theta_2} dW_t, \qquad X_0 = x_0, \tag{4.3}$$

and the natural parameter space is $\Theta = \Theta_1 \times \Theta_2$, with $\Theta_1 = \mathbb{R}$ and $\Theta_2 = (0, \infty)$.

4.1 Observation over a whole interval

A) The mathematically simplest case is when the whole path $t \mapsto Y_t$ is observed over an interval $[0, T_n]$: that is $\mathcal{G}_n = \mathcal{Y}_{T_n} := \sigma(Y_s : 0 \leq t \leq T_n)$. The theory has been established mainly by Kutoyants (see [25] and many subsequent papers by this author).

This observation scheme has an immediate and obvious drawback: it is *stricto sensu* impossible to achieve, since any conceivable mean of observation will end up with a finite set of numbers. And even if this set of numbers is very large, it is difficult to obtain a good approximation of the path $t \mapsto X_t$, which is quite irregular: it is Hölder continuous with arbitrary index $\alpha < 1/2$, but not with index $1/2$. Nevertheless, it is in principle possible to achieve an approximation of the path which is as good as one wishes by, say, discrete observations on a grid with sufficently small mesh. So, even if this continuous-time observation scheme is not feasible strictly speaking, it can be viewed as an idealization of real observation schemes. In this sense it has a lot of mathematical interest, because it gives an "upper limit" of what can be achieved by observing X an any (regular or irregular) grid inside the interval $[0, T_n]$.

As seen in Subsection 2.4, the measures $\mathbb{P}_\theta$ are mutually singular if the diffusion coefficients $c(\theta, .)$ differ for distinct values of θ. And, by computing the quadratic variation $\int_0^{T_n} c(\theta, s, Y_s) ds$ (a theoretically possible computation if the whole path is known), we can compute exactly the true value of θ. In other words the statistical problem is completely solved, with no estimation error.

The situation is totally different for the drift coefficient. So let us assume that $c(\theta, .) = c(.)$ does not depend on θ, and further that it is invertible and that the functions $b_{\zeta/\theta}(.) = c(.)^{-1}(a(\zeta, .) - a(\theta, .))$ are, say, locally bounded for all θ, ζ. Then the measures $\mathbb{P}_\theta$ are all equivalent on $\mathcal{G}_n$, and the likelihood $Z_n(\zeta/\theta)$ is given by (2.15) with $b_{\zeta/\theta}$ instead of b and $T = T_n$.

The asymptotic is then when $T_n \to \infty$: for getting any kind of results we need some nice behaviour of the diffusion process at infinity. In practice, we need our diffusion processes to be *homogeneous and ergodic*, with an invariant probability measure μ_θ whose support is $\mathbb{R}^d$ (or, the domain D over which all our diffusions live). Then, using the ergodic theorem and the associated central limit theorem, and if further the matrix c is everywhere invertible, one can prove that, under mild smoothness assumptions on the coefficients, *we*

have the LAN property with rate $1/\sqrt{T_n}$, and with asymptotic Fisher information matrix

$$I(\theta)_{i,j} = \int \frac{\partial}{\partial\theta_i} a(\theta,x)^* c(x)^{-1} \frac{\partial}{\partial\theta_j} a(\theta,x)\mu_\theta(dx). \tag{4.4}$$

Further the maximum likelihood estimators (MLE) are asymptotically efficient, as soon as for example the parameter set Θ is compact. Using the explicit expression (2.15) we can in principle compute the MLE, but this involves computing two integrals, one being a stochastic integral; for this one needs to do some approximation, like a Riemann type approximation, and it is difficult to keep track of the errors introduced through these approximations: a lot of papers have been devoted to methods allowing practical approximations of the likelihood or to alternative methods.

But, appart from the drawback stated at the beginning and from the difficulty of concrete calculations involving (2.15), one should emphasize the assumption that our diffusions are ergodic: the examples 3, 4 and 5 of Section 2 have this property if and only if the parameter μ is positive, while the Black-Scholes diffusion is never ergodic; more important even, all these examples are 1-dimensional, but if the diffusion is multi-dimensional it is much more difficult to have ergodic properties. And further, it is very unlikely that accurate models in finance can be at all ergodic (or even homogeneous) since there are obvious trends, at least for assets prices: so modelling with a Vasicek model or a Cox-Ingersoll-Ross model can be good only over a finite horizon, and this is of course totally contradictory with the fact that $T_n \to \infty$ above.

The Ornstein-Uhlenbeck process: For the process (4.3), with θ_2 fixed and θ_1 unknown, we can write explicitely the likelihood ratio and find the MLE, which takes the form

$$\widehat{\theta}_{1,n} = -\int_0^{T_n} Y_s dY_s \;/\; \int_0^{T_n} Y_s^2 ds. \tag{4.5}$$

The process is ergodic if and only if $\theta_1 > 0$, in which case the stationary measure is $\mu_\theta = \mathcal{N}(0, \theta_2/2\theta_1)$. As said before, we have the LAN property with rate $1/\sqrt{T_n}$ at each point $\theta_1 > 0$, and the asymptotic Fisher information is $I(\theta) = 1/2\theta_1$, and of course the MLE is then asymptotically efficient.

What is interesting here is that we can also derive the local asymptotic properties of this model at the points $\theta_1 \leq 0$:

a At point $\theta_1 = 0$ we have the LAQ property with rate $1/T_n$, and the variable $I(\theta)$ is (3.6) has the law of the variable $\int_0^1 W_s^2 dW_s$, where W is a Brownian motion. Observe that for $\theta_1 = 0$ the diffusion is just a non-standard Brownian motion, recurrent but not ergodic.

b At all points $\theta_1 < 0$ we have the LAMN property with rate $e^{\theta_1 T_n}$, and the conditional Fisher information $I(\theta)$ has the law of the square of an $\mathcal{N}(\left(\frac{x_0}{\sqrt{-2\theta_1\theta_2}}, \frac{1}{4\theta_1^2}\right)$ random variable. Observe that for $\theta_1 < 0$, the diffusion is transient: this explains why the starting point x_0 has an impact on the asymptotic behaviour.

In all cases the MLE is asymptotically efficient. But the rate of convergence of $\widehat{\theta}_{1,n}$ to θ_1, which is $\sqrt{T_n}$ in the ergodic case, is *much faster* in the other cases, especially in

the transient case. This is very specific to the O-U process, and for other transient or null-recurrent diffusions essentially nothing is known, and in particular not the rates of convergence if they exist at all.

B) Another related problem. A closely related asymptotic problem is as follows. Instead of (4.1) we consider the equation

$$dX_t = a(\theta, X_t)dt + \varepsilon_n \sigma(X_t)dW_t, \qquad X_0 = x_0, \qquad t \in [0,T]. \tag{4.6}$$

Here, T is fixed and σ is a known function and the known parameter ε_n is supposed to be small, so the above equation is a "noisy" version of the ordinary differential equations

$$dX_t = a(\theta, X_t)dt, \qquad X_0 = x_0, \qquad t \in [0,T] \tag{4.7}$$

with a "small noise" $\varepsilon_n W$. Then as $\varepsilon_n \to 0$ the solutions of (4.6), say, converge to the (deterministic) solution of (4.7), under appropriate assumptions on a.

There is a big difference between the two settings (4.1) as $T_n \to \infty$ and (4.6) as $\varepsilon_n \to 0$: the first one corresponds to modelling an intrinsically random phenomenon, while the second one corresponds to modelling a deterministic phenomenon with a small random noise. However, although the second problem is somewhat easier to handle and requires much less assumptions on a and almost no assumption on σ, both problem present many mathematical similarities. For instance we get (under appropriate assumptions) the LAN property for the model associated with (4.6) with rate ε_n.

4.2 Discrete observations

Now we proceed to more realistic observation schemes. The process is observed on a regular grid, at n regularly spaced values in time, say at times $(0, \Delta_n, 2\Delta_n, \ldots, n\Delta_n)$. The observed σ-field is $\mathcal{G}_n = \sigma(Y_{i\Delta_n} : 0 \le i \le n)$. Then, as seen in Subsection 2.5, under mild assumptions the measure $\mathbb{P}_\theta$ are all equivalent on $\mathcal{G}_n$, and the likelihood ratios $Z_n(\zeta/\theta)$ are given by (2.16), which here take the following form ($p^\theta_{s,t}$ denoting the transition densities for the parameter θ):

$$Z_n(\zeta/\theta) = \prod_{i=1}^{n} \frac{p^\zeta_{(i-1)\Delta_n, i\Delta_n}(Y_{(i-1)\Delta_n}, Y_{i\Delta_n})}{p^\theta_{(i-1)\Delta_n, i\Delta_n}(Y_{(i-1)\Delta_n}, Y_{i\Delta_n})}. \tag{4.8}$$

Mathematically speaking, we are observing a realization of a Markov chain, and asymptotic statistical theory for Markov chains is well established, at least in the homogeneous and ergodic case. However we have two main problems here: first, it may happen that Δ_n actually depends on n, so indeed we observe different Markov chains for different values of n; and, more important, *we do not know explicitely the transition densities* of our Markov chain, so the classical techniques for Markov chains *cannot be used* here.

Let us review the most important situations:

A) Constant stepsize. The stepsize is $\Delta_n = \Delta$, independent on n. This setting is the most natural one (apparently at least), but also the most difficult because the transitions

are not explicitely known. Further, to provide asymptotic results we need homogeneity and ergodicity, a set of assumptions which is probably rather rare in finance, as already said before. We study this case in Section 6.

B) Decreasing stepsize. Another possibility is to let $\Delta_n \to 0$ as $n \to \infty$: this will be studied in details in Section 5. Then, although we do not know the densities $p^\theta_{s,t}$, we have good approximations for them as soon as $t - s$ is small: the first order approximation is a Gaussian kernel, and there exist approximations at any order under enough smoothness of the coefficients. So we have approximate expressions for the likelihood which are good when Δ_n is small, and we can expect to find concrete estimators which perform as well, or almost as well, as the MLE.

In fact, we can single out two very different situations:

1) The first one is when $n\Delta_n$ (that is the maximal length of the observed interval) is also big. This amounts to say that $n\Delta_n \to \infty$. The drawback is again that we need homogeneity and ergodicity. Let us mention right away that, in the setting of Equation (4.2), the rates of convergence differ for θ_1 and θ_2.

