# Estimating Continuous-Time Models with Discretely Sampled Data*

Yacine Aït-Sahalia

Princeton University and NBER

First Draft: June 2005. This Version: August 13, 2006.

**Abstract**

This lecture surveys the recent literature on estimating continuous-time models using discrete observations. I start with the simplest possible framework and review different possible approaches as I progressively relax the assumptions made, to include different data generating processes (such as multivariate diffusions or jump processes), different observation schemes (such as incorporating market microstructure noise) and different sampling schemes (such as allowing for random time intervals.)

*Invited Lecture, 2005 World Congress of the Econometric Society. I am grateful to Oliver Linton, the discussant, for many useful comments. Financial support from the NSF under grant SBR-0350772 is also gratefully acknowledged.

# 1 Introduction

Since Merton's seminal work in the 1970s, the continuous-time paradigm has proved to be an immensely useful tool in finance and more generally economics. Continuous time models are widely used to study issues that include the decision to optimally consume, save and invest; portfolio choice under a variety of constraints; contingent claim pricing; capital accumulation; resource extraction; game theory and more recently contract theory. The objective of this lecture is to review some of the developments in the econometric literature devoted to the estimation and testing of these models.

The unifying theme of the class of the problems I will discuss is that the data generating process is assumed to be a continuous-time process describing the evolution of state variable(s), but the process is sampled, or observed, at discrete time intervals. The issues that arise, and the problems that are of interest, at the interface between the continuous-time model and the discrete-time data are quite different from those that we typically encounter in standard time series analysis. As a result, there has been a large amount of research activity in this area.

I will start with the simplest possible model, under many assumptions that restrict its generality, and describe how different inference strategies can be developed to work under progressively richer settings, where I relax either some aspect of the model's specification and/or the manner in which the process is sampled. I will attempt to systematize the treatment of different methods, present them in a unified framework and, when relevant, compare them. If nothing else, the reader should find these seemingly disparate methods presented using a common notation!

Let me apologize from the onset for the necessarily selective coverage that I will be able to achieve in this limited amount of space. Without any doubt, I will not be able to do justice to important strands of the literature and will instead put the emphasis on a selected set of methods – not surprisingly, those primarily associated with my own work!

# 2 The Simplest Model

The basic dynamic model for the state variable(s) of interest $X_t$ is a stochastic differential equation

$$dX_t = \mu dt + \sigma dW_t \tag{2.1}$$

where $W_t$ a standard Brownian motion. The process $X_t$ is observed at the discrete times $\tau_0 = 0$, $\tau_1$, $\tau_2$, ..., $\tau_N$, in the interval $[0, T]$, and I write the sampling intervals as $\Delta_n = \tau_n - \tau_{n-1}$. In a typical example, $X_t$ will represent the log-price of an asset.

I define the simplest model as one where the following assumptions are satisfied:

**Assumption 1.** *$X$ is a Markov process.*

**Assumption 2.** *$X$ is a diffusion.*

**Assumption 3.** *$X$ is a univariate process.*

**Assumption 4.** *$X$ has constant drift $\mu$ and diffusion $\sigma$.*

**Assumption 5.** *$X$ is sampled at a fixed time interval $\Delta = T/N$.*

**Assumption 6.** *$X$ is observed without error.*

For this model, any reasonable estimation strategy (MLE, OLS, GMM using the variance as a moment condition) ends up with the same estimator for the diffusion parameter $\sigma^2$. Let's write it to be even simpler for $\mu = 0$; otherwise one can simply subtract the natural estimator of $\mu$ from the (discrete) log-returns $X_{\tau_{n+1}} - X_{\tau_n}$ . Indeed, the log-returns in this case are i.i.d. $N(0, \sigma^2\Delta)$, and so the MLE for $\sigma^2$ is

$$\hat{\sigma}^2 \equiv T^{-1}[X,X]_T = T^{-1}\sum_{n=1}^{N}(X_{\tau_{n+1}} - X_{\tau_n})^2, \tag{2.2}$$

and has the asymptotic distribution

$$N^{1/2}\left(\hat{\sigma}^2 - \sigma^2\right) \underset{N\longrightarrow\infty}{\longrightarrow} N(0, 2\sigma^4). \tag{2.3}$$

# 3 More Realistic Assumptions

In each of the Sections below, one or more of these assumptions will be relaxed; as a general rule, the above assumptions that are not explicitly stated as being relaxed are maintained.

The rest of the lecture is devoted to exploring what can be done when the assumptions of the simplest model are relaxed. I will relax Assumption 1 to test whether the process is Markovian; relax Assumption 2 to include jumps and test for their presence; relax Assumption 3 to cover multivariate models; relax Assumption 4 to cover general parametric models as well as nonparametric models; relax Assumption 5 to allow for random sampling; and relax Assumption 6 to allow for market microstructure noise.

Each generalization falls into one of the following three classes: different data generating processes; different observation schemes; different sampling schemes.

## 3.1 Different Data Generating Processes

More general processes will be considered:

**Assumption 7.** *$X$ is a multivariate process of dimension $m$.*

**Assumption 8.** *$X$ has drift $\mu(x;\theta)$ and diffusion $\sigma(x;\theta)$ which are known functions except for an unknown parameter vector $\theta$.*

**Assumption 9.** *The drift and diffusion functions of $X$ are unknown time-homogenous smooth functions $\mu(x)$ and $\sigma(x)$.*

I will allow for departure from the continuity of the sample paths according to either:

**Assumption 10.** *$X$ is a jump-diffusion process with drift $\mu(x;\theta)$, diffusion $\sigma(x;\theta)$ and (possibly random) jump amplitude $J_t$. The arrival intensity of the Poisson jumps is $\lambda(x;\theta)$. The Brownian motion, Poisson process and jump amplitude are all independent.*

**Assumption 11.** *$X$ is a Lévy process with characteristics $(b, c, \nu)$ where $b$ is the drift, $\sigma$ the volatility from the continuous part and $v$ the Lévy measure from the pure jump part.*

In many but not all instances, it will be necessary to assume that:

**Assumption 12.** *The drift and diffusion functions of $X$, and its initial distribution (of $X_0$).are such that $X$ is a stationary process.*

## 3.2 Different Observation Schemes

Different assumptions can be made regarding the way(s) in which the process $X$ is observed:

**Assumption 13.** *All components of the vector $X$ are observed.*

**Assumption 14.** *$X_t = [S_t; V_t]'$, where the $(m-q)-$dimensional vector $S_t$ is observed but the $q-$dimensional $V_t$ is not.*

And, especially when dealing with high frequency financial data, it will be important to allow for the fact that $X$ may be observed with a fair amount of noise, as this is an essential feature of the observed data:

**Assumption 15.** *$X$ is observed with additive error, $Y_{\tau_n} = X_{\tau_n} + \varepsilon_{\tau_n}$.*

The next two assumptions represent the baseline case for modeling additive noise:

**Assumption 16.** *$\varepsilon$ is independent of $X$.*

**Assumption 17.** *$\varepsilon_{\tau_n}$ in Assumption 15 is i.i.d.*

They can be relaxed by allowing $\varepsilon$ to be correlated with the $X$ process, and by allowing $\varepsilon$ to be autocorrelated:

**Assumption 18.** *$\varepsilon_{\tau_n}$ in Assumption 15 is (when viewed as a process in index n) stationary and strong mixing with the mixing coefficients decaying exponentially; and, there exists $\kappa > 0$ such that $E\varepsilon^{4+\kappa} < \infty$.*

An alternative model for the noise is the following:

**Assumption 19.** *$X$ is observed with rounding error at some level $\alpha > 0$: $Y_t = \alpha[X_t/\alpha]$ where $[.]$ denotes the integer part of a real number.*

Finally, an additional source of error, beyond market microstructure noise captured by the above assumptions, is the fact that, under Assumption 7 and the components of the vector $X$ represent different asset prices, these assets may not be observed at the same instants because the assets do not trade, or see their quotes being revised, synchronously. This gives rise to:

**Assumption 20.** *The ith component of $X$, $X^{(i)}$, is observed at instants $\tau_0 = 0$, $\tau_1^{(i)}$, $\tau_2^{(i)}$, ..., $\tau_{n^{(i)}}^{(i)}$ which may not coincide for all $i = 1, ..., m$.*

### 3.3 Different Sampling Schemes

Instead of Assumption 5, the sampling intervals at which the observations are recorded can be of different types:

**Assumption 21.** *$X$ is sampled at an increasingly finer time interval, $\Delta \to 0$.*

**Assumption 22.** *The sampling intervals $\Delta_n = \tau_n - \tau_{n-1}$ are drawn at time $\tau_{n-1}$ so as to be conditionally independent of the future of the process $X$ given the past data $(Y_{n-1}, \Delta_{n-1}, Y_{n-2}, \Delta_{n-2}, ..., Y_1, \Delta_1)$ where $Y_n = X_{\tau_n}$.*

## 4 Relaxing the Model Specification: Allowing for Multivariate Parametric Dynamics

Of course, many models of interest require that we relax Assumption 4. To save space, I will describe how to conduct likelihood inference for arbitrary multivariate diffusion models, i.e., relax Assumption 3 at the same time. That is, suppose now that we are under Assumption 7 and 8. For now, also assume that the vector $X$ is fully observed so that Assumption 13 holds.

The SDE driving the $X$ process takes the multivariate parametric form

$$dX_t = \mu(X_t; \theta)dt + \sigma(X_t; \theta)dW_t \tag{4.1}$$

where $W_t$ is an $m-$dimensional standard Brownian motion. I will first discuss the maximum likelihood estimation of $\theta$ before turning to alternative, non-likelihood, approaches. When conducting likelihood inference, Assumption 12 is not necessary. Stationarity will however be a key requirement for some of the other methods.

### 4.1 Likelihood Inference

One major impediment to both theoretical modeling and empirical work with continuous-time models of this type is the fact that in most cases little can be said about the implications of the instantaneous dynamics for $X_t$ under these assumptions for longer time intervals. That is, unlike the simple situation where (2.1) implied that log-returns are i.i.d. and normally distributed, one cannot in general characterize in closed form an object as simple (and fundamental for everything from prediction to estimation and derivative pricing) as the conditional density of $X_{t+\Delta}$ given the current value $X_t$. For a list of the rare exceptions, see Wong (1964). In finance, the well-known models of Black and Scholes (1973), Vasicek (1977) and Cox et al. (1985) rely on these existing closed-form expressions.

In Aït-Sahalia (1999) (examples and application to interest rate data), Aït-Sahalia (2002b) (univariate theory) and Aït-Sahalia (2001) (multivariate theory), I developed a method which produces very accurate approximations *in closed form* to the unknown transition function $p_X(x|x_0, \Delta; \theta)$, that is, the conditional density of $X_{n\Delta} = x$ given $X_{(n-1)\Delta} = x_0$ implied by the model in equation (4.1). (When there is no risk of ambiguity, I will write simply $p$ instead of $p_X$; but, as will become clear below, the construction of $p_X$ requires a change of variable from $X$ to some $Y$ so it is best to indicate explicitly which variable's transition density we are talking about.)

