# Monte Carlo likelihood-based inference for diffusions

**Omiros Papaspiliopoulos**
Universitat Pompeu Fabra
http://www.econ.upf.edu/∼omiros

# Prerequisites and textbook references

- Probability at the level say of [Williams, 1991]
- Basic stochastic processes (Markov chains, Brownian motion), e.g [Grimmett and Stirzaker, 2001]
- Computational statistics, at the level say of [Liu, 2008]

The theoretical background on diffusions can be found for example in [Øksendal, 1998] or [Kloeden and Platen, 1995].

The theme is a fusion of computational statistics and probabilistic inference for stochastic processes. Hence, we will generally we will iterate the presentation pattern

$$\text{stats motivation} \to \text{stoc. proc. theory} \to \text{comp. method}$$

# Stochastic differential equations (SDEs)

We model $d$-dimensional stochastic process $V \in R^d$ as the solution of an SDE of the type:

$$\mathrm{d}V_s = b(s, V_s; \theta)\,\mathrm{d}s + \sigma(s, V_s; \theta)\,\mathrm{d}B_s, \quad s \in [0, T] \; ; \qquad (1)$$

- $B$ is an $m$-dimensional standard Brownian motion
- $b(\cdot, \cdot) : R_+ \times R^d \to R^d$ is the *drift*
- $\sigma(\cdot, \cdot) : R_+ \times R^d \to R^{d \times m}$ is the *diffusion coefficient* and plays the role of the square root of the covariance matrix

$$\Gamma = \sigma\sigma^*$$

- $V_0$ can be taken as fixed or elicited with a distribution, depending on the context.
- In parametric setting the functionals are known only up to certain parameters $\theta$ (e.g a linear combination of basis functions).

# Applications

SDEs provide a natural model for processes which at least conceptually evolve continuously in time and have continuous sample paths (although natural extensions exist for introducing jump components)

Application areas are ever increasing: finance [Sundaresan, 2000, Eraker et al., 2003, Ait-Sahalia and Kimmel, 2007], biology [Golightly and Wilkinson, 2006], molecular kinetics [Horenko and Schütte, 2008, Kou et al., 2005], longitudinal data analysis [Taylor et al., 1994]...

SDEs in many ways are the stochastic counterpart of ODEs which are the special case when $\sigma = 0$. They are Markov processes therefore a natural model for prediction.

# Part Ia: introduction to SDEs

- Canonical representation
- Itô's lemma
- Solving SDEs: linear SDEs
- Generator of the diffusion, Kolmogorov backward equation, transition operator

# Interpretation of the equation (1)

- infinitesimall-time interpretation of the equation ( **non-linear time series**): for fixed $t$ and $\epsilon \approx 0$

$$V_{t+\Delta t} \approx V_t + b(t, V_t; \theta)\Delta t + \sigma(t, V_t; \theta)(B_{t+\Delta t} - B_t)\,. \quad (2)$$

- white noise interpretation (**dynamical system**):

$$\frac{\mathrm{d}V_s}{\mathrm{d}s} = b(s, V_s, \theta) + \sigma(s, V_s, \theta)\frac{\mathrm{d}B_s}{\mathrm{d}s}$$

- mathematically precise interpretation (semi-martingale): for $s < t$

$$V_t = V_s + \int_s^t b(u, V_u; \theta)\mathrm{d}u + \int_s^t \sigma(u, V_u; \theta)\mathrm{d}B_u \quad (3)$$

Note on notation: drop out all arguments of functionals unless it is necessary

# Brownian motion (BM)

$B$ is standard Brownian motion (e.g Ch.2 of [Øksendal, 1998]):

1. $B_0 = 0$
2. $B$ has independent stationary increments: for $t_1 < t_2 < t_3 < t_4$, $B_{t_4} - B_{t_3}$ is independent of $B_{t_2} - B_{t_1}$ ;
3. For $s < t$, $B_t - B_s \sim N(0; t - s)$

$B$ is Markov and a martingale. Of particular relevance is its sample path properties. It is continuous but rough.