2) The second one is when $n\Delta_n$ stays bounded, in which case we can as well suppose that $n\Delta_n = T$ is constant. In this situation we of course cannot do better than if we were observing the whole path of the diffusion over the interval $[0, T]$, and consequently we cannot have even consistent estimators for θ_1 in (4.2): so only θ_2 (that is, the volatility, which fortunately is the most crucial parameter in finance) can be consistently estimated. On the other hand, we need neither homogeneity nor ergodicity.

Yet another situation which frequently occurs in practice is when the process is observed on a regular grid, but with missing observations. This is akin with the situation where the observations are made on an irregular grid. The mathematics for theses situations is not really more complicated than in the regular grid case, and we will not touch upon this topic here.

4.3 Observations with errors

So far we have examined the cases with complete observation on an interval (an idealized situation), or observation along a regular grid. Even this is an idealization of the reality, since it is not so often than one can observe exactly the values at any particular time. More realistic schemes are as follows, and will be studied in Section 7 in the setting of discrete observations:

A) Additive errors. Each observation suffers from an additive (random) error: instead of $Y_{i\Delta_n}$ one observes $Y_{i\Delta_n} + \varepsilon_{n,i}$, where the variables $\varepsilon_{n,i}$ are independent of Y and usually i.i.d. and centered when i varies. Rates of convergence of estimators then depend on the variances of the additive errors.

B) Round-off errors. Another kind of errors can occur, especially in financial data:

instead of observing Y_t one observes a rounded-off value (this is also called space quantization). More precisely one fixes a step $\alpha_n > 0$, and instead of $Y_{i\Delta_n}$ one observes the smallest multiple of α_n which is smaller or equal (or alternatively closest) to $Y_{i\Delta_n}$. The rate of convergence of estimators then depend on α_n.

C) Partially observed diffusion. Another situation is when the diffusion X is, say, 2-dimensional and one observes only the firt component X^1 of X (according to one the schemes mentioned above). This happens for example in finance, where X^1 is the price of the asset and X^2 represents the "volatility" of this price. Such a setting is somehow related to Problem A) above. This is quite difficult to study on the mathematical level, and more or less hinges upon filtering theory and may also be viewed as an avatar of the theory of *hidden Markov chains*. Very few definitive results are known here, and one has to resort on procedures which are not known to be optimal. We will not touch upon this problem here.

5 Discrete observations with decreasing stepsize

We start studying the problem B of Subsection 4.2: we are in the setting of Section 4, and we observe our diffusion at times $i\Delta_n$ for $i = 0, 1, \ldots, n$, without any measurement error, and $\Delta_n \to 0$. As said before, we should single out the two cases where $T_n = n\Delta_n$ goes to infinity, or stays constant, and we begin with the second one, which is somewhat easier to grasp (although the proofs are more difficult).

5.1 Observations on a fixed interval

Here we suppose that $\Delta_n = T/n$ for some $T > 0$. The observed σ-fields $\mathcal{G}_n = \sigma(Y_{iT/n} : i = 0, \ldots, n)$ are not increasing, but "in the limit" the σ-field $\mathcal{G}_\infty$ is $\mathcal{Y}_T$: so we have no consistent estimators for θ_1, but only for θ_2 in (4.2), and it is natural to look at the following equations

$$dX_t = a(t, X_t)dt + \sigma(\theta, t, X_t)dW_t, \qquad X_0 = x_0. \tag{5.1}$$

The coefficient a is not specified at all (we are in a *semi-parametric* setting), except that we assume it is continuous. The coefficient σ (or equivalently $c = \sigma\sigma^*$) is smooth enough and with linear growth in x: for example twice continuously differentiable in all variables is more than enough for most results below. For simplicity we suppose that Θ is a compact interval of $\mathbb{R}$, but everything would work as well in the multidimensional case for θ.

We need also an identifiability assumption: there are several possibilities, but again for simplicity we assume the simplest one to check, namely that

$$\theta \neq \zeta \qquad \Rightarrow \qquad c(\theta, 0, x_0) \neq c(\zeta, 0, x_0) \tag{5.2}$$

(the minimal one would be that for any $\zeta \neq \theta$, the $\mathbb{P}_\theta$-probability that the two processes $t \mapsto c(\theta, t, Y_t)$ and $t \mapsto c(\zeta, t, Y_t)$ agree on the interval $[0, T]$ equals 0).

The first (theoretical rather than practical) question which arises is whether the local models around some value $\theta \in \Theta$ converge in the sense of subsection 3.1, with an appropriate rate. To solve this, and in addition to the smoothness of c, we need two extra assumptions: first, a mild assumption is that $\dot{c}$, the derivative of c in θ, is not identically 0 along the path of Y, for example we have $\dot{c}(\theta, 0, x_0) \neq 0$ (to be compared with (5.2)); asecond, a much stronger assumption is that the matrix $c(\theta, t, x)$ is invertible for all (t, x). Then, one can show that we have the LAMN property, with rate $u_n = \frac{1}{\sqrt{n}}$ and random Fisher information given by (since $\Theta \subset I\!R$, the random Fisher information is not a matrix but a random number):

$$I(\theta) = \frac{1}{2T} \int_0^T \mathrm{trace}(\dot{c}c^{-1}\dot{c}c^{-1})(\theta, s, Y_s)ds. \tag{5.3}$$

This result has been shown by Dohnal [5] when X is 1-dimensional, by using an explicit expression of the densities of the transitions in terms of the scale function and of an extra Brownian bridge; it was next extended in [10] in the d-dimensional case when the coefficient c derives from a potential (the same explicit expression being still available), and it was given its final form by Gobet [12], using Malliavin calculus.

Hence the rate-efficient estimators will converge at rate $\sqrt{n}$, and asymptotically efficient estimators will further be asymptotically mixed Gaussian centered around the true value and with conditional variance $I(\theta)^{-1}$.

The second question which arises, and *is* of much practical interest, is to find such rate-efficient or even asymptotically efficient estimators. This turns out to be relatively easy. We can for example construct the following *contrasts*:

$$V_n(\zeta) = \sum_{i=1}^{n} \left(\log\det\left(c\left(\zeta, \frac{i-1}{n}, Y_{\frac{(i-1)T}{n}}\right)\right) \right.$$
$$\left. + \frac{n}{T}(Y_{\frac{iT}{n}} - Y_{\frac{(i-1)T}{n}})^* c^{-1}\left(\zeta, \frac{i-1}{n}, Y_{\frac{(i-1)T}{n}}\right)(Y_{\frac{iT}{n}} - Y_{\frac{(i-1)T}{n}})\right). \tag{5.4}$$

Then one takes the following estimator:

$$\widehat{\theta}_n = \mathrm{ArgMin}\ V_n(.) \tag{5.5}$$

(the function $\zeta \mapsto V_n(\zeta)$, being continuous, admits an absolute minimum on the compact set Θ; if there are several, take any one of them in (5.5)). It can then be proved that the sequence $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P_\theta$ towards a centered mixed Gaussian variable with conditional variance $I(\theta)^{-1}$, provided θ is in the interior of Θ. The proofs are a bit complicated, but the reasons for both the LAMN property and the optimality of the above estimators are simple enough:

- If $c(\theta, t, x) = c(\theta, t)$ does not depend on x at all and if $a \equiv 0$, then under each $I\!P_\theta$ the process Y is Gaussian, with mean x_0, and the densities $p^\theta_{s,t}$ are explicitely known. Then a tedious but elementary computation shows that the LAN property holds, with rate $\frac{1}{\sqrt{n}}$ and limiting Fisher information $\frac{1}{2T}\int_0^T \mathrm{trace}(\dot{c}c^{-1}\dot{c}c^{-1})(\theta, s)ds$. In addition, the variable $V_n(\zeta)$ in (5.4) is $-\log Z_n(\zeta/\theta)$, up to a multiplicative constant; hence (5.5) gives the MLE, which in the Gaussian case is known to be the best estimator.

- Coming to the general case, under $I\!P_\theta$ and conditionally on $\mathcal{Y}_{(i-1)T/n}$ the variable $Y_{iT/n}$ is approximately Gaussian with mean $Y_{(i-1)T/n} + O_P(1/n)$ and variance $c(\theta, (i-1)T/n, Y_{(i-1)T/n})/n$. Then our statistical model behaves asymptotically like another model constructed as such: we have first our canonical process Y and the law $I\!P_\theta$; then we have another process U which, conditionally on the path $t \mapsto Y_t$, is Gaussian with mean x_0 and covariance $E(U_s U_t^*) - x_0 x_0^* = \int_0^s c(\theta, r, Y_r)dr$ for $s \leq t$; finally, we observe the variables $U_{iT/n}$. Therefore one can argue "conditionally on the process Y" and, at the heuristic level, we can apply the (elementary) results valid for Gaussian processes.

The previous results hold under the crucial hypothesis that the matrix c is everywhere invertible. If this fails, the formulae (5.3) and (5.4) make no sense. However it is still possible to obtain reasonable estimators, as shown in [9]. For instance, instead of defining V_n by (5.4), we can set

$$V_n(\zeta) = \sum_{i=1}^{n} \left(\left| Y_{\frac{iT}{n}} - Y_{\frac{(i-1)T}{n}} \right|^2 - \frac{T}{n} \sum_{i=1}^{d} c^{ii} \left(\zeta, \frac{(i-1)T}{n}, Y_{\frac{(i-1)T}{n}} \right) \right)^2 \tag{5.6}$$

and still define $\widehat{\theta}_n$ by (5.5). Then $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P_\theta$ to a centered mixed Gaussian variable with conditional variance

$$\frac{2T \int_0^T \left((\sum_i \dot{c}^{ii})^2 \sum_{i,j} (c^{ij})^2 \right)(\theta, s, Y_s)ds}{\left(\int_0^T (\sum_i \dot{c}^{ii})^2 (\theta, s, Y_s)ds \right)^2}. \tag{5.7}$$

One can check that, as it should be, this quantity is always bigger than $1/I(\theta)$ when c is invertible. Once more, the contrast (5.6) is only one among many different possibilities.