Bayes' rule combined with the Markovian nature of the process, which the discrete data inherit, imply that the log-likelihood function has the simple form

$$\ell_N\left(\theta,\Delta\right)\equiv\sum\nolimits_{n=1}^{N}l_X\left(X_{n\Delta}|X_{(n-1)\Delta},\Delta;\theta\right) \tag{4.2}$$

where $l_X\equiv\ln p_X$, and the asymptotically irrelevant density of the initial observation, $X_0$, has been left out. As is clear from (4.2), the availability of tractable formulae for $l_X$ is what makes likelihood inference feasible under these conditions.

In the past, a computational challenge for such estimation was the tractable construction of a likelihood function since existing methods required solving numerically the Fokker-Planck-Kolmogorov partial differential equation (see e.g., Lo (1988)), or simulating a large number of sample paths along which the process is sampled very finely (Pedersen (1995)). Both classes of methods result in a large computational effort since the likelihood must be recomputed for each observed realization of the state vector, and each value of the parameter vector $\theta$ along the maximization. By contrast, the closed form likelihood expressions that I will summarize below make MLE a feasible choice for estimating $\theta$.

Jensen and Poulsen (2002), Stramer and Yan (2005) and Hurn et al. (2005) conducted extensive comparisons of different techniques for approximating transition function and demonstrated that the method is both the most accurate and the fastest to implement for the types of problems and sampling frequencies one encounters in finance. The method has been extended to time inhomogeneous processes by Egorov et al. (2003) and to jump-diffusions by Schaumburg (2001) and Yu (2003) (described in Section 8 below). DiPietro (2001) has extended the methodology to make it applicable in a Bayesian setting. Bakshi and Yu (2005) proposed a refinement to the method in the univariate case.

Identification of the parameter vector must be ensured. In fact, identifying a multivariate continuous-time Markov process from discrete-time data can be problematic when the process is not reversible, as an aliasing problem can be present (see Philips (1973) and Hansen and Sargent (1983).) As for the distributional properties of the resulting estimator, a fixed interval sample of a time-homogenous continuous-time Markov process is a Markov process in discrete time. Given that the Markov state vector is observed and the unknown parameters are identified, properties of the ML estimator follow from what is known about ML estimation of discrete-time Markov processes (see Billingsley (1961)). In the stationary case of Assumption 12, the MLE will under standard regularity conditions converge at speed $n^{1/2}$ to a normal distribution whose variance is given by the inverse of Fisher's information matrix. The nonstationary case is discussed below, in Section 7.

### 4.1.1 Reducibilty

Whenever possible, it is advisable to start with a change of variable. As defined in Aït-Sahalia (2001), a diffusion $X$ is *reducible* if and if only if there exists a one-to-one transformation of the diffusion $X$ into a diffusion $Y$ whose diffusion matrix $\sigma_Y$ is the identity matrix. That is, there exists an invertible function

$\gamma\left(x;\theta\right)$ such that $Y_t \equiv \gamma\left(X_t;\theta\right)$ satisfies the stochastic differential equation

$$dY_t = \mu_Y\left(Y_t;\theta\right)dt + dW_t. \tag{4.3}$$

Every univariate diffusion is reducible, through the simple transformation $Y_t = \int^{X_t} du/\sigma(u;\theta)$. The basic construction in Aït-Sahalia (2002b) in the univariate case proceeds by first changing $X$ into $Y$, and showing that the density of $Y$ can be approximated around a $N(0,1)$ distribution in the form of an expansion in Hermite polynomials. The unit diffusion in (4.3) is what yields the $N(0,1)$ distribution in the limit where $\Delta$ becomes small. The choice of Hermite (as opposed to other) polynomials is dictated by the fact that they are the orthonormal family in $L^2$ for the normal density. The key aspect is then that the coefficients of the expansion can be computed in closed form in the form of a Taylor expansion in $\Delta$.

However, not every multivariate diffusion is reducible. As a result, the passage from the univariate to the multivariate cases does not just reduce to a simple matter of scalars becoming matrices. Whether or not a given multivariate diffusion is reducible depends on the specification of its $\sigma$ matrix. Proposition 1 of Aït-Sahalia (2001) provides a necessary and sufficient condition for reducibility: the diffusion $X$ is reducible if and only if the inverse diffusion matrix $\sigma^{-1} = \left[\sigma^{-1}_{i,j}\right]_{i,j=1,\ldots,m}$ satisfies the condition that

$$\frac{\partial\sigma^{-1}_{ij}\left(x;\theta\right)}{\partial x_k} = \frac{\partial\sigma^{-1}_{ik}\left(x;\theta\right)}{\partial x_j} \tag{4.4}$$

for each triplet $(i,j,k) = 1,\ldots,m$ such that $k > j$, or equivalently

$$\sum_{l=1}^{m}\frac{\partial\sigma_{ik}\left(x;\theta\right)}{\partial x_l}\sigma_{lj}\left(x;\theta\right) = \sum_{l=1}^{m}\frac{\partial\sigma_{ij}\left(x;\theta\right)}{\partial x_l}\sigma_{lk}\left(x;\theta\right). \tag{4.5}$$

Whenever a diffusion is reducible, an expansion can be computed for the transition density $p_X$ of $X$ by first computing it for the density $p_Y$ of $Y$ and then transforming $Y$ back into $X$ . When a diffusion is not reducible, the situation is more involved (see Section 4.1.2 below). The expansion for $l_Y$ is of the form

$$l_Y^{(K)}\left(y|y_0,\Delta;\theta\right) = -\frac{m}{2}\ln\left(2\pi\Delta\right) + \frac{C_Y^{(-1)}\left(y|y_0;\theta\right)}{\Delta} + \sum\nolimits_{k=0}^{K} C_Y^{(k)}\left(y|y_0;\theta\right)\frac{\Delta^k}{k!}. \tag{4.6}$$

As shown in Theorem 1 of Aït-Sahalia (2001), the coefficients of the expansion are given explicitly by:

$$C_Y^{(-1)}\left(y|y_0;\theta\right) = -\frac{1}{2}\sum\nolimits_{i=1}^{m}\left(y_i - y_{0i}\right)^2 \tag{4.7}$$

$$C_Y^{(0)}\left(y|y_0;\theta\right) = \sum\nolimits_{i=1}^{m}\left(y_i - y_{0i}\right)\int_0^1 \mu_{Yi}\left(y_0 + u\left(y - y_0\right);\theta\right)du \tag{4.8}$$

and, for $k \geq 1$,

$$C_Y^{(k)}\left(y|y_0;\theta\right) = k\int_0^1 G_Y^{(k)}\left(y_0 + u\left(y - y_0\right)|y_0;\theta\right)u^{k-1}du \tag{4.9}$$

where $G_Y^{(k)}\left(y|y_0;\theta\right)$ is given explicitly as a function of $\mu_{Yi}$ and $C_Y^{(j-1)}$, $j = 1,\ldots,k$.

Given an expansion for the density $p_Y$ of $Y$, an expansion for the density $p_X$ of $X$ can be obtained by a direct application of the Jacobian formula:

$$l_X^{(K)}\left(x|x_0,\Delta;\theta\right) = -\frac{m}{2}\ln\left(2\pi\Delta\right) - D_v\left(x;\theta\right) + \frac{C_Y^{(-1)}\left(\gamma\left(x;\theta\right)|\gamma\left(x_0;\theta\right);\theta\right)}{\Delta}$$
$$+\sum\nolimits_{k=0}^{K} C_Y^{(k)}\left(\gamma\left(x;\theta\right)|\gamma\left(x_0;\theta\right);\theta\right)\frac{\Delta^k}{k!} \tag{4.10}$$

from $l_Y^{(K)}$ given in (4.6), using the coefficients $C_Y^{(k)}$, $k = -1, 0, ..., K$ given above, and where

$$D_v\left(x;\theta\right) \equiv \frac{1}{2}\ln\left(Det[v(x;\theta)]\right). \tag{4.11}$$

with $v\left(x;\theta\right) \equiv \sigma\left(x;\theta\right)\sigma^T\left(x;\theta\right)$.

### 4.1.2 The Irreducible Case

In the irreducible case, no such $Y$ exists and the expansion of the log likelihood $l_X$ has the form

$$l_X^{(K)}\left(x|x_0,\Delta;\theta\right) = -\frac{m}{2}\ln\left(2\pi\Delta\right) - D_v\left(x;\theta\right) + \frac{C_X^{(-1)}\left(x|x_0;\theta\right)}{\Delta} + \sum\nolimits_{k=0}^{K} C_X^{(k)}\left(x|x_0;\theta\right)\frac{\Delta^k}{k!}. \tag{4.12}$$

The approach is to calculate a Taylor series in $(x - x_0)$ of each coefficient $C_X^{(k)}$, at order $j_k$ in $(x - x_0)$. Such an expansion will be denoted by $C_X^{(j_k,k)}$ at order $j_k = 2(K - k)$, for $k = -1, 0, ..., K$.

The resulting expansion will then be

$$\tilde{l}_X^{(K)}\left(x|x_0,\Delta;\theta\right) = -\frac{m}{2}\ln\left(2\pi\Delta\right) - D_v\left(x;\theta\right) + \frac{C_X^{(j_{-1},-1)}\left(x|x_0;\theta\right)}{\Delta} + \sum\nolimits_{k=0}^{K} C_X^{(j_k,k)}\left(x|x_0;\theta\right)\frac{\Delta^k}{k!} \tag{4.13}$$

Such a Taylor expansion was unnecessary in the reducible case: the expressions given above provide the explicit expressions of the coefficients $C_Y^{(k)}$ and then in (4.10) we have the corresponding ones for $C_X^{(k)}$. However, even for an irreducible diffusion, it is still possible to compute the coefficients $C_X^{(j_k,k)}$ explicitly.