$$\lim_{n \to \infty} \sum_{i=1}^{n} (B_{it/n} - B_{(i-1)t/n})^2 = t \tag{4}$$

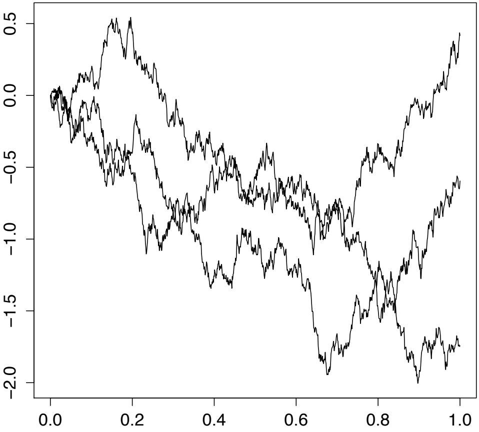
almost surely. The expression above is known as the quadratic variation. Multidimensional BM.

# Canonical representation

A useful and intuitive representation for a probability space on which Brownian motion can be defined is the following:

Let $\Omega = C([0;1); R^m)$ the space of continuous paths, and $\mathcal{F}$ the corresponding cylinder $\sigma$-algebra, and $\omega = (\omega_s; s \in [0;\infty))$ a typical element of $\Omega$. Then, BM is the measurable space $(\Omega, \mathcal{F})$ equipped with certain probability measures $\mathbb{W}^x$ , where $B_0 = x$ $\mathbb{W}^x$-a.s. (i.e $B(\omega) = \omega$).

# Typical sample paths (first thought about simulation)



# Making sense of (3)

- Stochastic integral $\int_0^t \sigma \mathrm{d}B$ mean and variance (**Itô isometry**)
- V is defined implicitly, thus we require existence and uniqueness of solutions - connection to ODEs

Provided the latter holds, its solution is called a diffusion process:

$$V_t = f_t(B; V_0) \tag{5}$$

where $f_t$ is an $\mathcal{F}_t$-measurable function of the Brownian path (strong-weak solutions). The solution has continuous sample paths and it is a Markov process

# Itô's lemma

The foundation of stochastic calculus; applies to Itô processes and it involves the quadratic variation process

$$
\begin{aligned}
X_t &= f(t, V_t) \quad \text{where } f : R \times R^d \to R^l \\
\mathrm{d}X_t &= \frac{\partial f_k}{\partial t} + \sum_i \frac{\partial f_k}{\partial v_i}\mathrm{d}V_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 f_k}{\partial v_i \partial v_j}\mathrm{d}V_i\mathrm{d}V_j
\end{aligned}
$$

- Rules for applying the previous formula: $\mathrm{d}V_i\mathrm{d}V_j$ . When $V$ is the diffusion (1), and $f(t; v) = f(v)$ is real-valued, we get

$$
\mathrm{d}X_t = \left(\sum_{i=1}^{d}\frac{\partial f}{\partial v_i}b_i + \frac{1}{2}\sum_{i,j}\frac{\partial^2 f}{\partial v_i \partial v_j}\Gamma_{i,j}\right)\mathrm{d}t + \sum_i \frac{\partial f}{\partial v_i}\sum_j \sigma_{ij}\mathrm{d}B_j \tag{6}
$$

- formula for $d = l = 1$
- Observation: X might not be a diffusion

# Solving SDEs

There are limited families of SDEs which we can solve analytically, and obtain explicitly $V_t$ as a function of the Brownian path. Itô's lemma is a valuable tool in those cases. See for example Ch.4 of [Kloeden and Platen, 1995]

In many cases, even when the solution can be found analytically, in the sense of (5) with an explicitly given $f_t$ , it might not be very useful for practical purposes, since ft will be too complicated, e.g the result of nested solutions of SDEs.

For statistical purposes, we are interested in cases where the solution allows us to find the conditional distribution of $V_t$ given $V_{t_0}$ for $t > t_0$ (see later for discrete-time dynamics of diffusions).