Remarks: 1) The same method accomodates the case where there are missing data, or where the observations take place at an irregular grid. For example if the observations are at times $0 = t(n,0) < t(n,1) < \ldots < t(n,n) = T_n$, we can take the following contrast (extending (5.4), and with the notation $\Delta(n,i) = t(n,i) - t(n,i-1)$):

$$V_n(\zeta) = \sum_{i=1}^{n} \left(\log\det \left(c\left(\zeta, t(n,i-1), Y_{t(n,i-1)}\right)\right) \right.$$
$$\left. + \frac{1}{\Delta(n,i)} (Y_{t(n,i)} - Y_{t(n,i-1)})^* c^{-1} \left(\zeta, t(n,i-1), Y_{t(n,i-1)}\right) (Y_{t(n,i)} - Y_{t(n,i-1)}) \right). \tag{5.8}$$

Then we have exactly the same asymptotic result for $\widehat{\theta}_n$ given by (5.5), provided $T_n \to T$ and the "empirical" measure $\frac{1}{n}\sum_{i=1}^n \varepsilon_{t(n,i)}$ of the observation times converges weakly to the uniform measure over the interval $[0,T]$.

2) When c is not invertible, the estimators minimizing (5.6) converge with the rate $\sqrt{n}$, but this does not mean that they are rate-efficient: although this remains an open question, it might happen that the singularity of c induces the LAMN property with a smaller rate, that is a rate u_n such that $u_n\sqrt{n} \to 0$: then rate-efficient estimators $\widehat{\theta}'_n$ are such that the

sequence $\frac{1}{u_n}(\widehat{\theta}'_n - \theta)$ is tight under $I\!P_\theta$, while of course the sequence $\frac{1}{u_n}(\widehat{\theta}_n - \theta)$ with $\widehat{\theta}_n$ as above is not tight under $I\!P_\theta$, and $\widehat{\theta}_n$ is not rate-efficient. Nevertheless, the estimators $\widehat{\theta}_n$ stay "reasonable" in all cases.

The Ornstein-Uhlenbeck process: Let us consider the process (4.3). In the present setting, θ_1 is a nuisance parameter and θ_2 is the parameter we wish to estimate. It turns out that we have not only the LAMN, but even the LAN property, with rate $\frac{1}{\sqrt{n}}$ and Fisher information $I(\theta) = 1/2\theta_2^2$. The contrast (5.4) takes the form

$$V_n(\zeta_2) = n\left(\log\zeta_2 + \frac{1}{T\zeta_2}\sum_{i=1}^{n}(Y_{\frac{iT}{n}} - Y_{\frac{(i-1)T}{n}})^2\right),$$

and the minimum contrast estimator is

$$\widehat{\theta}_{2,n} = \frac{1}{T}\sum_{i=1}^{n}(Y_{\frac{iT}{n}} - Y_{\frac{(i-1)T}{n}})^2.$$

Note that this works whenever the value of θ_1 is. Now, if further θ_1 is known, one can derive the genuine MLE, say $\widehat{\theta}'_{2,n}$, which takes the following form, to be compared with $\widehat{\theta}_{2,n}$ above:

$$\widehat{\theta}'_{2,n} = \begin{cases} \frac{2\theta_1}{n(1-e^{-2\theta_1 T/n})}\sum_{i=1}^{n}(Y_{\frac{iT}{n}} - e^{-\theta_1 T/n}Y_{\frac{(i-1)T}{n}}))^2 & \text{if } \theta_1 \neq 0 \\ \frac{1}{T}\sum_{i=1}^{n}(Y_{\frac{iT}{n}} - Y_{\frac{(i-1)T}{n}}))^2 & \text{if } \theta_1 = 0. \end{cases}$$

The sequence $(\widehat{\theta}'_{2,n})$ is also asymptotically efficient, and indeed $\sqrt{n}(\widehat{\theta}_{2,n} - \widehat{\theta}'_{2,n})$ goes to 0 in $I\!P_\theta$-probability.

5.2 Observations on an increasing interval

Now we suppose that at the same time $\Delta_n \to 0$ and $T_n = n\Delta_n \to \infty$. Here again the observed σ-fields $\mathcal{G}_n = \sigma(Y_{i\Delta_n} : i = 0, 1, \ldots, n)$ are not increasing, but "in the limit" $\mathcal{G}_\infty = \mathcal{Y}$: so as soon as the minimal identifiability assumption is met, namely that the measures $I\!P_\theta$ are all mutually singular on the largest σ-field $\mathcal{Y}$, we can hope for consistent estimators for θ.

Constructing estimators which are, first, consistent, and further with a reasonable (or optimal) rate of convergence is however a very different matter. This is where the place where the parameter comes in (in the drift term or in the diffusion term) makes a lot of differences. This is why we write the equations in the form (4.2), and some preliminary remarks are more or less obvious:

a - For θ_2 we can apply the previous method; by keeping only the first $l_n = [1/\Delta_n]$ observations (where $[x]$ denotes the integer part of x) and discarding the others, we are in the previous setting with l_n observations, and we can construct estimators which converge to the true value θ_2 with the rate $\sqrt{l_n} \sim 1/\sqrt{\Delta_n}$; this is of course not very good if Δ_n goes slowly to 0, but at least it gives consistent estimators.

Using all the data we can of course hope for better estimators for θ_2, but then to derive any kind of asymptotic properties we have to assume that our diffusions are homogeneous and ergodic.

b - For θ_1 it is quite another matter: first, to obtain any kind of reasonable result we must again assume that our diffusions are homogeneous ergodic. Second, we cannot do better than if the whole path of our process had been observed over the interval $[0, T_n]$, and we have seen already that the best possible rate in the later case is $\sqrt{T_n}$.

Therefore in the rest of this subsection we assume that the coefficients a and c depend on θ_1 and θ_2 respectively, and on x, but not on time. We suppose that the set Θ of all possible values for $\theta = (\theta_1, \theta_2)$ is a compact subset of $\mathbb{R}^2$ (higher dimension for θ is purely a notational problem). Finally we assume that the diffusions are ergodic, and that the unique invariant probabilities $\mu_\theta = \mu_{\theta_1,\theta_2}$ all have $\mathbb{R}^d$ as their support. Within this setting, the minimal identifiability assumption stated above has a simple expression in terms of the coefficient:

$$\left.\begin{array}{lll} a(\theta_1, x) = a(\zeta_1, x) \quad \text{for all } x & \Rightarrow & \theta_1 = \zeta_1 \\ c(\theta_2, x) = c(\zeta_2, x) \quad \text{for all } x & \Rightarrow & \theta_2 = \zeta_2. \end{array}\right\} \tag{5.9}$$

There are several ways of finding good estimators in this setting: see e.g. Florens-Zmirou [8] or Yoshida [37]; here we expound a method due to Kessler [22]: this method has been derived for the 1-dimensional case, so we suppose here that our diffusions are 1-dimensional, but there would be no difficulty to extend it to the multi-dimensional case, except for very cumbersome notation. It is based on the consideration of the infinitesimal generator A_θ of the diffusion, which takes the form (see (2.5)):

$$A_\theta f(x) = a(\theta_1, x) f'(x) + \frac{1}{2} c(\theta_2, x) f''(x),$$

and its iterates A_θ^m for $m = 2, \ldots$, and A_θ^0 is by convention the identity operator. For taking $A_\theta^m f$ we need of course f to be of class $C2m$, while the coefficients a and c should be at least of class $C^{2(m-1)}$ in x: for simplicity we assume that they are infinitely differentiable in x, and also 3 times differentiable in θ, with all derivatives with polynomial growth and the first derivatives in x bounded. Finally, we assume that $\varepsilon \leq c(\theta_2, x) \leq 1/\varepsilon$ for some $\varepsilon > 0$ and all θ_2, x, and that the measure μ_θ has moments of all orders and that $\sup_t \mathbb{E}_\theta(|Y_t|^p) < \infty$ for all $p < \infty$. All these assumptions are satisfied in most applications, as soon as the processes are ergodic.

Let us introduce a number of notation. We denote by $\phi(x) = x$ the identity on $\mathbb{R}$. Then we define the functions

$$r_l(h, \theta, x) = \sum_{i=0}^{l} \frac{h^i}{i!} A_\theta^i \phi(x),$$

$$g_{\theta,x}^0(y) = (y - x)^2,$$

$$j \geq 1 \quad \Rightarrow \quad g_{\theta,x}^j(y) = 2(y - x)\frac{A_\theta^j \phi(x)}{j!} + \sum_{r,s \geq 1, r+s=j} \frac{A_\theta^r \phi(x)}{r!} \frac{A_\theta^s \phi(x)}{s!},$$

$$\Gamma_l(h,\theta,x) = \sum_{j=0}^{l} h^j \sum_{r=0}^{l-j} \frac{h^r}{r!} A_\theta^r g_{\theta,x}^j(x).$$

Γ_l is a polynomial of degree l in h, with no constant term and first order term equal to $hc(\theta_2,x)$, so $\Gamma_l'(h,\theta,x) = \frac{\Gamma_l(h,\theta,x)}{hc(\theta_2,x)}$ is a polynomial of degree $l-1$ in h with constant term equal to 1. Then we can denote by $d_{j,l}(\theta,x)$ and $e_{j,l}(\theta,x)$ the coefficients of order $j \geq 0$ of the Taylor expansion in h, around 0, of the functions $1/\Gamma_l'(h,\theta,x)$ and $\log \Gamma_l'(h,\theta,x)$ respectively. Finally, we consider the contrast

$$V_{l,n}(\zeta) = \sum_{i=1}^{n} \left(\log c(\zeta_2, Y_{(i-1)\Delta_n}) + \sum_{j=1}^{l} \Delta_n^j e_{j,l+1}(\theta, Y_{(i-1)\Delta_n}) \right.$$
$$\left. + \frac{1}{\Delta_n c(\zeta_2, Y_{(i-1)\Delta_n})} (Y_{i\Delta_n} - r_l(\Delta_n, \zeta, Y_{(i-1)\Delta_n}))^2 \left(1 + \sum_{j=1}^{l} \Delta_n^j d_{j,l+1}(\zeta, Y_{(i-1)\Delta_n}) \right) \right) \quad (5.10)$$

This expression looks complicated, all the more when l is large, but it must be observed that it is "explicit" in the sense that if one knows the functions a and c, everything in (5.4) can be explicitely computed. Then one consider the estimator

$$\widehat{\theta}_n^l = (\widehat{\theta}_{1,n}^l, \widehat{\theta}_{2,n}^l) = \text{ArgMin } V_{l,n}(.), \quad (5.11)$$

which exists since Θ is compact and $V_{l,n}$ is a continuous function.