As in the reducible case, the system of equations determining the coefficients is obtained by forcing the expansion (4.12) to satisfy, to order $\Delta^J$, the forward and backward Fokker-Planck-Kolmogorov equations, either in their familiar form for the transition density $p_X$, or in their equivalent form for $\ln p_X$. For instance, the forward equation for $\ln p_X$ is of the form:

$$\frac{\partial l_X}{\partial\Delta} = -\sum_{i=1}^{m}\frac{\partial\mu_i^P(x)}{\partial x_i} + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\partial^2\nu_{ij}(x)}{\partial x_i\partial x_j} + \sum_{i=1}^{m}\mu_i^P(x)\frac{\partial l_X}{\partial x_i} + \sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\partial\nu_{ij}(x)}{\partial x_i}\frac{\partial l_X}{\partial x_j}$$
$$+\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\nu_{ij}(x)\frac{\partial^2 l_X}{\partial x_i\partial x_j} + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\frac{\partial l_X}{\partial x_i}\nu_{ij}(x)\frac{\partial l_X}{\partial x_j}. \tag{4.14}$$

For each $k = -1, 0, ..., K$, the coefficient $C_X^{(k)}\left(x|x_0;\theta\right)$ in (4.12) solves the equation

$$f_X^{(k-1)}\left(x|x_0;\theta\right) = C_X^{(k)}\left(x|x_0;\theta\right) - \sum\nolimits_{i=1}^{m}\sum\nolimits_{j=1}^{m} v_{ij}\left(x;\theta\right)\frac{\partial C_X^{(-1)}\left(x|x_0;\theta\right)}{\partial x_i}\frac{\partial C_X^{(k)}\left(x|x_0;\theta\right)}{\partial x_j} - G_X^{(k)}\left(x|x_0;\theta\right) = 0$$

Each function $G_X^{(k)}$, $k = 0, 1, ..., K$ involves only the coefficients $C_X^{(h)}$ for $h = -1, ..., k - 1$, so this system of equation can be utilized to solve recursively for each coefficient at a time. Specifically, the equation $f_X^{(-2)} = 0$

determines $C_X^{(-1)}$; given $C_X^{(-1)}$, $G_X^{(0)}$ becomes known and the equation $f_X^{(-1)} = 0$ determines $C_X^{(0)}$; given $C_X^{(-1)}$ and $C_X^{(0)}$, $G_X^{(1)}$ becomes known and the equation $f_X^{(0)} = 0$ then determines $C_X^{(1)}$, etc. It turns out that this results in a system of linear equations in the coefficients of the polynomials $C_X^{(j_k,k)}$, so each one of these equations can be solved explicitly in the form of the Taylor expansion $C_X^{(j_k,k)}$ of the coefficient $C_X^{(k)}$, at order $j_k$ in $(x - x_0)$.

Aït-Sahalia and Yu (2005) developed an alternative strategy for constructing closed form approximations to the transition density of a continuous time Markov process. Instead of expanding the transition function in orthogonal polynomials around a leading term, we rely on the saddlepoint method, which originates in the work of Daniels (1954). We showed that, in the case of diffusions, it is possible by expanding the cumulant generating function of the process to obtain an alternative closed form expansion of its transition density. We also showed that this approach provides an alternative gathering of the correction terms beyond the leading term that is equivalent at order $\Delta$ to the irreducible expansion of the transition density given in Aït-Sahalia (2001).

## 4.2 Alternative Non-Likelihood Methods

Because of the Cramer-Rao lower bound, maximum likelihood estimation will, at least asymptotically, be efficient and should consequently be the method of choice especially given the availability of closed-form, numerically tractable, likelihood expansions. There are, however, a number of alternative methods available to estimate $\theta$.

### 4.2.1 Moment Conditions, Contrast Functions and Estimating Equations

Consider a class of estimators for the parameter vector $\theta$ under Assumption 8 obtained by minimizing a quadratic criterion function, or equivalent setting to zero its $\theta-$gradient. To estimate the $d$-dimensional parameter vector $\theta$, we can select a vector of $r$ moment conditions $h(y_1, y_0, \delta, \theta)$, $r \geq d$, which is continuously differentiable in $\theta$. Here $y_1$ plays the role of the forward state variable, $y_0$ the role of the backward state variable, $\delta$ the sampling interval.

At this level of generality, of course, everything is possible. The question is to find actual functions $h$ which are tractable. In particular to avoid any bias in the resulting estimator, $h$ must be such that

$$E\left[h(Y_1, Y_0, \Delta, \theta_0)\right] = 0. \tag{4.15}$$

Whenever $h$ depends on both $Y_1$ and $Y_0$, in a non-trivial manner, this requires that we be able to compute the expected value of a candidate $h$ under the law of $Y_1|Y_0$ (by first conditioning on $Y_0$ and applying the law of iterated expectations), i.e., under the transition density $p(Y_1|Y_0, \Delta; \theta)$. This is a tough requirement to fulfill in closed form, and one which severely limits the types of moment conditions which can be implemented.

Hansen and Scheinkman (1995) provide the only example of $h$ functions which can be computed under a general specification of Assumption 8 (as opposed to just for special cases of $\mu$ and $\sigma$, such as the Ornstein-Uhlenbeck process), and in fact apply under more general Markov processes (i.e., allow us to relax Assumption

2). These moment conditions are in the form of expectations of the infinitesimal generator, one unconditional (C1) and one conditional (C2), that are applied to test functions. Stationarity, i.e., Assumption 12, is a crucial requirement of the method.

The C1 moment condition takes a sufficiently differentiable function $\psi(y_0, \theta)$ in the domain of the infinitesimal generator $A_\theta$ and forms the moment condition

$$h_{C1}(y_1, y_0, \delta, \theta) \equiv A_\theta \cdot \psi(y_0, \theta) \equiv \mu(y_0, \theta) \frac{\partial \psi}{\partial y_0} + \frac{1}{2} \sigma^2(y_0; \theta) \frac{\partial^2 \psi}{\partial y_0^2}, \tag{4.16}$$

with the last equality written for convenience under Assumption 3.

The C1 estimating equation relies on the fact that we have the unbiasedness condition

$$E\left[A_{\theta_0} \cdot \psi(Y_0, \theta)\right] = 0. \tag{4.17}$$

Once $A_\theta$ is evaluated at $\theta_0$, this is true for any value of $\theta$ in $\psi$, including $\theta_0$. A consequence of this is that the resulting estimator is unbiased for any function $\psi$ (recall (4.15)).

The C2 method takes two functions $\psi_0$ and $\psi_1$, again satisfying smoothness and regularity conditions, and forms the "back to the future" estimating function

$$h_{C2}(y_1, y_0, \delta, \theta) \equiv \{A_\theta \cdot \psi_1(y_1, \theta)\} \times \psi_0(y_0, \theta) - \{A_\theta \cdot \psi_0(y_0, \theta)\} \times \psi_1(y_1, \theta). \tag{4.18}$$

In general, the second of the two $A_\theta$ should be replaced by the infinitesimal generator associated with the reverse time process, $A_\theta^*$. But under regularity conditions, univariate stationary diffusions are time reversible (see Kent (1978)) and so the infinitesimal generator of the process is self-adjoint .

The C2 estimating equation relies on the fact that

$$E\left[\{A_{\theta_0} \cdot \psi_1(Y_1, \theta)\} \times \psi_0(Y_0, \theta) - \{A_{\theta_0} \cdot \psi_0(Y_0, \theta)\} \times \psi_1(Y_1, \theta)\right] = 0. \tag{4.19}$$

Once $A_\theta$ is evaluated at $\theta_0$, this is true for any value of $\theta$ in $\psi$, including $\theta_0$.

Equation (4.19) is again a consequence of the stationarity of the process $X$. Namely, the expectation of any function of $(X_t, X_{t+\delta})$, such as $E_{X_t, X_{t+\delta}}\left[\psi_0(X_t, \theta)\psi_1(X_{t+\delta}, \theta)\right]$, does not depend upon the date $t$ (it can of course depend upon the time lag $\delta$ between the two observations): hence

$$\frac{\partial}{\partial t} E_{X_t, X_{t+\delta}}\left[\psi_0(X_t, \theta)\psi_1(X_{t+\delta}, \theta)\right] = 0$$

from which (4.19) follows.

Unlike the typical use of the method of moments, one cannot in general select as moment conditions the "natural" conditional moments of the process since explicit expressions for the conditional mean, variance, skewness, etc. are not available in closed-form. Rather, the moment conditions in this case, i.e., the $h$ functions, are in the form of the infinitesimal generator of the process applied to arbitrary test functions. One additional aspect of the method is that it does not permit full identification of all the parameters of the model since multiplying the drift and diffusion functions by the same constant results in identical moment conditions. So parameters are only identified up to scale.

Bibby and Sørensen (1995) proposed moment conditions in the form

$$h(y_1, y_0, \delta, \theta) = \frac{\partial \mu \left(y_0; \theta\right) / \partial\theta\partial\theta'}{\sigma^2 \left(y_0; \theta\right)} \left(y_1 - E\left[Y_1 | Y_0 = y_0\right]\right) \tag{4.20}$$

By construction, (4.15), but except in special cases, $E\left[Y_1|Y_0 = y_0\right]$ does not have an explicit form, and any approximation of misspecification of it will induce bias.

Kessler and Sørensen (1999) proposed to use a more general version of (4.20), namely

$$h(y_1, y_0, \delta, \theta) = a\left(y_0, \delta, \theta\right) \left(\phi\left(y_1, \theta\right) - E\left[\phi\left(Y_1, \theta\right)|Y_0 = y_0\right]\right) \tag{4.21}$$

where $a$ is a weighting function, and to select the function $\phi$ in such a way as to make the computation of $E\left[\phi\left(Y_1, \theta\right)|Y_0 = y_0\right]$ tractable. This can be achieved by considering the eigenfunctions of the infinitesimal operator: let $\lambda_j$ and $\phi_j$, $j = 0, 1, ...$, denote respectively the eigenvalues and eigenfunctions of the operator $A_\theta$ defined by

$$A_\theta \cdot \phi_j(y_0, \theta) = -\lambda_j(\theta)\phi_j(y_0, \theta)$$

This will imply that

$$E\left[\phi\left(Y_1, \theta\right)|Y_0 = y_0\right] = \exp\left(-\lambda_j(\theta)\Delta\right)\phi_j(y_0, \theta)$$

(for regularity conditions and related material on the infinitesimal generator, see e.g., Aït-Sahalia et al. (2002)). But, unfortunately, for the general parametric model of Assumption 8, these moment functions are not explicit either, because the $(\lambda_j, \phi_j)$'s cannot be computed in closed form except in special cases. Bibby et al. (2002) provides a detailed review of these and other types of estimating equations.

### 4.2.2 Simulation-Based Methods of Moments

A different type of moment functions is obtained by simulation. The moment function is the expected value under the model whose parameters one wishes to estimate of the score function from an auxiliary model. The idea is that the auxiliary model should be easier to estimate; in particular, the parameters of the auxiliary model appearing in the moment function are replaced by their quasi-maximum likelihood estimates. The parameters of the model of interest are then estimated by minimizing a GMM-like criterion. This method has the advantage of being applicable to a wide class of models. On the other hand, the estimates are obtained by simulations which are computationally intensive and they require the selection of an arbitrary auxiliary model, so the parameter estimates for the model of interest will vary based on both the set of simulations and the choice of auxiliary model.

Related implementations of this idea are due to Smith (1993), Gouriéroux et al. (1993) and Gallant and Tauchen (1996); see Gallant and Tauchen (2002) for a survey of this literature.