For 1-d SDEs we have a clear result: linear SDEs (in a broad sense) can be solved, as well as the SDEs which after a transformation become linear. Nevertheless, the function $f_t$ might be too complicated for practical use

Practically useful solution is available for **affine SDEs** with multiplicative noise and **general linear SDEs with additive noise**

It is much harder to give a similar result for multi-d SDEs, although for a special family of multi-d linear SDEs we have again a direct solution. Note that even linear SDEs with additive noise might not have practically simple solution

Linear SDEs play an important role in applications, but as we will see, instrumental role in MC for non-linear diffusions. Therefore, we will study them in some detail.

# Trivial case: scaled BM with drift

$$\mathrm{d}V_s = a_2(s)\mathrm{d}s + b_2(s)\mathrm{d}B_s$$

Then without knowledge of stochastic calculus (!), effectively by definition of the SDE, we have

$$V_t = V_{t_0} + \int_{t_0}^{t} a_2(s)\mathrm{d}s + \int_{t_0}^{t} b_2(s)\mathrm{d}B_s$$

This is a Gaussian process (conditionally on $V_{t_0}$); derive moments

# 1-d linear SDEs (broad sense)

Following terminology in [Kloeden and Platen, 1995] we call an 1-d SDE linear in a broad sense if it takes the form

$$\mathrm{d}V_s = (a_1(s)V_s + a_2(s))\mathrm{d}s + (b_1(s)V_s + b_2(s))\mathrm{d}B_s$$

Special cases of this family are the Ornstein-Uhlenbeck process (Vasicek model), the geometric BM (Black-Scholes model), the Brownian bridge, and trivially BM.

An explicit solution is available for this family.

# Case I: narrow sense, additive noise, useful solution

We first show the solution for the so-called linear SDEs in a narrow sense, which are obtained when $b_1 = 0$, therefore the SDE is driven by additive noise

Let

$$P(t_0, t) = \exp\left\{\int_{t_0}^{t} a_1(s)\mathrm{d}s\right\}$$

Note that $\mathrm{d}P^{-1}/\mathrm{d}t = -P^{-1}a_1(t)$. Then, applying Itô to $P(t_0, t)^{-1}V_t$ we obtain

$$\mathrm{d}(P(t_0, t)^{-1}V_t) = P(t_0, t)^{-1}a_2(t)\mathrm{d}t + P(t_0, t)^{-1}b_2(t)\mathrm{d}B_t$$

thus, we get the solution

$$V_t = P(t_0, t)\left(V_{t_0} + \int_{t_0}^{t} P(t_0, s)^{-1}a_2(s)\mathrm{d}s + \int_{t_0}^{t} P(t_0, s)^{-1}b_2(s)\mathrm{d}B_s\right)$$

Conditionally on $V_{t_0}$ this is also a Gaussian process; the moments can be derived in various ways (we will revisit this). At first instance, we can get the mean and variance directly from the solution, thus we get a complete characterization of the **conditional distribution** of $V_t$ given $V_{t_0}$

Actually, the tractability of $P$ will determine whether the solution yields directly the conditional moments. However, in any case we have made an important step towards understanding the discrete-time dynamics

# Case II: multiplicative noise

We now consider the general case with $b_1 \neq 0$, thus we have multiplicative noise.

Note that the affinne SDE

$$\mathrm{d}V_t = a_1(t)V_t\mathrm{d}t + b_1(t)V_t\mathrm{d}B_t$$

can be solved by our previous result since it reduces to a scaled BM with drift by taking the log-transform and using Itô:

$$V_t = V_{t_0} \exp\left\{\int_{t_0}^{t}\left(a_1(s) - \frac{1}{2}b_1^2(s)\right)\mathrm{d}s + \int_{t_0}^{t} b_1(s)\mathrm{d}B_s\right\}$$

The general case, with $a_2$ and/or $b_2$ different from 0 can be solved, in the sense of expressing $V_t$ as a (very complicated!) function of the Brownian path; the solution involves the time and stochastic integrals with coefficients depending on the solution of the affine SDE driven by the same BM.