Now the result is as follows: suppose that Δ_n is such that $n\Delta_n^{l/2} \to 0$ for some integer $l \geq 2$, and also that θ is in the interior of Θ. Then the pair $(\sqrt{n\Delta_n}(\widehat{\theta}_{1,n}^l - \theta_1), \sqrt{n}(\widehat{\theta}_{2,n}^l - \theta_2))$ converges in law, under $I\!P_\theta$, towards a pair (U,V) of independent centered Gaussian variables with respective variances given by

$$\left(\int \frac{\dot{a}(\theta_1,x)^2}{c(\theta_2,x)} \mu_\theta(dx) \right)^{-1}, \qquad 2 \left(\int \frac{\dot{c}(\theta_2,x)^2}{c(\theta_2,x)^2} \mu_\theta(dx) \right)^{-1}. \quad (5.12)$$

Remarks: 1) The contrast (5.10) is an approximation of the log-likelihood (up to a multiplicative negative constant), which converges to the true log-likelihood as $l \to \infty$ for each fixed n. This is why the estimators based on this contrast work well when l is large enough relatively to the size of Δ_n, a fact expressed by the property $n\Delta_n^{l/2} \to 0$.

2) We do not know about the optimality of the second component $\widehat{\theta}_{2,n}^l$, although it is certainly rate-efficient at least. But the first component $\widehat{\theta}_{1,n}^l$ is asymptotically efficient: in fact, if instead of the values $Y_{i\Delta_n}$ one observes the whole path of Y over $[0,T_n]$, from the results of Subsection 4.1 we know that θ_2 is known exactly and that for θ_1 the LAN property holds with rate $1/\sqrt{n\Delta_n}$ and asymptotic Fisher information being the inverse of the first expression in (5.12) (compare with (4.4)): so $\widehat{\theta}_{1,n}^l$ performs as well as the asymptotically efficient estimators for θ_1 when we observe the whole path over $[0,T_n]$.

3) Let us emphasize once more the two different rates we get for the estimation of the two components θ_1 and θ_2.

4) How to apply the previous method, and in particular how to choose l ? This is of course a crucial point. On the theoretical level, the sequence Δ_n is given, and we have $n\Delta_n^{l/2} \to 0$ for all l bigger than some l_0, in which case one should take $l = l_0$, or it may also happen that $n\Delta_n^p \to \infty$ for all $p < \infty$, in which case the previous method breaks down. In practice it is quite a different matter, since indeed n and Δ_n are given ! ; hopefully n is large and Δ_n is small, and one may perform the previous estimations for increasing values of l, until the estimators $\widehat{\theta}^l_{1,n}$ and $\widehat{\theta}^l_{2,n}$ more or less stabilize.

To accomodate more precisely this sequential procedure, one may also give an adaptive version of the previous estimators (see the thesis of Kessler [21] for a precise definition) where the computation of $\widehat{\theta}^l_n$ is based on the previous value $\widehat{\theta}^{l-1}_n$: then one stops when $\widehat{\theta}^l_{1,n} - \widehat{\theta}^{l-1}_{1,n}$ is small w.r.t. $1/\sqrt{T_n}$ and $\widehat{\theta}^l_{2,n} - \widehat{\theta}^{l-1}_{2,n}$ is small w.r.t. $1/\sqrt{n}$.

The Ornstein-Uhlenbeck process: Let us consider the process (4.3). By looking at the explicit form for the likelihoods, one can prove that in our setting (the ergodic case, so that $\theta_1 > 0$) the LAN property holds, with rate $\frac{1}{\sqrt{T_n}}$ for the θ_1-component and $\frac{1}{\sqrt{n}}$ for the θ_2-component, and with asymptotic Fisher information matrix

$$I(\theta)_{1,1} = \frac{1}{2\theta_1}, \qquad I(\theta)_{2,2} = \frac{1}{2\theta_2^2}, \qquad I(\theta)_{1,2} = I(\theta)_{2,1} = 0. \tag{5.13}$$

Comparing with (5.12), we observe that the covariance matrix of the centered Gaussian variable (U, V) introduced just before this formula is exactly $I(\theta)^{-1}$: in other words, as soon as $n\Delta_n^{l/2} \to 0$, the sequence of estimators $(\widehat{\theta}_n)$ is *asymptotically efficient* for estimating θ (with of course different rates for the two comonents).

Now, it turns out that we have also local asymptotic properties when $\theta_1 \leq 0$, exactly as for observations over a whole interval (Section 4): if $\theta_1 = 0$ we have the LAQ property with rates $\frac{1}{T_n}$ for θ_1 and $\frac{1}{\sqrt{n}}$ for θ_2; if $\theta_1 < 0$ we have the LAMN property with rates $e^{\theta_1 T_n}$ for θ_1 and $\frac{1}{\sqrt{n}}$ for θ_2; further the components $I(\theta)_{i,j}$ of the associated random matrix $I(\theta)$ are given by by (5.13) if (i, j) is either $(1, 2)$ or $(2, 1)$ or $(2, 2)$, while $I(\theta)_{1,1}$ is as given in Section 4. As a matter of fact, one could prove that the MLE is asymptotically efficient in these two cases as well.

6 Discrete observations with constant stepsize

The setting is the same as in Section 5, except that here $\Delta_n = \Delta$. In a sense, this scheme of observations seems the most natural one when observations are discrete in time. However, as in Subsection 5.2, we must assume that the diffusions are homogeneous ergodic, with unique invariant probability measures μ_θ whose supports are $I\!R^d$ (or D if all diffusions live on the domain D). Observe that here, in accordance with the results of this subsection, the rates for θ_1 and θ_2 should both be $\sqrt{n}$: so there is no reason to single out these two components, and we come back to Equation (4.1), with coefficients not depending on time, and some given starting point x_0 (the same for all θ's).

Next, about the necessary identifiability assumption: at first glance one should take

the analogue of (5.9), that is

$$a(\zeta,x)=a(\theta,x) \quad \text{and} \quad c(\zeta,x)=c(\theta,x) \quad \text{for all } x \qquad \Rightarrow \qquad \zeta=\theta. \tag{6.1}$$

However this turns out to be not enough: indeed "at the limit" we have $\mathcal{G}_\infty=\sigma(Y_{i\Delta}:i=0,1,\ldots)$, and the restriction of $I\!P_\theta$ to $\mathcal{G}_\infty$ is entirely characterized by the kernel P_Δ^θ, where $(P_t^\theta))_{t\geq 0}$ is the semi-group of the diffusion with parameter θ. So the right identifiability assumption is in fact

$$P_\Delta^\zeta(x,.)=P_\Delta^\theta(x,.) \quad \text{for all } x \qquad \Rightarrow \qquad \zeta=\theta, \tag{6.2}$$

an assumption which is strictly stronger than (6.1), and which unfortunately cannot be read in a simple way from the coefficients (in the examples 3, 4 and 5 of Section 2, however, this identifiability assumption is satisfied).

Apparently the problem is much simpler here than before, since we observe the sequence $(Y_{i\Delta})_{0\leq i\leq n}$, an homogeneous Markov chain with transition P_Δ^θ under each $I\!P_\theta$. And as soon as the matrices $c(\theta,x)$ are invertible these transitions admits positive densities $p_\Delta^\theta(x,.)$ w.r.t. Lebesgue measure. So the likelihood on the σ-field $\mathcal{G}_n$ is given by (4.8), i.e.

$$Z_n(\zeta/\theta)=\prod_{i=1}^n \frac{p_\Delta^\zeta(Y_{(i-1)\Delta},Y_{i\Delta})}{p_\Delta^\theta(Y_{(i-1)\Delta},Y_{i\Delta})}. \tag{6.3}$$

Then, since our Markov chains are in addition ergodic, it is well known since a long time ago (see for example the book of Roussas [34]) that under some reasonable smoothness assumptions on the densities p_Δ^θ (implied by suitable smoothness of a and c), we have the LAN property with rate $1/\sqrt{n}$, with asymptotic Fisher information matrix at point θ given by

$$I(\theta)_{i,j}=\int \frac{\frac{\partial}{\partial\theta_i}p_\Delta^\theta(x,y)\frac{\partial}{\partial\theta_j}p_\Delta^\theta(x,y)}{p_\Delta^\theta(x,y)}\mu_\theta(dx)dy. \tag{6.4}$$

Now the problems begin, since we aim to getting asymptotically efficient estimators, if possible. The MLE is of course asymptotically efficient, but it is also unavailable because we have no explicit exprssion for the densities p_Δ^θ in terms of the coefficients of the equations. So we have to resort on other methods.

6.1 Approximating the likelihood

A first method consists in computing an approximation of the likelihood (6.3) and then maximizing in ζ this approximation; more precisely we have to compute an approximation $\tilde{p}_\Delta^\zeta(x,y)$ for all pairs of the form $(x,y)=(Y_{(i-1)\Delta},Y_{i\Delta})$ and all values of ζ, and minimize

$$\zeta \mapsto \tilde{V}_n(\zeta)=\prod_{i=1} n\tilde{p}_\Delta^\zeta(Y_{(i-1)\Delta},Y_{i\Delta}).$$

The key point is to compute $\tilde{p}_\Delta^\zeta$: for this, we can use expansions of p_Δ^ζ as a power series in Δ and stop the expansion at some prescribed degree. This is for example the point of view taken by Ait-Sahalia [1]: he first makes a space transorm which render p_Δ

relatively close to a Gaussian kernel, and then uses an expansion in Hermite polynomials, but unfortunately this works only in dimension 1; other expansions are also possible. But in a sense this is not much different from the underlying idea behind the method explained in Subsection 5.2, and it is likely to work only for relatively "small" values of Δ.

The nice thing about such methods is that they give right away the function $\tilde{p}_\Delta^\zeta(x,y)$ (as a function of x, y, and ζ), and often also allow to keep track of the error $p_\Delta^\zeta - \tilde{p}_\Delta^\zeta$.

Another possibility is to use Monte-Carlo techniques to approximate p_Δ^ζ: this has been developped by Pedersen in [30], [31]: these work for any value of Δ, but the error $\tilde{P}_\Delta^\zeta - p_\Delta^\zeta$ is difficult to control.