### 4.2.3 Two-Step Estimation Based on the Continuous Record Likelihood

Phillips and Yu (2005) proposed the following method. Suppose that the parameter vector is split according to $\theta = (\theta_1, \theta_2)'$ where $\mu$ depends on $\theta$ through $\theta_1$ only and $\sigma$ through $\theta_2$ only. If the full continuous time path were observable, and $\theta_2$ were known, then the log-likelihood for $\theta_1$ is given by Girsanov's Theorem:

$$\ell_n(\theta_1) = \int_0^T \frac{\mu(X_t, \theta_1)}{\sigma^2(X_t, \theta_2)} dX_t - \frac{1}{2}\int_0^T \frac{\mu^2(X_t, \theta_1)}{\sigma^2(X_t, \theta_2)} dt \tag{4.22}$$

Suppose now that we also have $\sigma(X_t, \theta_2) = \theta_2$ constant. Then the quadratic variation of the process is $\langle X, X\rangle_T = \int_0^T \sigma^2(X_t, \theta_2)\, dt = \theta_2^2 T$. And the realized volatility $[X, X]_T$, defined as in (2.2), makes it possible to estimate $\theta_2$ as $\hat{\theta}_2 = T^{-1/2}[X, X]_T^{1/2}$. Plug that into the continuous-record likelihood

$$\ell_n(\theta_1) = \int_0^T \frac{\mu(X_t, \theta_1)}{\sigma^2\left(X_t, \hat{\theta}_2\right)} dX_t - \frac{1}{2}\int_0^T \frac{\mu^2(X_t, \theta_1)}{\sigma^2\left(X_t, \hat{\theta}_2\right)} dt,$$

approximate the integral with a sum

$$\sum_{i=2}^{n} \frac{\mu\left(X_{(i-1)\Delta}, \theta_1\right)}{\sigma^2\left(X_{(i-1)\Delta}, \hat{\theta}_2\right)} \left(X_{i\Delta} - X_{(i-1)\Delta}\right) - \frac{1}{2}\sum_{i=2}^{n} \frac{\mu^2\left(X_{(i-1)\Delta}, \theta_1\right)}{\sigma^2\left(X_{(i-1)\Delta}, \hat{\theta}_2\right)} \Delta$$

and maximize over $\theta_1$. Approximations to the continuous-record likelihood are discussed in Sørensen (2001).

When $\sigma(X_t, \theta_2) = \theta_2 f(X_t)$ and $f$ is a known function, the same method applies: $\langle X, X\rangle_T = \int_0^T \sigma^2(X_t, \theta_2)\, dt$ $= \theta_2^2 \int_0^T f(X_t)\, dt$, estimate $\int_0^T f(X_t)\, dt$ using $\sum_{i=2}^n f\left(X_{(i-1)\Delta}\right)\Delta$, and using $[X, X]_T$ to estimate $\langle X, X\rangle_T$, we have an estimator of $\theta_2$ again. For more general $\sigma$ functions, one can estimate $\theta_2$ by minimizing a scaled distance (for instance quadratic) between $\langle X, X\rangle_T = \int_0^T \sigma^2(X_t, \theta_2)\, dt$ (or rather a sum that approximates the integral) and $[X, X]_T$ inside each subinterval; then proceed as before by plugging that into the continuous record likelihood and maximize over $\theta_1$.

# 5 Relaxing the Model Specification: Allowing for a Partially Observed State Vector

Under Assumption 13, the vector $X_t$ is fully observed. This assumption can be violated in many practical situations when some components of the state vector are latent. I now examine likelihood inference for models where Assumption 13 is replaced by Assumption 14. Unlike likelihood-based inference, however, relatively little changes for the simulation-based methods described in Section 4.2.2 above; they remain applicable in this context (although without asymptotic efficiency). Estimating equations that are applicable to partially observed state vectors are discussed in Genon-Catalot et al. (1999).

Two typical examples in finance consist of stochastic volatility models, where $V_t$ is the volatility state variable(s), and term structure models, where $V_t$ is a vector of factors or yields. I will discuss how to conduct likelihood inference in this setting, without resorting to the statistically sound but computationally infeasible integration of the latent variables from the likelihood function. The idea is simple: write down in closed form an expansion for the log-likelihood of the state vector $X$, including its unobservable components. Then enlarge the observation state by adding variables that are observed and functions of $X$. For example, in the stochastic volatility case, an option price or an option-implied volatility; in term structure models, as many bonds as there are factors. Then, using the Jacobian formula, write down the likelihood function of the pair consisting of the observed components of $X$ and the additional observed variables, and maximize it.

## 5.1 Likelihood Inference for Stochastic Volatility Models

In a stochastic volatility model, the asset price process $S_t$ follows

$$dS_t = (r - \delta) S_t dt + \sigma_1 (X_t; \theta) dW_t^Q \tag{5.1}$$

where $r$ is the riskfree rate, $\delta$ is the dividend yield paid by the asset (both taken to be constant for simplicity only), $\sigma_1$ denotes the first row of the matrix $\sigma$ and $Q$ denotes the equivalent martingale measure (see e.g., Harrison and Kreps (1979)). The volatility state variables $V_t$ then follow a SDE on their own. For example, in the Heston (1993) model, $m = 2$ and $q = 1$:

$$dX_t = d \begin{bmatrix} S_t \\ V_t \end{bmatrix} = \begin{bmatrix} (r - \delta) S_t \\ \kappa (\gamma - V_t) \end{bmatrix} dt + \begin{bmatrix} \sqrt{(1-\rho^2) V_t} S_t & \rho \sqrt{V_t} S_t \\ 0 & \sigma \sqrt{V_t} \end{bmatrix} d \begin{bmatrix} W_1^Q(t) \\ W_2^Q(t) \end{bmatrix} \tag{5.2}$$

The model is completed by the specification of a vector of market prices of risk for the different sources of risk ($W_1$ and $W_2$ here), such as

$$\Lambda (X_t; \theta) = \left[ \lambda_1 \sqrt{(1-\rho^2) V_t}, \ \lambda_2 \sqrt{V_t} \right]', \tag{5.3}$$

which characterizes the change of measure from $Q$ back to the physical probability measure $P$.

Likelihood inference for this and other stochastic volatility models is discussed in Aït-Sahalia and Kimmel (2004). Given a time series of observations of both the asset price, $S_t$, and a vector of option prices (which, for simplicity, we take to be call options) $C_t$, the time series of $V_t$ can then be inferred from the observed $C_t$. If $V_t$ is multidimensional, sufficiently many options are required with varying strike prices and maturities to allow extraction of the current value of $V_t$ from the observed stock and call prices. Otherwise, only a single option is needed. For reasons of statistical efficiency, we seek to determine the joint likelihood function of the observed data, as opposed to, for example, conditional or unconditional moments. We employ the closed-form approximation technique described above, which yields in closed form the joint likelihood function of $[S_t; V_t]'$. From there, the joint likelihood function of the observations on $G_t = [S_t; C_t]' = f(X_t; \theta)$ is obtained simply by multiplying the likelihood of $X_t = [S_t; V_t]'$ by the Jacobian term $J_t$:

$$\ln p_G (g|g_0, \Delta; \theta) = -\ln J_t (g|g_0, \Delta; \theta) + l_X (f^{-1}(g; \theta) | f^{-1}(g_0; \theta); \Delta, \theta) \tag{5.4}$$

with $l_X$ obtained in Section 4.

If a proxy for $V_t$ is used directly, this last step is not necessary. Indeed, we can avoid the computation of the function $f$ by first transforming $C_t$ into a proxy for $V_t$. The simplest one consists in using the Black-Scholes implied volatility of a short-maturity at-the-money option in place of the true instantaneous volatility state variable. The use of this proxy is justified in theory by the fact that the implied volatility of such an option converges to the instantaneous volatility of the logarithmic stock price as the maturity of the option goes to zero. An alternate proxy (which we call the integrated volatility proxy) corrects for the effect of mean reversion in volatility during the life of an option. If $V_t$ is the instantaneous variance of the logarithmic stock price, we can express the integral of variance from time $t$ to $T$ as

$$V\left(t,T\right)=\int_{t}^{T}V_{u}du \tag{5.5}$$

If the volatility process is instantaneously uncorrelated with the logarithmic stock price process, then we can calculate option prices by taking the expected value of the Black-Scholes option price (with $V(t,T)$ as implied variance) over the probability distribution of $V(t,T)$ (see Hull and White (1987)). If the two processes are correlated, then the price of the option is a weighted average of Black Scholes prices evaluated at different stock prices and volatilities (see Romano and Touzi (1997)).

The proxy we examine is determined by calculating the expected value of $V(t,T)$ first, and substituting this value into the Black-Scholes formula as implied variance. This proxy is model-free, in that it can be calculated whether or not an exact volatility can be computed and results in a straightforward estimation procedure. On the other hand, this procedure is in general approximate, first because the volatility process is unlikely to be instantaneously uncorrelated with the logarithmic stock price process, and second, because the expectation is taken before substituting $V(t,T)$ into the Black-Scholes formula rather than after and we examine in Monte Carlo simulations the respective impact of these approximations, with the objective of determining whether the trade-off involved between simplicity and exactitude is worthwhile.

The idea is to adjust the Black-Scholes implied volatility for the effect of mean reversion in volatility, essentially undoing the averaging that takes place in equation (5.5). Specifically, if the $Q$-measure drift of $Y_t$ is of the form $a+bY_t$ (as it is in many of the stochastic volatility models in use), then the expected value of $V(t,T)$ is given by:

$$E_{t}\left[V\left(t,T\right)\right]=\left(\frac{e^{b(T-t)}-1}{b}\right)\left(V_{t}+\frac{a}{b}\right)-\frac{a}{b}\left(T-t\right) \tag{5.6}$$

A similar expression can be derived in the special case where $b=0$. By taking the expected value on the left-hand side to be the observed implied variance $V_{\text{imp}}(t,T)$ of a short maturity $T$ at-the-money option, our adjusted proxy is then given by:

$$V_{t}\approx\frac{bV_{\text{imp}}(t,T)+a\left(T-t\right)}{e^{b(T-t)}-1}-\frac{a}{b}. \tag{5.7}$$

Then we can simply take $[S_t;V_{\text{imp}}(t,T)]'$ as the state vector, write its likelihood from that of $[S_t;V_t]'$ using a Jacobian term for the change of variable (5.7).

It is possible to refine the implied volatility proxy by expressing it in the form of a Taylor series in the "volatility of volatility" parameter $\sigma$ in the case of the CEV model, where the $Q$-measure drift of $Y_t$ is of the form $a+bY_t$, and the $Q$-measure diffusion of $Y_t$ is of the form $\sigma Y_t^{\beta}$ (Lewis (2000)). However, unlike (5.7), the relationship between the observed $V_{\text{imp}}(t,T)$ and the latent $Y_t$ is not invertible without numerical computation of the parameter-dependent integral.