In general however, the solution will not be practically useful (for simulation, statistical inference)

# Multivariate linear SDEs with additive noise

$$\mathrm{d}V_s = (a_1(s)V_s + a_2(s))\mathrm{d}s + b_2(s)\mathrm{d}B_s$$

where $a_1$, $a_2$, $b_2$ are matrices of appropriate dimensions. Unlike the 1-d case, the solution is hard to obtain in this case. This is in accordance with ODEs.

It is easy when $a_1$ is constant. In that case, working as before we obtain the same solution as in 1-d case with

$$P(t_0; t) = \exp\{a_1(t - t_0)\}$$

# Example: Brownian bridge

Consider the following SDE:

$$\mathrm{d}V_s = \frac{y - V_s}{T - s}\mathrm{d}s + \mathrm{d}B_s \,, s \in [0, T] \tag{7}$$

i.e $a_1(s) = -1/(T - s)I$, $a_2 = y/(T - s)$, $b_2 = I$, $b_1 = 0$. This is an instance of multi-d linear SDE with non-constant $a_1$ which can be solved, since $a_1$ has a very simple form.

Working as before, we get for $0 < t_1 < t_2 < T$, as

$$V_{t_2} \mid V_{t_1} \sim N\left(V_{t_1} + \frac{t_2 - t_1}{T - t_1}(y - V_{t_1}),\ \frac{(t_2 - t_1)(T - t_2)}{T - t_1} I_d\right) . \tag{8}$$

# Generator

Homogeneous Markov processes: ODEs, discrete and continuous-time Markov chains, SDEs. Common structure

Genetic code of a Markov process enconded in the "one-step" distribution: intuition from Markov chains

For continuous-time processes it is an infinitesimall step.

# Generator of diffusions

We concentrate on homogeneous processes:

$$\mathrm{d}V_s = b(V_s)\mathrm{d}_s + b(V_s)\mathrm{d}B_s$$

Following a similar definition for general Markov processes the generator is defined as an operator acting on an appropriate space of real-valued functions $f : R^d \to R$, and characterised as follows:

$$Af(v) = \lim_{\epsilon \to 0} \frac{\mathbb{E}[f(V_\epsilon)|V_0 = v] - f(v)}{\epsilon}, \quad v \in R^d \tag{9}$$

Note that by continuity of $V$, both nominator and denominator tend to 0. For convenience we denote (this is standard) $\mathbb{E}[f(V_\epsilon)|V_0 = v] = \mathbb{E}^v[f(V_\epsilon)]$.

Note that drectly by Itô we have

$$\mathbb{E}^v[f(V_\epsilon)] = f(v) + \int_0^\epsilon \mathbb{E}\left[\sum_i \frac{\partial f}{\partial v_i} b_i + \frac{1}{2}\sum_{i,j} \frac{\partial^2 f}{\partial v_i \partial v_j}\Gamma_{ij}\right] \mathrm{d}s$$

therefore the generator is

$$Af = \sum_i \frac{\partial f}{\partial v_i} b_i + \frac{1}{2}\sum_{i,j} \frac{\partial^2 f}{\partial v_i \partial v_j}\Gamma_{ij}$$

for sufficiently regular set of functions, which includes all bounded $C^2$ functions. Note that A is a **linear operator**

Combining our new expressions together with Itô we get for real-valued $f$:

$$\mathrm{d}f(V_t) = Af(V_t)\mathrm{d}t + \sum_{i=1}^{d} \frac{\partial f}{\partial v_i} \sum_j \sigma_{i,j} \mathrm{d}B_{j,t} \quad \text{(Itô for diffusions)} \tag{10}$$

$$\mathbb{E}^v[f(V_t)] = f(v) + \int_0^t \mathbb{E}^v[AF(V_s)]\mathrm{d}s \quad \text{(version of Dynkin's formula)}$$

Note the special role that the **null space** of A plays, and its relation with martingale functions of the diffusion