This method is relatively efficient, but its main drawback is that it allows to nicely approximate $p_\Delta^\zeta(x,y)$ for any given individual value of (x,y,ζ), but not as a function of these variables, which we need because we have to maximize $\tilde{V}_n$. So we can either compute $\tilde{p}_\Delta^\zeta$ for all ζ in a finite set Θ_n consisting in grid with a mesh mush smaller than $\frac{1}{\sqrt{n}}$ (since we want an estimate whose error is of order $\frac{1}{\sqrt{n}}$), or we can use a gradient method or more sophisticated minimization methods, which necessitate e.g. the approximation of the derivatives of p_Δ^ζ in ζ. All these are again much computing-intensive. And once again, it is extremely difficult to keep track of the approximation errors of the method.

6.2 Contrast functions and estimating functions

Another idea, which has many similarities with the method explained in Section 5, consists in using a contrast function of the form

$$V_n(\zeta) = \sum_{i=1}^{n} F(Y_{(i-1)\Delta}, Y_{i\Delta}, \zeta) \tag{6.5}$$

for a suitable smooth function F, and to take for estimator $\widehat{\theta}_n$ the value, or one of the values, which minimize $V_n(.)$, as in (5.5).

When F is differentiable in ζ and when $\widehat{\theta}_n$ above is not on the boundary of Θ, then $\widehat{\theta}_n$ also solves the system of equations $\frac{\partial}{\partial\zeta_i} V_n(\zeta) = 0$ for $i = 1, \ldots, q$ (q is the dimension of θ. That is, with G denoting the gradient of F (as a function of the parameter), $\widehat{\theta}_n$ solves the following equation, called *an estimating equation*:

$$W_n(\widehat{\theta}_n) = 0, \tag{6.6}$$

where W_n is

$$W_n(\zeta) = \sum_{i=1}^{n} G(Y_{(i-1)\Delta}, Y_{i\Delta}, \zeta). \tag{6.7}$$

In the sequel, we suppose for simplicity that θ is 1-dimensional, so the function G above is also 1-dimensional; but everything would work in the multidimensional case as well. First, the ergodic theorem says that

$$\frac{1}{n} W_n(\zeta) \;\to\; W(\theta,\zeta) := \int \mu_\theta(dx) P_\Delta^\theta(x,dy) G(x,y,\zeta) \tag{6.8}$$

in $I\!P_\theta$-probability. Therefore if G is smooth enough and chosen in such a way that

$$\int \mu_\theta(dx) P_\Delta^\theta(x,dy) G(x,y,\zeta) = 0 \qquad \Leftrightarrow \qquad \theta = \zeta, \tag{6.9}$$

then the sequence $\widehat{\theta}_n$ of (6.6) converge in $I\!P_\theta$-probability to θ, for every $\theta \in \Theta$: that is, the estimators are weakly consistent. Similarly one has that

$$\frac{1}{n}\dot{W}_n(\zeta) \;\to\; \dot{W}(\theta,\zeta) := \int \mu_\theta(dx) P_\Delta^\theta(x,dy) \dot{G}(x,y,\zeta) \tag{6.10}$$

in $I\!P_\theta$-probability, and this convergence holds even uniformly in ζ (for each fixed θ; as usual a "dot" means taking the derivative in θ).

Suppose next that G satisfies in addition

$$\int G(x,y,\theta) P_\Delta^\theta(x,dy) = 0 \qquad \text{for all } \; x,\theta. \tag{6.11}$$

We then say that the estimating function W_n is a *martingale estimating function*, because the summands in (6.7) are martingale increments (note that (6.11) yields the implication from right to left in (6.9)). Then the central limit theorem for ergodic Markov chains and the martingale property imply that the sequence $\frac{1}{\sqrt{n}} W_n(\theta)$ converges in law under $I\!P_\theta$ to a centered Gaussian variable with variance $\int \mu_\theta(dx) P_\Delta^\theta(x,y) G(x,y,\theta)^2$, under suitable assumptions. Now, suppose that θ is in the interior of Θ; for n large enough, $\widehat{\theta}_n$ is also in the interior of Θ, so we can write $0 = W_n(\widehat{\theta}_n) = W_n(\theta) + (\widehat{\theta}_n - \theta)\dot{W}_n(\zeta_n)$, where ζ_n is between θ and $\widehat{\theta}_n$ (and random). These facts, together with (6.10), yield that the sequence $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P_\theta$ to a centered Gaussian variable with variance

$$\alpha = \int \mu_\theta(dx) P_\Delta^\theta(x,y) G(x,y,\theta)^2 \;/\; \left(\int \mu_\theta(fx) P_\Delta^\theta(x,y) \dot{G}(x,y,\theta) \right)^2. \tag{6.12}$$

In other words, the sequence of estimators $\widehat{\theta}_n$ defined by (6.6) is rate-efficient, provided the function G is smooth enough and satisfies (6.9) and (6.11). Then, among all possible choices for G, one should choose one which minimizes the variance α in (6.12).

Of course (6.9) and (6.11) are not so easy to fulfill, and especially the last one, since one still does not know the transitions P_Δ^ζ ! One may think of several possibilities:

1 - For each θ take a function $\phi(.,\theta)$ which is an eigenfuntion of the generator A_θ of our diffusion, with eigenvalue $\lambda(\theta)$ (note that $\lambda(\theta) < 0$). Then $G(x,y,\theta) = g(x,\theta)(\phi(y,\theta) - e^{-\lambda(\theta)}\phi(x,\theta))$ will satisfies (6.11), whichever the function g is. Linear combinations of such functions also do the job. This is done e.g. by Kessler and Sørensen [24]. Observe that since A_θ is given in terms of the coefficients $a(\theta,.)$ and $c(\theta,.)$, finding eigenfunctions for A_θ is in principle easier than finding P_Δ^θ.

2 - We take an arbitrary smooth function f on $I\!R^d \times I\!R^d$ and let $G(x,y,\theta) = f(x,y) - f'(x,\theta)$, where $f'(x,\theta) = \int P_\Delta^\theta(x,dy) f(x,y)$, so (6.11) is obvious. The function f' is, once more, not explicit; but it is possible to approximate it by Monte-Carlo techniques for example, with the same drawbacks than the method using approximated likelihoods. This method is akin with the so-called "simulated moments method" developped by many authors in practical studies.

3 - One can relax (6.11), for instance by taking $G(x,y,\theta) = g(x,\theta)$ not depending on y, but such that $\int \mu_\theta(dx)g(x,\theta) = 0$ for all θ: this property replaces (6.11) and is enough to obtain that $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P_\theta$ to some centered Gaussian variable, although with a variance having a much more complicated form than (6.12). The advantage is that finding functions g satisfying $\int \mu_\theta(dx)g(x,\theta) = 0$ is much easier than finding functions satisfying both (6.10) and (6.11), because μ_θ is reasonably often explicitely known (when the diffusion process is 1-dimensional for example). The main disadvantage is that the minimal variance we can thus obtain by appropriately choosing g is always bigger than the minimal variance α in (6.12) when G is appropriately chosen. This was done by Kessler [23], who in particular stidied the Ornstein-Uhlenbeck case completely.

Many other possibilities are indeed available in the literature: one can consult Prakasa Rao [33] for an extensive account on this, and of course the paper [2] of Bibby, Jacobsen and Sørensen in this volume for many more details about estimating functions. But one must say that indeed there is no universal method in this setting, working for all diffusions: the comparison between the various methods is largely empirical and done only for special diffusion processes.

A last remark: if one wants to get rid of the identifiability problem stated at the beginning (that is, replace (6.2) by (6.1)), a possibility is to assume that the observations take place at times $T_1, T_2, \ldots, T_n$, the occurence times of a Poisson process independent of the diffusion, and with parameter $1/\Delta$: this seems strange at first glance, but in fact it is compatible with many sets of data, in which the inter-observation times are not really regularly spaced (see Duffie and Singleton [6]).

7 Observations with errors

When there are errors of various kinds, or incomplete observations, very little is known so far. The problem becomes difficult because we loose the Markov struture of the process and introduce complicated dependencies between the observed variables.

We will give very few elements here, and only for a single observation scheme: namely when the process is observed at times iT/n for $i = 0, 1, \ldots, n$. In view of the discussion in subsection 5.1, we consider Equation (5.1):

$$dX_t = a(t, X_t)dt + \sigma(\theta, t, X_t)dW_t, \qquad X_0 = x_0$$

with a not depending on θ. For simplicity we also assume that X and θ as well are 1-dimensional, that σ (or $c = \sigma^2$) does not vanish, and that a and c are smooth enough in all variables. Further we assume some kind of identifiability assumption, say that $c(\theta, 0, x_0) \neq c(\zeta, 0, x_0)$ whenever $\theta \neq \zeta$ (as said before this could be much weakened). This choice of our observation scheme is due to the fact that this is almost the only situation for which the influence of error measurements has been studied so far.

7.1 Additive errors

The simplest possible kind of error, if not the most natural one, is when each observation is blurred with an additive error, all errors being i.i.d. and independent of the diffusion itself and with a known distribution. As a matter of fact, we can only deal with Gaussian errors. So we suppose that the actual observations are of the form

$$Z_i^n = Y_{iT/n} + \sqrt{\rho_n}\varepsilon_i, \tag{7.1}$$

where ρ_n is a known positive number, and the ε_i are i.i.d. variables with law $\mathcal{N}(0,1)$, independent of the process X.

We let the error variance ρ_n depend on n: this is because we are interested again in asymptotic properties, and it may seem natural that the measurement error be small when there are many observations; on the other hand, the case where $\rho_n = \rho$ does not depend on n may also seem quite natural.

Mathematically speaking, the statistical model at hand may be described as follows: let still $(\Omega, \mathcal{Y}, (I\!P_\theta)_{\theta\in\Theta})$ be the canonical space with the canonical process Y and the weak solutions $I\!P_\theta$ of our diffusion equations; let $(\Omega'', \mathcal{Y}'', I\!Q)$ be another probability space on which are defined i.i.d. $\mathcal{N}(0,1)$ variables ε_n; then we take the statistical model $(\Omega', \mathcal{Y}', (I\!P'_\theta)_{\theta\in\Theta})$, where

$$\Omega' = \Omega \times \Omega'', \qquad \mathcal{Y}' = \mathcal{Y} \otimes \mathcal{Y}'', \qquad , I\!P'_\theta = I\!P_\theta \otimes I\!Q.$$

Further we define the variables Z_i^n on this space by (7.1). The observed σ-field at stage n is then $\mathcal{G}_n = \sigma(Z_i^n : i = 0, 1, \ldots, n)$. Here, not only the σ-fields $\mathcal{G}_n$ are not increasing, but as said before we have lost the Markov property for the chain $(Z_i^n)_{i\geq 0}$.