## 5.2 Likelihood Inference for Term Structure Models

Another example of model for which inference must be conducted under Assumption 14 consist of term structure models. A multivariate term structure model specifies that the instantaneous riskless rate $r_t$ is a deterministic function of an $m-$dimensional vector of state variables, $X_t$:

$$r_t = r\left(X_t;\theta\right). \tag{5.8}$$

Under the equivalent martingale measure $Q$, the state vector follows the dynamics given in (4.1). In order to avoid arbitrage opportunities, the price at $t$ of a zero-coupon bond maturing at $T$ is given by the Feynman-Kac representation:

$$P\left(x,t,T;\theta\right) = E^{Q}\left[\exp\left(-\int_t^T r_u du\right)\middle| X_t = x\right] \tag{5.9}$$

An affine yield model is any model where the short rate (5.8) is an affine function of the state vector and the risk-neutral dynamics (4.1) are affine:

$$dX_t = \left(\tilde{A} + \widetilde{B}X_t\right)dt + \Sigma\sqrt{S\left(X_t;\alpha,\beta\right)}dW_t^Q \tag{5.10}$$

where $\tilde{A}$ is an $m$–element column vector, $\widetilde{B}$ and $\Sigma$ are $m \times m$ matrices, and $S\left(X_t;\alpha,\beta\right)$ is the diagonal matrix with elements $S_{ii} = \alpha_i + X_t'\beta_i$, with each $\alpha_i$ a scalar and each $\beta_i$ an $m \times 1$ vector, $1 \leq i \leq m$ (see Dai and Singleton (2000)).

It can then be shown that, in affine models, bond prices have the exponential affine form

$$P\left(x,t,T;\theta\right) = \exp\left(-\gamma_0\left(\tau;\theta\right) - \gamma\left(\tau;\theta\right)'x\right) \tag{5.11}$$

where $\tau = T - t$ is the bond's time to maturity. That is, bond yields (non-annualized, and denoted by $g\left(x,t,T;\theta\right) = -\ln\left(P\left(x,t,T;\theta\right)\right)$) are affine functions of the state vector:

$$g(x,t,T;\theta) = \gamma_0\left(\tau;\theta\right) + \gamma\left(\tau;\theta\right)'x. \tag{5.12}$$

Alternatively, one can start with the requirement that the yields be affine, and show that the dynamics of the state vector must be affine (see Duffie and Kan (1996)).

The final condition for the bond price implies that $\gamma_0\left(0;\theta\right) = \gamma\left(0;\theta\right) = 0$, while

$$r_t = \delta_0 + \delta'x. \tag{5.13}$$

Affine yield models owe much of their popularity to the fact that bond prices can be calculated quickly as solutions to a system of ordinary differential equations. Under non-linear term structure models, bond prices will normally be solutions to a partial differential equation that is far more difficult to solve.

Aït-Sahalia and Kimmel (2002) consider likelihood inference for affine term structure models. This requires evaluation of the likelihood of an observed panel of yield data for each parameter vector considered during a search procedure. The procedure for evaluating the likelihood of the observed yields at a particular value of the parameter vector consists of four steps. First, we extract the value of the state vector $X_t$ (which is not directly observed) from those yields that are treated as observed without error. Second, we evaluate the joint likelihood of the series of implied observations of the state vector $X_t$, using the closed-form approximations to the likelihood function described in Section 4. Third, we multiply this joint likelihood by a Jacobian term, to find the likelihood of the panel of observations of the yields observed without error. Finally, we calculate the likelihood of the observation errors for those yields observed with error, and multiply this likelihood by the likelihood found in the previous step, to find the joint likelihood of the panel of all yields.

The first task is therefore to infer the state vector $X_t$ at date $t$ from the cross-section of bond yields at date $t$ with different maturities. Affine yield models, as their name implies, make yields of zero coupon bonds affine functions of the state vector. Given this simple relationship between yields and the state vector, the likelihood function of bond yields is a simple transformation of the likelihood function of the state vector.

If the number of observed yields at that point in time is smaller than the number $N$ of state variables in the model, then the state is not completely observed, and the vector of observed yields does not follow a Markov process, even if the (unobserved) state vector does, enormously complicating maximum likelihood estimation. On the other hand, if the number of observed yields is larger than the number of state variables, then some of the yields can be expressed as deterministic functions of other observed yields, without error. Even tiny deviations from the predicted values have a likelihood of zero. This problem can be avoided by using a number of yields exactly equal to the number of state variables in the underlying model, but, in general, the market price of risk parameters will not all be identified. Specifically, there are affine yield models that generate identical dynamics for yields with a given set of maturities, but different dynamics for yields with other maturities. A common practice (see, for example, Duffee (2002)) is to use more yields than state variables, and to assume that certain benchmark yields are observed precisely, whereas the other yields are observed with measurement error. The measurement errors are generally held to be i.i.d., and also independent of the state variable processes.

We take this latter approach, and use $N+H$ observed yields, $H \geq 0$, in the postulated model, and include observation errors for $H$ of those yields. At each date $t$, the state vector $X_t$ is then exactly identified by the yields observed without error, and these $N$ yields jointly follow a Markov process. Denoting the times to maturity of the yields observed without error as $\tau_1, ..., \tau_N$, the observed values of these yields, on the left-hand side, are equated with the predicted values (from (5.12)) given the model parameters and the current values of the state variables, $X_t$:

$$g_t = \Gamma_0(\theta) + \Gamma(\theta)' X_t. \tag{5.14}$$

The current value of the state vector $X_t$ is obtained by inverting this equation:

$$X_t = \left[\Gamma(\theta)'\right]^{-1}\left[g_t - \Gamma_0(\theta)\right]. \tag{5.15}$$

While the only parameters entering the transformation from observed yields to the state variables are the parameters of the risk-neutral (or $Q$-measure) dynamics of the state variables, once we have constructed our time series of values of $X_t$ sampled at dates $\tau_0$, $\tau_1, ..., \tau_n$, the dynamics of the state variable that we will be able to infer from this time series are the dynamics under the physical measure (denoted by $P$). The first step in the estimation procedure is the only place where we rely on the tractability of the affine bond pricing model. In particular, we can now specify freely (that is, without regard for considerations of analytical tractability) the market prices of risk of the different Brownian motions

$$\begin{aligned} dX_t &= \mu^P\left(X_t;\theta\right)dt+\sigma\left(X_t;\theta\right)dW_t^P \\ &= \left\{\mu^Q\left(X_t;\theta\right)+\sigma\left(X_t;\theta\right)\Lambda\left(X_t;\theta\right)\right\}dt+\sigma\left(X_t;\theta\right)dW_t^P. \end{aligned} \tag{5.16}$$

We adopt the simple specification for the market price of risk

$$\Lambda\left(X_t;\theta\right)=\sigma\left(X_t;\theta\right)'\lambda \tag{5.17}$$

with $\lambda$ an $m\times 1$ vector of constant parameters, so that under $P$, the instantaneous drift of each state variables is its drift under the risk-neutral measure, plus a constant times its volatility squared. Under this specification, the drift of the state vector is then affine under both the physical and risk-neutral measures, since

$$\mu^P\left(X_t;\theta\right)=\left(\tilde{A}+\widetilde{B}X_t\right)+\Sigma S\left(X_t;\beta\right)'\Sigma'\lambda\equiv A+BX_t. \tag{5.18}$$

An affine $\mu^P$ is not required for our likelihood expansions. Since we can derive likelihood expansions for arbitrary diffusions, $\mu^P$ may contain terms that are non-affine, such as the square root of linear functions of the state vector, as in Duarte (2004) for instance. Duffee (2002) and Cheridito et al. (2005) also allow for a more general market price of risk specifications than Dai and Singleton (2000), but retain the affinity of $\mu^Q$ and $\mu^P$ (and also of the diffusion matrix). However, we do rely on the affine character of the dynamics under $Q$ because those allow us to go from state to yields in the tractable manner given by (5.15).

# 6 Relaxing the Model Specification: Allowing for Nonparametric Dynamics

I now relax Assumption 4 to Assumption 9. I first describe specification tests that are designed to test any particular model, given by Assumption 8, against the nonparametric alternative Assumption 9. Then I discuss the non and semiparametric estimation of the functions $\mu(x)$ and $\sigma(x)$. Throughout this Section, Assumption 12 holds. Conditions on the function $\mu$ and $\sigma$ can be given to ensure that Assumption 12 is satisfied (see e.g. Aït-Sahalia (1996c)).

## 6.1 Specification Testing Against Nonparametric Alternatives

Consider a given parametrization for a diffusion given by Assumption 8, that is, a joint parametric family:

$$\mathcal{P}\equiv\left\{\left(\mu\left(\cdot,\theta\right),\sigma^2\left(\cdot,\theta\right)\right)\ \mid\theta\in\Theta\right\} \tag{6.1}$$

where $\Theta$ is a compact subset of $R^K$. If we believe that the true process is a diffusion with drift and diffusion functions $\left(\mu_0(\cdot)\ ,\sigma_0^2(\cdot)\right)$, a specification test asks whether there are values of the parameters in $\Theta$ for which the parametric model $\mathcal{P}$ is an acceptable representation of the true process, i.e., do the functions $\left(\mu_0(\cdot)\ ,\sigma_0^2\ (\cdot)\right)$ belong to the parametric family $\mathcal{P}$? Formally, the null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \ \exists\, \theta_0 \in \Theta \text{ such that.} \mu(\cdot,\theta_0) = \mu_0(\cdot) \text{ and } \sigma^2(\cdot,\theta_0) = \sigma_0^2(\cdot) \\ H_1 &: \ \left(\mu_0(\cdot), \sigma_0^2(\cdot)\right) \notin \mathcal{P} \end{aligned} \tag{6.2}$$

Because $\left(\mu_0(\cdot), \sigma_0^2(\cdot)\right)$ cannot easily be estimated directly from discretely sampled data, Aït-Sahalia (1996c) proposed to test the parametric specification using an indirect approach. Let $\pi(\cdot,\theta)$ denote the marginal density implied by the parametric model in Assumption 8, and $p(\Delta,\cdot|\cdot,\theta)$ the transition density. Under regularity assumptions, $\left(\mu(\cdot,\theta), \sigma^2(\cdot,\theta)\right)$ will uniquely characterize the marginal and transition densities over discrete time intervals. For example, the Ornstein-Uhlenbeck process $dX_t = \beta(\alpha - X_t)\,dt + \gamma dW_t$ specified by Vasicek (1977) generates Gaussian marginal and transitional densities. The square-root process $dX_t = \beta(\alpha - X_t)\,dt + \gamma X_t^{1/2} dW_t$ used by Cox et al. (1985) yields a Gamma marginal and non-central chi-squared transitional densities.