# Kolmogorov backward equation

Let $f$ be bounded $C^2$ function, $V_t$ a time homogeneous diffusion with generator $A$, and

$$u(t, v) = \mathbb{E}^v[f(V_t)]$$

Then $u$ belongs in the domain of $A$ for every $t$ and

$$\begin{aligned} \frac{\partial u}{\partial t} &= Au \\ u(0, v) &= f(v) \end{aligned} \tag{11}$$

(and in fact, it is the unique solution of the above PDE)

# Sketch of the proof (based on Markov property)

Observation 1: Note that by Dynkin:

$$u(t, v) = f(v) + \int_0^t \mathbb{E}^v[Af(V_s)]\mathrm{d}s$$

which shows that $t \mapsto u(t, v)$ is differentiable with

$$\dot{u} = \mathbb{E}^v[Af(V_t)]$$

Fix some t, then we check by definition if $u(t, v)$ is in the domain of $A$. For clarity, let $g(v) = u(t, v)$

$$
\begin{aligned}
\frac{\mathbb{E}^v[g(V_r)] - g(v)}{r} &= \frac{\mathbb{E}^v[\mathbb{E}^{V_r}[f(V_t)]] - f(v)}{r} \\
&= \frac{\mathbb{E}^v[\mathbb{E}^v[f(V_{t+r}) \mid \mathcal{F}_r]] - f(v)}{r} \\
&= \frac{\mathbb{E}^v[f(V_{t+r})] - f(v)}{r} \\
&= \frac{u(t+r, v) - u(t, v)}{r} \to \dot{u}
\end{aligned}
$$

This equation gives a linear PDE for the evolution of expectations of the diffusions; note that the corresponding ODE is non-linear (when we fix v and wish to compute a given expectation)

Thus, although this is a nice mathematical result, to be used in practice numerical solutions will be necessary

# Transition operator and BKE

Let $P^t(v, \cdot)$ be the transition operator:

$$P^t f(v) = \mathbb{E}^v[f(V_t)] = u(t, v) \tag{12}$$

Then, we have shown that

$$\frac{\mathrm{d}}{\mathrm{d}t} P^t f = P^t(Af) = A(P^t f) \tag{13}$$

thus in some sense $A$ and $P^t$ commute

# Part Ib: Statistical challenges: observation schemes and probabilistic inference

# Fully observed case

- Continuous-time observations over a compact time interval $[0, T]$. Realistically, we might have data at arbitrarily high frequency
- No observation noise
- The model makes sense at that high frequency!

We wish to estimate the drift and diffusion functionals. We will see that parametric inference for this observation regime is rather straightforward. In fact, there are still interesting non-trivial developments for non-parametric inference.

# Partially observed case

- Discretely observed diffusions and parameter estimation: (1) observed at discrete time points $0 = t_0 < t_1 < \cdots < t_n$ with observed values $\mathbf{v} = \{v_0, v_1, \ldots, v_n\}$, e.g V is the vector of predator-pray population sizes, the configuration of a molecule). In this case we wish to estimate the drift and diffusion functionals. Parametric/non-parametric
- Partially observed diffusions, state and parameter estimation: (1) is not fully observed at discrete time points (e.g V is the the price and stochastic volatility, V is only the pray size, V is the position of a tracked object observed with noise, higer order models...). In this case we wish to estimate unknown parameters driving the process and make inference for the unobserved components. Filtering-smoothing

# Paradigm followed in the course: probabilistic inference

We are interested in likelihood-based inference for the unknown parameters, i.e maximum likelihood and Bayesian methods; and probabilistic inference for the unobserved processes, i.e inference according to the conditional law of the process given the observed data, where the prior law is given by the SDE specification (1). To simplify the presentation we will refer to such estimation procedures as probabilistic inference.