7.1.1 Neglecting the errors

The first try to estimate might be to neglect the measurement errors and to use the method explained in Subsection 5.1. Let us try this in the *very simple* case where $x_0 = 0$ and $a \equiv 0$ and $c(\theta, t, x) = \theta$ and $\Theta = (0, \infty)$. Then of course our Equation (5.1) reduces to $X = \sqrt{\theta}W$, where W is a Brownian motion.

If we observe without error (which amounts to taking $\rho_n = 0$), we have the LAN property with rate $\frac{1}{\sqrt{n}}$ and asymptotic Fisher information $I(\theta) = \frac{1}{2\theta^2}$; this can be seen from (5.3), but it also reduces to very old results since here we observe equivalently the normalized increments $U_i^n = \sqrt{\frac{n}{T}}(Y_{(i-1)T/n} - Y_{iT/n})$, which are i.i.d. $\mathcal{N}(0, \theta)$. The contrast (5.4) writes as

$$V_n(\zeta) = n \log \zeta + \frac{n}{T} \sum_{i=1}^{n} \frac{(Y_{iT/n} - Y_{(i-1)T/n})^2}{\zeta},$$

whose minimum is achieved at the point

$$\widehat{\theta}_n = \frac{1}{T} \sum_{i=1}^{n} (Y_{iT/n} - Y_{(i-1)T/n})^2 = \sum_{i=1}^{n} (U_i^n)^2. \tag{7.2}$$

Since the U_i^n are i.i.d. $\mathcal{N}(0, \theta)$, it is also well known that $\widehat{\theta}_n$ is optimal for estimating θ in all possible senses, not only asymptotically but for every n.

Now we have measurement errors, but we still use the contrast above, just replacing the unobserved $Y^n_{iT/n}$ by the variables Z^n_i of (7.1). This amounts to taking the estimate $\widehat{\theta}_n$ of (7.2) with Z^n_i instead of $Y^n_{iT/n}$. Since all variables are Gaussian, it is elementary to check that

- $\widehat{\theta}_n \to \theta$ in $I\!P'_\theta$-probability if and only if $n\rho_n \to 0$: so if $n\rho_n$ does not go to 0 the sequence $\widehat{\theta}_n$ is not even consistent, and in fact it goes to $+\infty$!
- $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P'_\theta$ to an $\mathcal{N}(0, 2\theta^2) = \mathcal{N}(0, 1/I(\theta))$ random variable if and only if $n^{3/2}\rho_n \to 0$: in this case, the sequence $\widehat{\theta}_n$ is asymptotically efficient.
- $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P'_\theta$ to an $\mathcal{N}(v, 2\theta^2) = \mathcal{N}(v, 1/I(\theta))$ random variable if and only if $n^{3/2}\rho_n \to v \in [0, \infty)$: if $v > 0$ the sequence $\widehat{\theta}_n$ is rate-efficient, but not asymptotically efficient because of the bias.
- If $n^{3/2}\rho_n \to \infty$, then the sequence $\sqrt{n}(\widehat{\theta}_n - \theta)$ is not tight under $I\!P'_\theta$: so in this case the sequence $\widehat{\theta}_n$ is not rate-efficient.

The main point coming out from this analysis is that it is *very dangerous* to forget about measurement errors: if these are "small enough", meaning that $n^{3/2}\rho_n$ is small, then there is no harm (this is obvious from a heuristic point of view, except for the power 3/2 which comes from precise calculations), but otherwise one gets bad estimators, and even *inconsistent* estimators when $n\rho_n$ does not go to 0.

7.1.2 Taking care of the errors

In view of what precedes we should take the errors into consideration, at least to get consistent estimators, and if possible to find asymptotically efficient estimators. For this, we reproduce some (unfortunately not quite complete) results from [11].

Let us first single out three cases corresponding to different asymptotic behaviour of ρ_n (Case 3 below accomodates the situation where $\rho_n = \rho$ does not depend on n), and introduce some notation:

$$\left.\begin{array}{lll} \textbf{Case 1:} & n\rho_n \to u = 0, & \text{then set} \quad u_n = 1/\sqrt{n} \\ \textbf{Case 2:} & n\rho_n \to u \in (0,\infty), & \text{then set} \quad u_n = 1/\sqrt{n} \\ \textbf{Case 3:} & n\rho_n \to u = \infty, \quad \sup_n \rho_n < \infty, & \text{then set} \quad u_n = (\rho_n/n)^{1/4}, \end{array}\right\} \tag{7.3}$$

$$\phi_u(x,y) = \begin{cases} \frac{y^2}{2x^2} & \text{if } u = 0 \\ \frac{y^2(2+x/u)}{2\sqrt{u}x^{3/2}(4+x/u)^{3/2}} & \text{if } 0 < u < \infty \\ \frac{y^2}{8x^{3/2}} & \text{if } u = \infty \end{cases} \tag{7.4}$$

Next, in agreement with the case without errors, we can hope for the LAMN property to hold, perhaps. We have been unable to prove this, but we can prove it in the particular

case where $a \equiv 0$ and $c(\theta, t, x) = c(\theta, t)$ does not depend on x (then the solution of (5.1) is a Gaussian process with constant mean x_0). We also assume that c is smooth and that both c and $\dot{c}$ do not vanish (the last assumption may be somehow relaxed, but we want to keep things simple here). In this case, and if further $\rho_n = 0$, according to Subsection 5.1 we have the LAN property with asymptotic Fisher information

$$I(\theta) = \frac{1}{2T} \int_0^T \frac{\dot{c}(\theta, s)^2}{c(\theta, s)^2} ds. \tag{7.5}$$

If measurement errors are present and if ρ_n is such that we are in one of the three cases above, one can then prove that the LAN property hold *with rate* u_n, and asymptotic Fisher information

$$I(\theta) = \frac{1}{T} \int_0^T \phi_u(Tc(\theta, s), T\dot{c}(\theta, s)) ds. \tag{7.6}$$

Observe that (7.6) and (7.5) agree when $u = 0$. Observe also that the rate is $\frac{1}{\sqrt{n}}$ as soon as $n\rho_n \to u < \infty$: in this case, and even when $n^{3/2}\rho_n$ does not go to 0, one should be able to find asymptotically efficient estimators with this rate $\frac{1}{\sqrt{n}}$, a property not enjoyed by the estimators (7.2).

Now, let us turn to constructing estimators. We go back to the general situation of Equation (5.1). We first choose a sequence k_n of integers in such a way that $nu_n^2/k_n \to 0$ and $k_n/nu_n \to 0$, by taking for example

$$k_n = \begin{cases} [n^{1/4}] & \text{in Cases 1 and 2} \\ [n^{5/8}\rho_n^{3/8}] & \text{in Case 3.} \end{cases}$$

We also set $l_n = [n/k_n]$, and we take n large enough to have $k_n \geq 2$ and $l_n \geq 2$. Next we consider the $(k_n - 1) \times (k_n - 1)$-matrix D^n whose entries are

$$D^n_{i,j} = \begin{cases} 2\rho_n & \text{if } i = j \\ -\rho_n & \text{if } |i - j| = 1 \\ 0 & \text{otherwise} \end{cases}$$

and whose eigenvalues are

$$\lambda_i^n = 2\rho_n \left(1 - \cos \frac{i\pi}{k_n}\right), \qquad i = 1, \ldots, k_n - 1,$$

and we write $D^n = P^n L^n P^{n,\star}$ where L^n is diagonal with entries given above and P^n is orthogonal. Next we set $s_m^n = \frac{k_n(m-1)T}{n}$ for $m = 1, \ldots, l_n$, and, recalling the observations Z_i^n of (7.1),

$$F_j^{n,m} = \sum_{i=1}^{k_n - 1} P_{ij}^n (Z^n_{k_n(m-1)+j} - Z^n_{k_n(mp-1)+j-1}), \qquad j = 1, \ldots, k_n - 1, \quad m = 1, \ldots, l_n,$$

$$S_m^n = \frac{1}{k_n - 1} \sum_{i=1}^{k_n - 1} Z^n_{k_n(m-2)+i}, \qquad m = 2, \ldots, l_n,$$

$$\Phi_j^{n,m}(\zeta) = 2\left(1 - \cos\frac{j\pi}{k_n}\right) + \frac{Tc(\zeta, s_m^n, S_m^n)}{n\rho_n}, \qquad j = 1, \ldots, k_n - 1, \quad m = 2, \ldots, l_n.$$

Then at this point we can write a contrast function as

$$W_n(\zeta) = \sum_{m=2}^{l_n} \sum_{j=1}^{k_n-1} \left(\frac{(F_j^{n,m})^2}{\rho_n \Phi_j^{n,m}(\zeta)} + \log \Phi_j^{n,m}(\zeta) \right). \tag{7.7}$$

Finally $\widehat{\theta}_n$ is a point achieving the minimum of $W_n(.)$ over Θ: observe that $W_n(\zeta)$, hence $\widehat{\theta}_n$, can actually be computed (in principle) from the observations.

Then one can prove that, provided θ is in the interior of Θ, the sequence $\frac{1}{u_n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P'_\theta$ towards a centered mixed Gaussian variable with conditional variance $I(\theta)^{-1}$, where

$$I(\theta) = \frac{1}{T}\int_0^T \phi_u(Tc(\theta, s, Y_s), T\dot{c}(\theta, s, Y_s))ds. \tag{7.8}$$

Remarks: 1) When $c(\theta, t, x) = c(\theta, t)$ does not depend on x, then (7.6) and (7.8) agree; so if futher $a \equiv 0$ the estimators $\widehat{\theta}_n$ above are asymptotically efficient.

2) Although we cannot prove the LAMN property in general, a comparison with the case $c(\theta, t, x) = c(\theta, t)$ strongly supports the fact that we indeed have the LAMN property with rate u_n and asymptotic conditional Fisher information $I(\theta)$ given by (7.8), together with the fact that the estimators $\widehat{\theta}_n$ are asymptotically efficient, also in the case of genuine diffusions (5.1).

3) It is noteworthy to observe that all ingredients above use the *known* value ρ_n and of course the observations themselves, but they do not depend on the case we are in (see (7.3)). This is of big practical importance because, although we know n and ρ_n, it is difficult to decide whether the product $n\rho_n$ is "very small" or moderate or big... In fact, in all cases our estimators will be optimal (asymptotically speaking), within the relevant asymptotic framework.