More generally, any parametrization $\mathcal{P}$ of $\mu$ and $\sigma^2$ corresponds to a parametrization of the marginal and transitional densities:

$$\left\{(\pi(\cdot,\theta), p(\Delta,\cdot,|\cdot,\theta)) \mid \left(\mu(\cdot,\theta),\, \sigma^2\ (\cdot,\theta)\right) \in \mathcal{P},\ \theta \in \Theta\right\} \tag{6.3}$$

While the direct estimation of $\mu$ and $\sigma^2$ with discrete data is problematic, the estimation of the densities explicitly take into account the discreteness of the data. The basic idea in Aït-Sahalia (1996c) is to use the mapping between the drift and diffusion on the one hand, and the marginal and transitional densities on the other, to test the model's specification using densities at the observed discrete frequency $(\pi, p)$ instead of the infinitesimal characteristics of the process $(\mu, \sigma^2)$:

$$\text{Drift and Diffusion} \quad \Longleftrightarrow \quad \text{Marginal and Transitional Densities} \tag{6.4}$$

Aït-Sahalia (1996c) proposed two tests, one based on the marginal density $\pi$, the other on the transition density $p$ and derived their asymptotic distributions under the assumption 12.

### 6.1.1 Testing the Marginal Specification

I start with the first specification test proposed for discretely sampled diffusions, the marginal density-based specification test in Aït-Sahalia (1996c). Let $\{\pi(\cdot,\theta) \mid \left(\mu(\cdot,\theta),\, \sigma^2\ (\cdot,\theta)\right) \in \mathcal{P},\ \theta \in \Theta\}$ denote the parametric family of marginal densities implied by the specification of the parametric model in Assumption 8. This family is characterized by the fact that the density $\pi(\cdot,\theta)$ corresponding to the pair $\left(\mu(\cdot,\theta), \sigma^2(\cdot,\theta)\right)$ is:

$$\pi(x,\theta) = \frac{\xi(\theta)}{\sigma^2(x,\theta)} \exp\left\{\int^x \frac{2\mu(u,\theta)}{\sigma^2(u,\theta)} du\right\} \tag{6.5}$$

where the choice of the lower bound of integration in the interior of the domain of the diffusion is irrelevant, and is absorbed in the normalization constant $\xi(\theta)$ determined to insure that the density integrates to one. If we let the true marginal density of the process be

$$\pi_0(x) = \frac{\xi_0}{\sigma_0^2(x)} \exp\left\{\int^x \frac{2\mu_0(u)}{\sigma_0^2(u)} du\right\} \tag{6.6}$$

we can then test

$$\begin{aligned} H_{M0} &: \quad \exists\, \theta_0 \in \Theta \text{ such that } \pi(\cdot, \theta_0) = \pi_0(\cdot) \\ H_{M1} &: \quad \pi_0(\cdot) \notin \Pi_M \end{aligned} \tag{6.7}$$

It is necessary that $H_{M0}$ be true for $H_0$ to be true. If the true density $\pi_0(\cdot)$ were known, we could simply check to see if it belonged to the proposed parametric class. Since it is unknown, we must estimate it, and do so with an estimator that does not already assume that the null hypothesis is correct (otherwise there is obviously no way of testing it). We use for that purpose a nonparametric estimator –that is, free of all parametric assumptions regarding $\mu$ and $\sigma^2$ – which will converge to the true density whether or not the parametric model in Assumption 8 is correctly specified. Now consider a parametric estimator of the implied density $\pi(\cdot, \theta_0)$. It will converge to the true density only if the model is correctly specified. Therefore the parametric and nonparametric density estimators should be close together if the parametric model is correct, and far from each other otherwise. A measure of distance $M$ between the two density estimates provides a natural statistic to test the null hypothesis of correct parametric specification. Aït-Sahalia (1996c) suggested to test $H_{M0}$ using the distance measure between the densities:

$$M \equiv \min_{\theta \in \Theta} \int_{\underline{x}}^{\bar{x}} (\pi(x, \theta) - \pi_0(x))^2 \pi_0(x) w(x) dx \tag{6.8}$$

where $w$ is a weight (or trimming) function, for instance $w(x) = 1\{|x| \leq c\}$ for some constant $c$ or $w(x) = 1$. The former is used to analyze the fixed-point trimming, and the second no trimming or the fixed-percentage trimming. When the state space has finite boundaries, trimming can be advisable.

The test statistic is based on

$$\hat{M} = nh \min_{\theta \in \Theta} \int_D (\pi(x, \theta) - \hat{\pi}(x))^2 \hat{\pi}(x) w(x) dx, \tag{6.9}$$

where $\hat{\pi}$ is the nonparametric density estimate of $\pi_0$, computed with bandwidth $h$. The null hypothesis is therefore rejected when the test statistic $\hat{M}$ is large enough. The test evaluates the distance between the densities at the "best possible" parametric estimator:

$$\hat{\theta}_M \equiv \arg\min_{\theta \in \Theta} \int_D (\pi(x, \theta) - \hat{\pi}(x))^2 \hat{\pi}(x) w(x) dx. \tag{6.10}$$

The test rejects the parametric specification $\pi(\cdot, \theta)$ if $\hat{M}$ takes a large value. The intuition for the test is straightforward. If $\pi(\cdot, \theta)$ is the correct specification, $\hat{M}$ should be small.

*Under stationarity*, $\hat{M}$ is shown in Aït-Sahalia (1996c) to be asymptotically Normally distributed and the critical value $c(\alpha)$ of the size $\alpha$ test is given by

$$\hat{c}(\alpha) = \hat{E}_M + z_{1-\alpha} h^{1/2} \hat{V}_M^{-1/2},$$

where $z_{1-\alpha}$ is the one-sided Normal cutoff for $\alpha$, and

$$\hat{E}_M = c_1 \int_{-\infty}^{\infty} \hat{\pi}^2(x) w(x)\, dx \quad \text{and} \quad \hat{V}_M = c_2 \int_{-\infty}^{\infty} \hat{\pi}^4(x) w(x)\, dx$$

for constants $c_1$ and $c_2$ dependent upon the kernel function only, and given by

$$\begin{aligned} c_1 &= \int_{-\infty}^{+\infty} K^2(x)\, dx \\ c_2 &= 2\int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} K(u)\, K(u+x)\, du \right\}^2 dx \end{aligned}$$

Following that work, Pritzker (1998) noted that, in small samples, the near-nonstationarity of empirical interest rate data can lead to over rejection of the null hypothesis when critical values are computed based on the asymptotic distribution derived under stationarity (see also Chapman and Pearson (2000), and Chen and Gao (2004b) for a critique of the critique). I will revisit this issue below in Section 7.

### 6.1.2 Testing the Specification of Transitions: An Indirect Approach

The stationary marginal density of the process, as just discussed, does not summarize all the information available in the data. Aït-Sahalia (1996c) shows how to exploit stationarity (in fact, just time-homogeneity) by combining the Kolmogorov forward and backward equations characterizing the transition function $p(y, t|x, s)$ in a way that eliminates the time derivatives of $p$ –which are unobservable with discrete data. Consider the forward Kolmogorov equation, with the natural initial condition ($x$ and $s$ are fixed):

$$\frac{\partial p(y,t|x,s)}{\partial t} = -\frac{\partial}{\partial y}\left(\mu(y) p(y,t|x,s)\right) + \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2(y) p(y,t|x,s)\right) \tag{6.11}$$

for all $y$ in the interior of the domain $\mathcal{S}$ of the diffusion and $t$ such that $t > s$. The backward equation ($y$ and $t$ are fixed) is:

$$-\frac{\partial p(y,t|x,s)}{\partial s} = \mu(x)\frac{\partial}{\partial x}\left(p(y,t|x,s)\right) + \frac{1}{2}\sigma^2(x)\frac{\partial^2}{\partial x^2}\left(p(y,t|x,s)\right) \tag{6.12}$$

for all $x$ in $\mathcal{S}$ and $s$ such that $0 \leq s < t$.

Unfortunately, these two equations cannot be used as such to estimate the parameters because their left-hand-side contains the derivative of the transition density with respect to time. Time derivatives cannot be estimated without observations on how the process changes over small intervals of time. But we can work around this problem by getting rid of the time derivatives as follows. Under time-homogeneity, $p(y,t|x,s) = p(y, t-s|x, 0) \equiv p(y|x, t-s)$ and therefore: $\partial p/\partial t = -\partial p/\partial s$. Combining the two equations (6.11)-(6.12) then yields the transition discrepancy:

$$\begin{aligned} K\left(y|x,\Delta\right) \equiv & \left\{ \frac{1}{2}\frac{\partial^2}{\partial y^2}\left(\sigma^2\left(y\right)p\left(y|x,\Delta\right)\right) - \frac{\partial}{\partial y}\left(\mu\left(y\right)p\left(y|x,\Delta\right)\right)\right\} \\ & - \left\{\mu\left(x\right)\frac{\partial}{\partial x}\left(p\left(y|x,\Delta\right)\right) + \frac{1}{2}\sigma^2\left(x,\theta\right)\frac{\partial^2}{\partial x^2}\left(p\left(y|x,\Delta\right)\right)\right\} \end{aligned} \tag{6.13}$$

where $\Delta = t - s$. For every $(x, y)$ in $\mathcal{S}^2$ and $\Delta > 0$, $K\left(y|x,\Delta\right)$ must be zero. Note that this must hold for every time interval $\Delta$, not just small ones.

If we now parametrize the diffusion process, then $K$ (with $\mu$ and $\sigma^2$ replaced by their assumed parametric form $\mu\left(\cdot,\theta\right)$ and $\sigma^2(\cdot,\theta)$ respectively) must be zero at the true parameter value under the null of correct parametric specification. Given nonparametric estimates of the transition function, $K = 0$ provides a testable implication. While the above discussion focuses on diffusions, the Kolmogorov equations have natural extensions for more general Markov processes (such as processes with jumps) and the corresponding transition discrepancy can be defined.

### 6.1.3 Testing the Specification of Transitions: A Direct Approach

In what follows, I discuss different approaches to understanding the behavior of transition functions. I consider specification tests for any given parametrization; tests for the Markov hypothesis; tests for the continuity of sample paths; all are based on nonparametric estimation of the transition functions.

Aït-Sahalia et al. (2005a) propose to develop an alternative specification test for the transition density of the process, based on a direct comparison of the nonparametric estimate of the transition function to the assumed parametric transition function instead of the indirect transition discrepancy criterion, that is in the form

$$H_0 : p(y|x,\Delta) = p(y|x,\Delta,\theta) \quad \text{vs.} \quad H_1 : p(y|x,\Delta) \neq p(y|x,\Delta,\theta). \tag{6.14}$$

instead of the indirect approach

$$H_0 : K(y|x,\Delta,\theta) = 0 \quad \text{vs.} \quad H_1 : K(y|x,\Delta,\theta) \neq 0 \tag{6.15}$$

What makes this new direct testing approach possible is the subsequent development described in Section 4 of closed form expansions for $p(y|x,\Delta,\theta)$, which, as noted above, is rarely known in closed form; this is required for (6.14), whereas (6.15) does not require knowledge of $p(y|x,\Delta,\theta)$. A complementary approach to this is the one proposed by Hong and Li (2005), who use the fact that under the null hypothesis, the random variables $\{P(X_{i\Delta}|X_{(i-1)\Delta},\Delta,\theta)\}$ are a sequence of i.i.d. uniform random variables; see also Thompson (2004), Chen and Gao (2004a) and Corradi and Swanson (2005). Using for $P$ the closed form approximations of Aït-Sahalia (2002b), they proposed to detect the departure from the null hypothesis by comparing the kernel-estimated bivariate density of $\{(Z_i, Z_{i+\Delta})\}$ with that of the uniform distribution on the unit square, where $Z_i = P(X_{i\Delta}|X_{(i-1)\Delta},\Delta,\theta)$. They need to deal with the boundary issues for the kernel density estimates on a finite region.