# Why likelihood-based statistics?

Many classical (ie well-established) results in Statistics support the notion that inference should be based solely on the likelihood (and prior):

- Neymann-Pearson, the likelihood principle.
- Cramer-Rao lower bound, asymptotic normality of MLE in regular problems
- Coherence and access to a ready-made calculus for inference and decision making, called probability theory (Bayesian)

Of course there are arguments against likelihood-based inference:

- All the mathematical arguments for likelihood-based inference rely on the assumption the model is correct which is of course almost always untrue.
- Some non-likelihood based methods can also achieve the Cramer-Rao lower bound (or get close to it) (efficiency).
- There are less clear-cut results about the “optimality” of MLEs for finite data sets
- Likelihood inference can be hard!

A very rich and interesting literature exists for parametric inference for discretely-observed diffusions using non-likelihood methods.

Later in the course, time permitting, we will give a summary of alternative approaches to inference for SDEs

# Why probabilistic inference?

It is quite undebated in the literature that provided a joint model is assumed for observed and unobserved variables, inference for the unobserved variables should be based on the conditional law (this is basic probability theory and conditioning).

There is less consent about how to estimate unknown parameters in presense of unobserved processes (missing data). Nevertheless, integrating out the unobserved processes propertly accounts for the lack of information. Bayesian approaches are more natural and effective in these incomplete data model-based frameworks

Note that for diffusions there is always missing data, since we will note be able to record the whole path, thus the process in-between observation times will be unobserved, even in the fully-observed case.

Of course, we expect this not to be a serious problem if the frequency of observation is appropriately "high"

# Part II: Likelihood inference for the fully observed case

- Quadratic variation
- Girsanov theorem and likelihoods in the path space
- Plug-in MLE

# Setup

We observe $(v_s; s \in [0; T])$ from (1) and wish to estimate the unknown parameters. We will see shortly that the drift and diffusion coefficient are treated very differently in the fully observed case

A developed mathematical framework is available for statistical analyses in this high frequency regime, see for example [Prakasa Rao, 1999]. Two main components of this theory is the quadratic variation identity and the Cameron-Martin-Girsanov change of measure.

# Quadratic variation identity

local characteristics of the SDE can be completely identified given an observed path:

$$\lim_{\Delta\to 0}\sum_{t_j\le t}(V_{t_{j+1}}-V_{t_j})(V_{t_{j+1}}-V_{t_j})^* = \int_0^t \Gamma(s, V_s)\mathrm{d}s \tag{14}$$

in probability for any partition $0 = t_0 \le t_1 \le \cdots \le t_n = t$, whose mesh is $\Delta$.

This generalizes the simpler version for the BM (4)

# Implications for volatility estimation

- from high frequency data we can consistently estimate the diffusion coefficient. In fact it is the square of it which is identifiable: any square root of $\Gamma$ is indistinguishable by the data
- Therefore, unless one has very good reasons it is not necessary to model parametrically the diffusion coefficient; we can treat it fully non-parametrically, thus work with

$$\mathrm{d}V_s = b(s, V_s; \theta)\,\mathrm{d}s + \sigma(s, V_s)\,\mathrm{d}B_s$$

  and estimate $\sigma$ consistently using the QV identity to get $\hat{\sigma}$ (which effectively is very close to $\sigma$ if the mesh is sufficiently fine)
- The topic of estimation of $\sigma$ under various model mispecifications has received a great amount of interest lately

# Implications for estimating drift parameters

We have seen that a path on a compact interval completely identifies the diffusion coefficient of the SDE. This clearly implies that the probability laws generated by the diffusions

$$\begin{aligned} \mathrm{d}V_t &= b\mathrm{d}t + \sigma_1 \mathrm{d}B_t \\ \mathrm{d}V_t &= \alpha\mathrm{d}t + \sigma_2 \mathrm{d}B_t \end{aligned}$$

are mutually singular if $\Gamma_1 \neq \Gamma_2$. Therefore, the problem of estimating parameters in the drift can be informally cast as: estimating $\theta$ among the probability laws corresponding to the same $\sigma$.