4) Assuming that the errors are Gaussian is rather strong, but we know nothing about more general errors.

The Ornstein-Uhlenbeck process: We just mention here the case of the process (4.3), to see more explicitely on an example the value taken by the conditional Fisher information (7.8). We get in fact a deterministic quantity, given by

$$I(\theta) = \begin{cases} \frac{1}{2\theta_2^2} & \text{if} \quad u = 0 \\ \frac{\sqrt{T}(2+T\theta_2/u)}{4\sqrt{u}\theta_2^{3/2}(4+T\theta_2/u)^{3/2}} & \text{if} \quad 0 < u < \infty \\ \frac{\sqrt{T}}{8\theta_2^{3/2}} & \text{if} \quad u = \infty \end{cases}$$

One can observe that T comes in explicitely, except when $u = 0$.

7.2 Round-off errors

Another sort of error consists in round-off errors: instead of the true value x of the diffusion at some time, only a rounded-off value of x, at some level $\alpha > 0$, is available to the statistician: that is, instead of x one observes the value $\alpha[x/\alpha]$ (recall that $[v]$ denotes the integer part of $v \in \mathbb{R}$). This sort of measurement is particularly relevant for financial data, where one models prices or interest rates with a diffusion, although the actual values in the market are always multiples of some basic currency (dollars, or cents, or 0.1%'s,...).

Recall that everything here is 1-dimensional (θ as well as the diffusion), and a and c are smooth and c does not vanish. Exactly as in the previous subsection where the error level ρ_n was possibly depending on n, here the round-off level will also possibly depend on n, say α_n. That is, at stage n we observe the variables

$$Z_i^n = \alpha_n \left[Y_{\frac{iT}{n}} / \alpha_n \right]. \tag{7.9}$$

Contrarily to the previous case there is no need to enlarge our probability space: the statistical model is thus $(\Omega, \mathcal{Y}, (\mathbb{P}_\theta)_{\theta \in \Theta})$ (the canonical space with the canonical process Y), but the observed σ-field is $\mathcal{G}_n = \sigma(Z_i^n : i = 0, 1, \ldots, n)$.

7.2.1 Neglecting the errors

Here again we can first try to use the method of Subsection 5.1 without taking care of the errors, and we again do this in the simple case where $x_0 = 0$ and $a \equiv 0$ and $c(\theta, t, x) = \theta$ and $\Theta = (0, \infty)$, that is $X = \sqrt{\theta} W$, where W is a Brownian motion.

Recall again that without round-off error we have the LAN property with rate $\frac{1}{\sqrt{n}}$ and asymptotic Fisher information $I(\theta) = \frac{1}{2\theta^2}$, and the optimal estimators are given by (7.2), that is

$$\widehat{\theta}_n = \frac{1}{T} \sum_{i=1}^{n} (Y_{iT/n} - Y_{(i-1)T/n})^2. \tag{7.10}$$

Now we have round-off errors. If we just use $\widehat{\theta}_n$ above with Z_i^n given by (7.9) instead of $Y_{iT/n}^n$, we get the following asymptotic behaviour (see [17]):

- If $\alpha_n \sqrt{n} \to 0$, then $\widehat{\theta}_n \to \theta$ in $\mathbb{P}_\theta$-probability.
- The sequence $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $\mathbb{P}_\theta$ towards an $\mathcal{N}(0, 2\theta^2)$ random variable if and only if $\alpha_n n \to 0$; and it is tight if and only if the sequence $\alpha_n n$ is bounded.
- If $\alpha_n \sqrt{n} \to \beta \in (0, \infty)$, then $\widehat{\theta}_n$ converges in $\mathbb{P}_\theta$-probability to some constant depending on β, which is strictly bigger than θ: so the estimators $\widehat{\theta}_n$ are not consistent.
- If $\alpha_n \sqrt{n} \to \infty$ and $\alpha_n \to 0$, then $\frac{1}{\alpha_n \sqrt{n}} \widehat{\theta}_n$ converges in $\mathbb{P}_\theta$-probability to some positive constant: so the estimators $\widehat{\theta}_n$ are not consistent, and even converge to $+\infty$.

- If $\alpha_n \to \alpha \in (0,\infty)$, then the sequence $\frac{1}{\sqrt{n}}\widehat{\theta}_n$ converges in $I\!P_\theta$-probability towards a constant times the sum of the values of the local time of Y taken at all level $k\alpha$ for $k \in \mathbf{Z}$, taken at time T.

We can draw the same conclusion from this analysis than for additive errors: it is *very dangerous* to forget about round-off errors: if these are "small enough", meaning that $n\alpha_n$ is small, then there is no harm in doing that, but otherwise one gets bad estimators, and even *inconsistent* estimators when $\alpha_n\sqrt{n}$ does not go to 0.

7.2.2 Taking care of the errors

Now we take the round-off errors into consideration and we exhibit asymptotically efficient estimators. The method explained below is due to Delattre: see [3] and [4].

First, it is possible to prove the LAMN property when $\alpha_n\sqrt{n} \to \beta \in [0,\infty)$. Describing the asymptotic random Fisher information is a bit lengthy, and we need some preliminary notation. First, consider a Brownian motion W over $I\!R$ (such that $W_0 = 0$) and a random variable U which is uniform over $[0,1]$ and independent from W; for $\alpha > 0$, let $\mathcal{H}_\alpha$ be the σ-field generated by all variables of the form $\alpha\left[U + \frac{W_i}{\alpha}\right]$, $i \in \mathbf{Z}$; then for all $i \in I\!N$ and all $\alpha \in (0,\infty)$ we define the random variables $\xi_i^\alpha = E((W_i - W_{i-1})^2 - 1|\mathcal{H}_\alpha)$. The following formula defines a positive function over $I\!R_+$:

$$J(\alpha) = E((\xi_1^\alpha)^2) + 2\sum_{i=2}^{\infty} E(\xi_1^\alpha \xi_i^\alpha),$$

and $J(0) := \lim_{\alpha\downarrow 0} J(\alpha)$ equals 2 and J strictly decreases from 2 to 0 when α increases from 0 to ∞. Then, assuming that a and c are smooth enough and that c does not vanish, if $\alpha_n\sqrt{n} \to \beta \in [0,\infty)$ we have the LAMN property with rate $\frac{1}{\sqrt{n}}$ and conditional Fisher information given by

$$I(\theta) = \frac{1}{4T}\int_0^T \frac{\dot{c}(\theta,s,Y_s)^2}{c(\theta,s,Y_s)^2}\, J\left(\frac{\beta}{\sqrt{c(\theta,s,Y_s)}}\right) ds. \tag{7.11}$$

Observe that if $\beta = 0$, and since $J(0) = 2$, the above $I(\theta)$ is also the value given in (5.3), corresponding to observations without errors. If $\beta > 0$ then the above $I(\theta)$ is strictly smaller than the value given in (5.3), which corresponds to the intuitive idea that if the round-off error is "big" then we obtain less information on the process.

When $\alpha_n\sqrt{n} \to \infty$ and $\alpha_n \to 0$ it is also possible to prove the LAMN property, but the rate is now α_n and the conditional Fisher information takes yet another form.

Now let us come to constructing estimators. Here again, $\widehat{\theta}_n$ will be a point achieving the minimum of a contrast function which takes the form

$$W_n(\zeta) = \sum_{i=1}^{n} F(\alpha_n\sqrt{n}, c(\zeta, \frac{(i-1)T}{n}, Z_i^n + \frac{\alpha_n}{2}), \sqrt{n}(Z_i^n - Z_{i-1}^n)), \tag{7.12}$$

where F is a suitable (known) function on $I\!R_+ \times I\!R_+ \times I\!R$, so $\widehat{\theta}_n$ can actually be computed (in principle) from the observations.

Then when $\alpha_n\sqrt{n} \to \beta \in [0,\infty)$ one can prove the following, as soon as the function F is smooth, with polynomial growth at most, even in the last variable (i.e. $F(\alpha,z,x) = F(\alpha,z,-x)$), and such that the function $z \mapsto \int_0^1 du \int h(y)F(\alpha,z,\alpha[u+z'y/\alpha])h(y)dy$ (where h is the density of the law $\mathcal{N}(0,1)$) admits a unique minimum at point $z'=z$: provided θ is in the interior of Θ, the sequence $\sqrt{n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P_\theta$ towards a centered mixed Gaussian variable with conditional variance $\Sigma_{F,\beta}(\theta)$ (an expression in terms of F and its derivatives and of β, looking a bit like (5.7) in a more complicated way). Examples of possible such functions F are

$$F_p(\alpha,z,x) = \frac{|x|^p}{\gamma_p(z,\alpha)} + \log\gamma_p(z,\alpha), \quad \text{where} \quad \gamma_p(z,\alpha) = \int_0^1 du \int h(y)\left|\alpha[u+\frac{z}{\alpha}]\right|^p dy.$$

The estimators such constructed are thus rate-efficient; they are usually nor asymptotically efficient, and indeed one does not know how to choose F in such a way that $\widehat{\theta}_n$ becomes asymptotically efficient (i.e. $\Sigma_{F,\beta} = I(\theta)^{-1}$, see (7.11)). However with F_2 as above, we obtain $\Sigma_{F,0} = I(\theta)^{-1}$, so the associated estimators are asymptotically efficient when $\alpha_n\sqrt{n} \to 0$ at least.

When $\alpha_n\sqrt{n} \to \infty$ and $\alpha_n \to 0$, one can use the same contrasts, except that we need some additional assumptions on the behaviour of the function $F(\alpha,z,x)$ as $\alpha \to \infty$. Then the sequence $\frac{1}{\alpha_n}(\widehat{\theta}_n - \theta)$ converges in law under $I\!P_\theta$ towards a centered mixed Gaussian variable with conditional variance $\Sigma_{F,\infty}(\theta)$ (which indeed is the limit of $\frac{1}{\alpha^2}\Sigma_{F,\alpha}$ as $\alpha \to \infty$). So again these estimators are rate-efficient.