As in the parametric situation of (4.2), note that the logarithm of the likelihood function of the observed data $\{X_1, \cdots, X_{n+\Delta}\}$ is

$$\ell(p) = \sum_{i=1}^{n} \ln p(X_{(i+1)\Delta}|X_{i\Delta}, \Delta),$$

after ignoring the stationary density $\pi(X_1)$. A natural test statistic is to compare the likelihood ratio under the null hypothesis and the alternative hypothesis. This leads to the test statistic

$$T_0 = \sum_{i=1}^{n} \ln \left( \hat{p}(X_{(i+1)\Delta}|X_{i\Delta}, \Delta)/p(X_{(i+1)\Delta}|X_i, \Delta, \hat{\theta}) \right) w(X_{i\Delta}, X_{(i+1)\Delta}). \tag{6.16}$$

where $w$ is a weight function. Note that this is not the maximum likelihood ratio test, since the estimate under the alternative model is not derived from the maximum likelihood estimate.

Under the null hypothesis (6.14), the parametric and nonparametric estimators are approximately the same. Using for $p$ the locally linear estimator described below, the null distribution of $T_0$ is usually obtained by Taylor's expansion:

$$\begin{aligned} T_0 \approx & \sum_{i=1}^{n} \frac{\hat{p}(X_{(i+1)\Delta}|X_{i\Delta}, \Delta) - p(X_{i+\Delta}|X_{i\Delta}, \Delta, \hat{\theta})}{p(X_{(i+1)\Delta}|X_i, \Delta, \hat{\theta})} w(X_{i\Delta}, X_{(i+1)\Delta}) \\ & + \frac{1}{2} \sum_{i=1}^{n} \left\{ \frac{\hat{p}(X_{(i+1)\Delta}|X_{i\Delta}) - p(X_{(i+1)\Delta}|X_{i\Delta}, \Delta, \hat{\theta})}{p(X_{(i+1)\Delta}|X_{i\Delta}, \Delta, \hat{\theta})} \right\}^2 w(X_{i\Delta}, X_{(i+1)\Delta}). \end{aligned}$$

As the local linear estimator $\hat{p}$ is not an MLE, it is not clear whether the first term is asymptotically negligible. To avoid unnecessary technicality, we will consider the following $\chi^2$-test statistic

$$T_1 = \sum_{i=1}^{n} \left\{ \frac{\hat{p}(X_{(i+1)\Delta}|X_{i\Delta}, \Delta) - p(X_{(i+1)\Delta}|X_{i\Delta}, \Delta, \hat{\theta})}{p(X_{(i+1)\Delta}|X_{i\Delta}, \Delta, \hat{\theta})} \right\}^2 w(X_{i\Delta}, X_{(i+1)\Delta}). \tag{6.17}$$

A natural alternative method to $T_2$ is

$$T_2 = \sum_{i=1}^{n} \{\hat{p}(X_{(i+1)\Delta}|X_{i\Delta}, \Delta) - p(X_{(i+1)\Delta}|X_{i\Delta}, \Delta, \hat{\theta})\}^2 w(X_{i\Delta}, X_{(i+1)\Delta}). \tag{6.18}$$

The transition-density based test depends on two smoothing parameters $h_1$ and $h_2$. It is somewhat hard to implement. Like the Cramer-von Mises test, a viable alternative method is to compare the discrepancies between transition distributions. The leads to the test statistic

$$T_3 = \sum_{i=1}^{n} \{\hat{P}(X_{(i+1)\Delta}|X_{i\Delta}, \Delta) - P(X_{(i+1)\Delta}|X_{i\Delta}, \Delta, \hat{\theta})\}^2 w(X_{i\Delta}, X_{(i+1)\Delta}). \tag{6.19}$$

As we describe in the paper, the tests $T_1$ and $T_2$ are more powerful for detecting local departures from the null model, while $T_3$ is more powerful for detecting the global departures. The former involves an additional choice of bandwidth $h_2$. We derive the distribution of these statistics and investigate their power properties in departure directions that are relevant for empirical work in finance. The null distribution of $T_3$ can be better approximated by its asymptotic counterpart, as $T_3$ only localizes in the $x$-direction and hence uses many more local data points. This also makes its implementation more stable.

## 6.2 Nonparametric Estimation

### 6.2.1 Nonparametric Global Estimation

In the case where sampling occurs at a fixed $\Delta > 0$, i.e., under Assumption 5, nonparametric estimation options are limited. Under Assumption 12, Aït-Sahalia (1996b) shows that

$$\sigma^2(x) = \frac{2}{\pi(x)} \int^x \mu(u)\,\pi(u)\,du. \tag{6.20}$$

where $\pi$ is the stationary (marginal) density of the process. Suppose we parameterize $\mu$ to be affine in the state variable, $\mu(x) = \beta(\sigma - x)$. In that case,

$$E[X_\Delta | X_0] = \alpha + e^{-\beta\Delta}(X_0 - \alpha). \tag{6.21}$$

This conditional moment condition applies for any $\Delta > 0$. As a consequence, $(\alpha, \beta)$ can be recovered by estimating a first order scalar autoregression via least squares from data sampled at any interval $\Delta$. The implied drift estimator may be plugged into formula (6.20) to produce a semiparametric estimator of $\sigma^2(x)$. Since neither (6.21) nor (6.20) does not require that the time interval be small, this estimator of $\sigma^2(x)$ can be computed from data sampled at any time interval $\Delta$, not just small ones.

### 6.2.2 Nonparametric Local Estimation

Under Assumption 21, the infinitesimal drift and diffusion functions can be defined as the limits of the conditional mean and variance of the process:

$$\mu(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} \int_{|y - x^*| < \varepsilon} (y - x)\, p(y|x, \Delta) dy \tag{6.22}$$

$$\sigma^2(x) = \lim_{\Delta \to 0} \frac{1}{\Delta} \int_{|y - x^*| < \varepsilon} (y - x)^2 p(y|x, \Delta) dy. \tag{6.23}$$

These local characterizations suggest a regression-based estimation of $\mu$ and $\sigma$, since they can be rewritten as

$$\begin{aligned} E\left[\Delta^{-1/2}\xi_{i\Delta} | X_{(i-1)\Delta} = x\right] &= \mu(x) + o(\Delta) \\ E\left[\xi_{i\Delta}^2 | X_{(i-1)\Delta} = x\right] &= \sigma^2(x) + o(\Delta) \end{aligned}$$

where $\xi_{i\Delta} = \Delta^{-1/2}(X_{i\Delta} - X_{(i-1)\Delta})$ denotes the scaled increments of the process.

Kernel estimators of the type

$$\hat{\mu}(x) = \frac{\sum_{i=1}^n \Delta^{-1/2}\xi_{i\Delta} K_h\left(X_{(i-1)\Delta} - x\right)}{\sum_{i=1}^n K_h\left(X_{(i-1)\Delta} - x\right)} \text{ and } \hat{\sigma}^2(x) = \frac{\sum_{i=1}^n \xi_{i\Delta}^2 K_h\left(X_{(i-1)\Delta} - x\right)}{\sum_{i=1}^n K_h\left(X_{(i-1)\Delta} - x\right)}$$

where $K$ is the kernel function, $h$ the bandwidth and $K_h(z) = h^{-1}K(z/h)$, have been considered by Florens-Zmirou (1989), Stanton (1997), Jiang and Knight (1997), Bandi and Phillips (2003) and Kristensen (2004).

Instead of a kernel estimator (which is a locally constant estimator), Fan and Yao (1998) apply locally linear regression to this problem. This results in an estimator which has better boundary behavior than the kernel estimator.

These estimators are based on a form of state-domain smoothing, that is, they rely on using information on observations that are locally close together in space. But it is possible to use instead information on observations that are locally close together in time. And to combine both sources of information for potential efficiency gains.

### 6.2.3 Nonparametric Estimation of Transition Densities

To estimate the transition density $p(y|x,\Delta)$ nonparametrically using discretely sampled data on $X$ at time intervals $\Delta$, we can use a Nadaraya-Watson kernel estimator for the conditional density, or use a locally linear estimator (see Fan et al. (1996)). A kernel estimator is simply based on the fact that $p(y|x,\Delta)$ is the ratio of the join density of $(X_\Delta, X_0)$ to the marginal density of $X_0$, that is

$$\hat{p}(y|x,\Delta) = \frac{\sum_{i=1}^n K_h\left(X_{i\Delta}-y\right) K_h\left(X_{(i-1)\Delta}-x\right)}{\sum_{i=1}^n K_h\left(X_{(i-1)\Delta}-x\right)}. \tag{6.24}$$

Consider now two bandwidths $h_1$ and $h_2$ and two kernel functions $K$ and $W$. The locally linear estimator of $p$ is based on observing that as $h_2 \to 0$

$$E\left\{K_{h_2}(X_{(i+1)\Delta}-y)|X_{i\Delta}=x\right\} \approx p(y|x), \tag{6.25}$$

where $K_h(z) = K(z/h)/h$. The left-hand side of (6.25) is the regression function of the random variable $K_{h_2}(X_{(i+1)\Delta}-y)$ given $X_{i\Delta}=x$. Hence, the local linear fit can be used to estimate this regression function. For each given $x$, one minimizes

$$\sum_{i=1}^n \left\{K_{h_2}(X_{(i+1)\Delta}-y) - \alpha - \beta(X_{i\Delta}-x)\right\}^2 W_{h_1}(X_{i\Delta}-x) \tag{6.26}$$

with respect to the the local parameters $\alpha$ and $\beta$. The resulting estimate of the conditional density is simply $\hat{\alpha}$. The estimator can be explicitly expressed as

$$\hat{p}(y|x,\Delta) = \frac{1}{nh_1h_2}\sum_{i=1}^n W_n\left(\frac{X_{i\Delta}-x}{h_1};x\right) K\left(\frac{X_{(i+1)\Delta}-y}{h_2}\right), \tag{6.27}$$

where $W_n$ is the effective kernel induced by the local linear fit. Explicitly, it is given by

$$W_n(z;x) = W(z)\frac{s_{n,2}(x) - zs_{n,1}(x)}{s_{n,0}(x)s_{n,2}(x) - s_{n,1}(x)^2},$$

where

$$s_{n,j}(x) = \frac{1}{nh_1}\sum_{i=1}^n \left(\frac{X_{i\Delta}-x}{h_1}\right)^j W\left(\frac{X_{i\Delta}-x}{h_1}\right).$$

Note that the effective kernel $W_n$ depends on the sampling data points and the location $x$.