# Likelihood ratios in the path space

Under weak conditions the laws which correspond to SDEs with the same diffusion coefficient but different drifts are equivalent and a simple expression for the Radon-Nikodym derivative (likelihood ratio) is available. This is the context of the Cameron-Martin-Girsanov theorem for Itô processes, e.g Theorem 8.6.6 of [Øksendal, 1998]. (it is very easy to understand what goes on as a Gaussian change of measure)

Consider functionals $h$ and $\alpha$ of the dimensions of $b$ and assume that $h$ solves the equation:

$$\sigma(s,x)h(s,x) = b(s,x) - \alpha(s,x)$$

Additionally, let $\mathbb{P}_b$ and $\mathbb{P}_\alpha$ be the probability laws implied by the SDE with drift $b$ and $\alpha$ respectively.

Then, under certain conditions $\mathbb{P}_b$ and $\mathbb{P}_\alpha$ are equivalent with density (continuous time likelihood) on $\mathcal{F}_t = \sigma(B_s; s \leq t)$, $t \leq T$, given by

$$\left.\frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_\alpha}\right|_t = \exp\left\{\int_0^t h(s, V_s)^* \mathrm{d}B_s - \frac{1}{2}\int_0^t [h^* h](s, V_s)\mathrm{d}s\right\}. \quad (15)$$

In this expression, $B$ is the $\mathbb{P}_\alpha$ Brownian motion, and although this is the usual probabilistic statement of the Cameron-Martin-Girsanov theorem, it is not a natural expression to be used in statistical inference, and alternatives are necessary. (we want a variable measurable w.r.t the information generated by $V$).

For example, note that when $\sigma$ can be inverted, the expression can be considerably simplified:

$$\exp\left\{\int_0^t [(b-\alpha)^*\Gamma^{-1}](s, V_s)\mathrm{d}V_s \right. \\ \left. -\frac{1}{2}\int_0^t [(b-\alpha)^*\Gamma^{-1}(b+\alpha)](s, V_s)\mathrm{d}s\right\} . \tag{16}$$

For statistical inference about the drift (1) the Girsanov theorem is used with $\alpha = 0$. Any unknown parameters in the drift can be estimated by using (16) as a likelihood function. In practice, the integrals in the density are approximated by sums, leading to an error which can be controlled provided the data are available at arbitrarily high frequency.

# Framework for statistical inference in the high-freq regime

- Estimate $\sigma$ non parametrically using QV
- Plug in the estimate in (16) (with $\alpha = 0$) and use the latter as a likelihood function for $\theta$

This is the approach in [Polson and Roberts, 1994], who also discuss Bayesian approaches (using this likelihood framework), estimating Bayes factors for model comparison in simple financial applications (mean reverting models). In mainstream statistics, this was the state of the art for implementation of likelihood inference until mid-90s (with negative results for discretely observed data from the 80s)

# Appendix: understading Girsanov via Gaussian change of measure

The concept of change of measure is very central statistics and MC for stochastic processes. We give a simplified presentation of the change of measure between two Gaussian laws, and to the various ways this result might be put in use. It is easy to see the correspondence between the expressions we obtain here and those for diffuusions

Let $(\Omega, \mathcal{F})$ be a measure space with elements $\omega \in \Omega$, $B : \Omega \to R^m$ a random variable on that space, let $\sigma$ be a $d \times m$ matrix, $\Gamma = \sigma\sigma^*$, $a, b$, be $d \times 1$ vectors, and define a random variable $V$ via the equation

$$V(\omega) = b + \sigma B(\omega) \, .$$

Let $\mathbb{R}_b$ be the probability measure on $(\Omega, \mathcal{F})$ such that $B$ is a standard Gaussian vector. Therefore, under this measure $V$ is a Gaussian vector with mean $b$ (hence the indexing of the measure by $b$). Assume now that we can find a $m \times 1$ vector $h$ which solves the equation

$$\sigma h = (b - a)\,, \tag{17}$$

and define $\hat{B}(\omega) = B(\omega) + h$. Thus, we have the alternative representation

$$V(\omega) = a + \sigma \hat{B}(\omega)\,,$$

which follows directly from the definitions of $V$ and $h$.