An example of function F which works for all cases ($\alpha_n\sqrt{n}$ bounded, or going to infinity) is

$$F(\alpha,z,x) = \frac{1}{1 \bigvee \alpha}\,\frac{|x|^2}{\gamma_2(z,\alpha)} + \log\frac{\gamma_2(z,\alpha)}{1 \bigvee \alpha}. \tag{7.13}$$

Remarks: 1) Exactly as in Remark 3 of Subsection 6.8.1, one sees that with e.g. the function in (7.13) we have estimators $\widehat{\theta}_n$ which do not depend on the asymptotic behaviour of the sequence $\alpha_n\sqrt{n}$: this is again of big practical importance.

2) The fact that α_n goes to 0 is crucial to all what precedes. If for example we take $\alpha_n = \alpha$ not depending on n, then apart from the convergence in probability of $\widehat{\theta}_n/\sqrt{n}$ for the estimators in (7.10) towards a sum of local times, essentially nothing is known, but even the identifiability of the parameter in this case could be a problem.

The Ornstein-Uhlenbeck process: Again, we mention the case of the process (4.3), to see more explicitely on an example the value taken by the conditional Fisher information (7.11). We get again a deterministic quantity, which is $I(\theta) = \frac{1}{2\theta_2^2}J(\beta/\sqrt{\theta_2})$: we see clearly on this formula the influence of the "asymptotic" round-off factor β, and that the key quantity is the quotient $\beta/\sqrt{\theta_2}$, as it should be by scaling arguments. And, contrarily to the case with additive errors, the quantity T does not come into the picture.

8 Concluding remarks

We now have seen a series of methods for etimating parameters in diffusion processes, mainly when the observations are regularly spaced at times $i\Delta_n$ for $i = 0, \ldots, n$. Obviously we have let aside a number of problems, even in this setting: first we have made assumptions on the coefficients which are not necessarily met in practice, like in particular the invertibility of the diffusion coefficient which plays an important rôle in several cases. Second, and probably more importantly, we have not really studied the case where the diffusions live on a domain D, and especially the case where the boundary ∂D can be attained. Third, the case where there are measurement errors has been studied in quite specific situations only, and there is obviously a need to go further in this topic.

Let us stick to the case of perfect observations on a regular grid. In a concrete situation we have a number n of observations, and a stepsize $\Delta = \Delta_n$. Then, which one among the various methods should we choose ? Since in practice n is (relatively) large, this question boils down to determining in which asymptotic situation we can reasonably assume we are: is Δ small, in which case we may assume that $\Delta = \Delta_n \to 0$? is it "really" small, in which case we can suppose that $n\Delta_n$ is more or less constant ? Related with this problem is, of course, the kind of parameter we wish to estimate: in the setting of Equation (4.2), is it θ_1, or θ_2, or both ? keeping in mind that the volatily is probably the most important parameter, we can think that in most cases we are interested essentially in θ_2; keeping in mind that as soon as $T_n = n\Delta_n \to \infty$ we essentially need the process to be ergodic, is it reasonable to believe that our phenomenon is truly stationary ?

All these questions are crucial in a sense, and so far there is no definitive answers. In fact there is a need for numerical experimentations on some case studies (for models more involved than the Ornstein-Uhlenbeck process which, because of its Gaussian property, might present too much specific structure to be truly representative of general diffusions): one should check the validity of the different methods with the same set of data, perhaps with simulated data to be sure of the underlying model. And also, before using a method which necessitates ergodicity, we should perhaps make a test of the stationarity of the process, or at least do the estimation on disjoint pieces of data and check whether the estimates on each piece are more or less consistent with one another, or whether there is a clear trend.

It might of course be the case that all reasonable methods give more or less the same results, at least as far as the second component θ_2 is concerned: a close examination of the methods suggests such a nice property (except for the methods using simulated moments or simulated likelihoods), but there is of course no garantee for that.

Finally, we end this paper with some words about inference for discontinuous processes, letting apart the case of point processes, which has been extensively studied but is not relevant for finance. There is a number of papers about general estimation problems for possibly discontinuous semimartingales, but mainly when the whole path of the process is observed on an interval $[0, T_n]$, the asymptotic being $T_n \to \infty$. More interesting would be to look at discontinuous processes observed on a regular grid, just as above: but the problems seem then to be quite difficult, and very few results have been obtained so far.

More precisely, if the grid has a constant stepsize Δ, and provided our discontinuous

processes are Markov, we are again in the situation of observing n successive values taken by a Markov chain, and if it is ergodic the methods of Section 6 can still be applied, with obvious modifications. If our processes are Lévy processes, and although they are never ergodic, some reasonably complete answers are available: see Jedidi [18].

However if the stepsize Δ_n goes to 0, then very unexpected phenomena appear. Assume for example that $\Delta_n = T/n$ and that the observed processes are $X_t = \theta Z_t$, where Z is a given process whose law is known, and θ is the parameter to estimate ($\theta \in (0, \infty)$). If Z is a stable process with index $\alpha \in (0, 2]$, then by the scaling property we easily find that the LAN property holds with rate $\frac{1}{\sqrt{n}}$. If now Z is the sum of a symmetric stable process of index $\alpha \in (0, 2]$ and of a standard Poisson process, then Far [7] proved that when $\alpha = 2$ (i.e. Z is the sum of a Brownian motion and a Poisson process) we have the LAMN property with rate $\frac{1}{\sqrt{n}}$ (and asymptotically efficient estimators can be derived and behave better than if we had a Brownian motion alone); and if $\alpha < 2$, then we have convergence of the local model with a "random rate" which is $\frac{1}{\sqrt{n}}$ with positive probability, and $\frac{1}{n^{1/\alpha}}$ with also positive probability: so asymptotically efficient estimators converge to the true value at a random rate which is $\sqrt{n}$ with positive probability, and $n^{1/\alpha}$ (much bigger than $\sqrt{n}$) also with positive probability. And of course nothing is known when the observed process is the solution of an SDE driven by, say, a Lévy process, and with a coefficient depending on the parameter of interest.

References

[1] Ait Sahalia Y. (2001): Maximum-likelihood estimation of discretely-sampled diffusions: a closed-form approximation approach. To appear in *Econometrica.*

[2] Bibby B.M., Jacobsen M., Sørensen M. (2001): Estimating functions for diffusions. This volume.

[3] Delattre S. (1997): Estimation du coefficient de diffusion pour un processus de diffusion en présence d'erreurs d'arrondi. *Thesis*, Université Paris-6.

[4] Delattre S. (1998): Estimation for a diffusion in the presence of round-off errors. To apper in *Scand. J. Statist.*.

[5] Dohnal G. (1987): On estimating the diffusion coefficient. *J. Applied Probab.*, **34**, 105-114.

[6] Duffie D., Glynn P. (1996): Estimation of continuous-time Markov processes sampled at random time intervals. Preprint.

[7] Far H. (2001): Propeiétés asymptotiques de modèmes paramétriques associés à l'observation discrétisée de processus de sauts. *Thesis*, Université Paris-6.

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

[9] Genon-Catalot V., Jacod J. (1993): On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. IHP-Proba.*, **29**, 119-151.

[10] Genon-Catalot V., Jacod J. (1994): Estimation of the diffusion coefficient for diffusion processes; random sampling. *Scand. J. Statistics*, **21**, 193-221.

[11] Gloter A., Jacod J. (2000): Diffusions with measurements errors: I, II (Prépublications du Laboratoire de Probabilités et modèles aléatoires, Paris-6,7).

[12] Gobet E. (2000): LAMN property for elliptic diffusion: a Malliavin calculus approach. Forthcoming in Bernoulli.

[13] Ibragimov I.A., Has'minskii R.Z. (1981): *Statistical estimation: asymptotic theory.* Springer Verlag: Berlin.

[14] Höpfner R., Jacod J., Ladelli L. (1990): Local asymptotic normality and mixed normality for Markov statistical models. PTRF **86**, 105-129.

[15] Ikeda N. and Watanabe S. (1981): *Stochastic differential equations and diffusion processes.* Noth Holland: New York.

[16] Jacod J., Shiryaev A.N. (1987): *Limit theorems for stochastic processes.* Springer Verlag: Berlin.

[17] Jacod J. (1996): La variation quadratique du brownien en présence d'erreurs d'arrondi. *Astérisque* **236**, 155-162.

[18] Jedidi W. (2001): Local Asymptotic Normality of Statistical Models Associated with Discrete Observations of Lévy Processes. Preprint.

[19] Jeganathan P. (1982): On the asymptotic theory of estimation when the limit of the loglikelihood ration is mixed normal. *Sankhya A*, **44**, 173-212.

[20] Karatzas I. and Shreve S. (1988): *Brownian motion and stochastic calculus.* Springer Verlag: Berlin.

[21] Kessler M. (1996): Estimation paramétrique des coefficients d'une diffusion ergodique à partir d'observations discrètes. *Thesis*, Université Paris-6.

[22] Kessler M. (1997): Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statistics* **24**, 211-229

[23] Kessler M. (1997): Simple and explicit estimating functions for a discretely observed diffusion process. *Scand. J. Statistics* 27, 65-82.

[24] Kessler M., Sørensen M. (1999): Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli* **5**, 299-314.

[25] Kutoyants Yu.A. (1984): *Parameter estimation for stochastic processes.* Heldermann Verlag: Berlin.

[26] LeCam L. (1986): *Asymptotic methods in statistical decision theory.* Springer Verlag: Berlin.

[27] LeCam L., Yang G.L. (1990): *Asymptotics in statistics: some basic concepts.* Springer Verlag: Berlin.

[28] Liptser, R.S. and Shiryayev, A.N. (1978): *Statistics of Stochastic Processes.* Springer Verlag: Berlin.

[29] Øksendal B. (1985): *Stochastic differential equations.* Springer Verlag: Berlin.

[30] Pedersen A.R. (1995): A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statistics* **22**, 55-71.

[31] Pedersen A.R. (1995): Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli* **1**, 257-279.

[32] Prakasa Rao B.L.S. (1999): *Semimartingales and their statistical inference.* Chapman & Hall: Boca Raton.

[33] Prakasa Rao B.L.S. (1999): *Statistical inference for diffusion type processes.* Arnold: London.

[34] Roussas G. (1972): *Contiguity of probability measures.* Cambridge University Press: London.

[35] Revuz D. and Yor M. (1991): *Continuous Martingales and Brownian Motion.* Springer Verlag: Berlin.

[36] Stroock D.W. and Varadhan S.R.S. (1979): *Miltidimensional diffusion processes.* Springer Verlag: Berlin.

[37] Yoshida N. (1992): Estimation for diffusion processes from discrete observations. *J. Multivariate Anal.* **41**, 220-242.