Let $\mathbb{R}_a$ be the measure defined by its density with respect to $\mathbb{R}_b$,

$$\frac{\mathrm{d}\mathbb{R}_a}{\mathrm{d}\mathbb{R}_b}(\omega) = \exp\left\{-h^* B(\omega) - h^* h/2\right\} , \qquad (18)$$

which is well-defined since the right-hand side has finite expectation with respect to $\mathbb{R}_b$. Notice that under this new measure, $\hat{B}$ is a standard Gaussian vector. To see this, notice that for any Borel set $A \subset R^m$,

$$\begin{aligned}
\mathbb{R}_a[\hat{B} \in A] &= \int_{\{\omega:\hat{B}(\omega)\in A\}} \exp\left\{-u^* B(\omega) - u^* u/2\right\} \mathrm{d}\mathbb{R}[\omega] \\
&= \int_{\{y:y+u\in A\}} \exp\left\{-u^* y - u^* u/2 - y^* y/2\right\} (2\pi)^{-m/2} \mathrm{d}y \\
&= \int_A e^{-v^* v/2} (2\pi)^{-m/2} \mathrm{d}v \,,
\end{aligned}$$

where the last equality follows from a change of variables.

Notice that directly from (18) we have

$$\frac{\mathrm{d}\mathbb{R}_b}{\mathrm{d}\mathbb{R}_a}(\omega) = \exp\{h^* B(\omega) + h^* h/2\} = \exp\{h^* \hat{B}(\omega) - h^* h/2\}\,. \quad (19)$$

Let $\mathbb{E}_b$ and $\mathbb{E}_a$ denote expectations with respect to $\mathbb{R}_b$ and $\mathbb{R}_a$ respectively. Thus, for any measurable $\mathbb{R}_b$-integrable function $f$ defined on $R^d$,

$$\begin{aligned} \mathbb{E}_b[f(V)] &= \mathbb{E}_a\left[f(V)\exp\{h^* B + h^* h/2\}\right] \\ &= \mathbb{E}_a\left[f(V)\exp\{h^* \hat{B} - h^* h/2\}\right]\,. \end{aligned} \quad (20)$$

Let $X$ be another random variable, defined as $X(\omega) = a + \sigma B(\omega)$. Since under $\mathbb{R}_a$, the pair $(V, \hat{B})$ has the same law as the pair $(X, B)$ under $\mathbb{R}_b$, we have that

$$\mathbb{E}_b[f(V)] = \mathbb{E}_b[f(X) \exp\{h^* B - h^* h/2\}] \,.$$

If further $\sigma$ is invertible we get

$$\mathbb{E}_b[f(V)] = \mathbb{E}_b \left[ f(X) \exp\left\{ (b-a)^* \Gamma^{-1} X - \frac{1}{2}(b-a)^* \Gamma^{-1}(b+a) \right\} \right] \,. \tag{21}$$

Let $\mathbb{P}_b$ and $\mathbb{P}_a$ be the law of $V$ implied by $\mathbb{R}_b$ and $\mathbb{R}_a$ respectively. Then, assuming that $\sigma$ is invertible and taking $\alpha = 0$, we can obtain from the previous expression the likelihood ratio between the hypotheses that $V$ has mean $b$ against that it has mean 0, but a Gaussian distribution with covariance $\Gamma$ in both cases. Therefore, we get the likelihood function for estimating $b$ on the basis of observed data $V$, while treating $\Gamma$ as known:

$$L(b) = \frac{\mathrm{d}\mathbb{P}_b}{\mathrm{d}\mathbb{P}_0}(V) = \exp\left\{ b^*\Gamma^{-1}V - \frac{1}{2}b^*\Gamma^{-1}b \right\}. \tag{22}$$

which compares directly with (16)

# Part III: discrete-time dynamics of diffusions and discrete-time likelihood

- Transition density, discrete-time likelihood, approximate dynamics based on discretizations
- Pseudo-likelihood approaches
- Exact simulation of diffusions using rejection sampling on the path space (Girsanov and a transformation)