From (6.27), a possible estimate of the transition distribution $P(y|x,\Delta) = P(X_{i+\Delta} < y|X_i = x,\Delta)$ is given by

$$\hat{P}(y|x,\Delta) = \int_{-\infty}^{y} \hat{p}(z|x)dz = \frac{1}{nh_1}\sum_{i=1}^{n} W_n\left(\frac{X_{i\Delta}-x}{h_1}; x\right)\bar{K}\left(\frac{X_{(i+1)\Delta}-y}{h_2}\right),$$

where $\bar{K}(u) = \int_u^\infty K(z)dz$.

### 6.2.4 Nonparametric Option-Based Estimation of Transition Densities

A different approach is to estimate the transition density using discretely sampled observations on option prices instead of the prices of the underlying asset. A method to achieve this either nonparametrically or semiparametrically was proposed by Aït-Sahalia and Lo (1998).

Under standard assumptions the price of a derivative, written on the underlying asset with price process $X$, and with European payoff function $g(.)$ to be received $\Delta$ units of time from now is simply

$$H(x,\Delta) = e^{-r\Delta}\int_0^{+\infty} g\left(y\right)p^*(y|x,\Delta)dy \tag{6.28}$$

where $r$ is the riskless rate and $p^*(y|x,\Delta)$ denotes the transition density corresponding to the risk-neutral dynamics of $X$. If the observed dynamics of $X$ are given by Assumption 9, then the risk neutral ones correspond to replacing $\mu(x)$ by $(r-q)x$ where $q$ denotes the dividend rate paid by the asset ($r$ and $q$ are assumed constant for simplicity.)

If we specialize (6.28) to the payoff of a call option with strike price $K$, then $g_K(y) = \max(0, y-K)$ and it follows that

$$p^*(y|x,\Delta) = e^{r\Delta}\left[\frac{\partial^2 H_K(x,\Delta)}{\partial K^2}\right]_{K=y} \tag{6.29}$$

where $H_K$ is the price of a call option with strike $K$ (see Breeden and Litzenberger (1978).) One can then estimate by nonparametric regression the call pricing function $H_K$ and differentiate twice with respect to $K$. Aït-Sahalia and Lo (1998) discuss the twin curses of dimensionality and differentiation as they relate to this problem and suggest various solutions, including a semiparametric regression at the level of implied volatilities.

Note that the theory imposes the restriction that the price of a call option must be a decreasing and convex function of the option's strike price. Aït-Sahalia and Duarte (2003) show how to impose these types of shape restrictions on $p^*$ as a simple modification of nonparametric locally polynomial estimators. Their estimators satisfy in all samples these shapes restrictions.

Aït-Sahalia and Lo (2000) show how these types of transition density estimators can be used to infer the representative agent's preferences that are implicit in the market prices of options. If we let $p(y|x,\Delta)$ denote the transition density of $X$ under its observed dynamics, then in an exchange economy where $X$ is the only consumption good the representative agent maximizes the utility of consumption at date $t$, $U_t(X)$, then

$$p^*(y|x,\Delta) = c\,\zeta_\Delta\left(y|x\right)\; p(y|x,\Delta).$$

where $\zeta_{\Delta}(y|x) = U'_{\Delta}(y) / U'_{0}(x)$ is the agent's marginal rate of substitution and $c$ is a constant so that the lhs integrates to one. The coefficient of local risk aversion of the agent is then

$$\gamma(y|x,\Delta) \equiv -\frac{U''_{\Delta}(y)}{U'_{0}(x)} = -\frac{\partial \zeta_{\Delta}(y|x) / \partial y}{\zeta_{\Delta}(y|x)} = \frac{\partial p(y|x,\Delta)/\partial y}{p(y|x,\Delta)} - \frac{\partial p^{*}(y|x,\Delta)/\partial y}{p^{*}(y|x,\Delta)}$$

which can be estimated directly from estimates of the two transition densities $p$ and $p^{*}$.

# 7 Relaxing the Model Specification: Allowing for Nonstationary Dynamics

In this Section, I discuss inference and testing when Assumption 12 is not satisfied.

## 7.1 Likelihood Estimation Under Nonstationarity

There is an extensive literature applicable to discrete-time stationary Markov processes starting with the work of Billingsley (1961). The asymptotic covariance matrix for the ML estimator is the inverse of the score covariance or information matrix where the score at date $t$ is $\partial \ln p(X_{t+\Delta}|X_t, \Delta, \theta)/\partial\theta$ where $\ln p(\cdot|x, \Delta, \theta)$ is the logarithm of the conditional density over an interval of time $\Delta$ and a parameter value $\theta$.

When the underlying Markov process is nonstationary, the score process inherits this nonstationarity. The rate of convergence and the limiting distribution of the maximum likelihood estimator depends upon growth properties of the score process (*e.g.* see Hall and Heyde (1980) Chapter 6.2). A nondegenerate limiting distribution can be obtained when the score process behaves in a sufficiently regular fashion. The limiting distribution can be deduced by showing that general results pertaining to time series asymptotics (see e.g., Jeganathan (1995)) can be applied to the present context. One first establishes that the likelihood ratio has the locally asymptotically quadratic (LAQ) structure, then within that class separates between the locally asymptotically mixed Normal (LAMN), locally asymptotically Normal (LAN) and locally asymptotically Brownian functional (LABF) structures. As we have seen, when the data generating process is stationary and ergodic, the estimation is typically in the LAN class. The LAMN class can be used to justify many of the standard inference methods given the ability to estimate the covariance matrix pertinent for the conditional normal approximating distribution. Rules for inference are special for the LABF case. These structures are familiar from the linear time series literature on unit roots and co-integration. Details for the case of a nonlinear Markov process can be found in Aït-Sahalia (2002b).

## 7.2 Nonparametric Local Estimation Under Nonstationarity

The properties of the estimators described in Section 6.2.2 have been studied by Bandi and Phillips (2003); see Bandi and Phillips (2002) for a survey of these results. Asymptotic results for these local estimators can be obtained, in the absence of a stationary distribution for the process, under conditions such as recurrence of the process.

### 7.3 Specification Testing When the Underlying Process is Nonstationary

Recall now the specification test I described in Section 6 above for stationary processes, based on the marginal density. Aït-Sahalia and Park (2005) study the behavior of the marginal based test in the local-to-unity framework. The idea is to draw a parallel with what happens in the unit root literature in time series: when studying an AR(1) model with autoregressive parameter $\rho$, in the nearly-integrated situation where $\rho$ is close to one, it is often the case that the small sample distribution of the parameter estimate of $\rho$ is better approximated by assuming that the process has a unit root (the discrete equivalent to our Brownian motion) than by assuming that the process has $\rho$ close to but strictly smaller than one (in which case the process is stationary). We investigate the ability of that limiting distribution to better approximate the small behavior of the test when in fact the data generating process is stationary but exhibits very low speeds of mean reversion – as is the case for US interest rates, for instance. This will involve extensive use of the local time of the process.

In other words, if one fits an AR(1) process (i.e., Ornstein-Uhlenbeck) process to US short term rates, $\rho = 0.99$ at the daily frequency, 0.95 at the monthly frequency. The process is still formally stationary (because of strong mean reversion at the extremes), but the asymptotics based on assuming stationarity are slow to converge. And it may well be that the asymptotic behavior of the test derived under nonstationarity is a better approximation to the small sample behavior of the test, just like the Dickey-Fuller distribution is a better approximation to the properties of $\hat{\rho}$ when the true value of $\rho$ is 0.95, say.

As before, $X$ is observed at intervals of length $\Delta$ over time $[0, T]$. We let $T \to \infty$ and possibly $\Delta \to 0$. Suppose we apply the marginal density-based test above, which is designed for stationary processes. Nonstationarity can be viewed as a different form of misspecification (compared to misspecifying the form of $\mu$ and $\sigma$). For nonstationary diffusions, there is no stationary density. Therefore, any parametric family of densities $\pi$ cannot correctly specify the underlying density. And the test must be able to reject the null of correct specification.

What does the kernel density estimator $\hat{\pi}$ now estimate? It turns out that

$$\hat{\pi}(x) = \frac{L(T,x)}{T} + o_p(1)$$

where $L$ is the local time of the process (see Akonom (1993)). This also plays a crucial role in the estimation of nonstationary processes (see Bandi and Phillips (2003)). The local time $L$ of $X$ at $x$ is:

$$L(T,x) = \lim_{\varepsilon \to 0} \frac{1}{2\varepsilon} \int_0^T 1\{|X_t - x| < \varepsilon\}\, dt$$

Intuitively, $L(T,x)$ denote the time spent by $X$ in the neighborhood of point $x$ between time 0 and time $T$. The occupation times formula states that

$$\int_0^T f(X_t)dt = \int_{-\infty}^{\infty} f(x) L(T,x)\, dx$$

for any nonnegative function $f$ on $\mathbb{R}$, allows one to switch between time and space integrals.

Of critical importance here will be the behavior of the local time of the process asymptotically in $T$. We asume that, for the processes of interest, there exists $\kappa \in [0,1]$ such that for each $x$, $L(T,x) = O_p(T^{\kappa})$ as $T \to \infty$. A stationary process should be expected to spend more time near any given value $x$ than an explosive process, which barely visits each point. Indeed, $\kappa = 1$ for stationary (or positive recurrent) diffusions: as $T \to \infty$,

$$\frac{L(T,x)}{T} \to \pi(x) \quad \text{a.s.}$$

By contrast, $\kappa = 0$ for transient processes. We also give examples illustrating the intermediary range $0 < \kappa < 1$, including Brownian motion for which $\kappa = 1/2$ and

$$L(T,x) =_d T^{1/2} L(1, T^{-1/2}x)$$

where for fixed $x$, as $T \to \infty$, we have

$$L(1, T^{-1/2}x) \to L(1,0).$$

due to the continuity of $L(1,\cdot)$. The distribution of $L(1,0)$ or more generally $L(T,0)$ is given by

$$\Pr\left(L(T,0) \geq u\right) = \sqrt{\frac{2}{\pi T}} \int_{u}^{+\infty} \exp\left(-\frac{y^2}{2T}\right) dy \tag{7.1}$$

A heuristic argument that gives the result goes as follows. Since the test is an integral of a power of $\hat{\pi}$, and $\hat{\pi} = \frac{L}{T} + o_p(1)$ we will have to study integrals of powers of the local time $L$. We have $L(T,x) = T^{\kappa}\ell_T(x)$. But we must also have $\int_{-\infty}^{\infty} L(T,x)dx \equiv T$ since the total time spent between 0 and $T$ in the neighborhood of *all* points is $T$. Therefore $\int_{-\infty}^{\infty} \ell_T(x)dx = T^{1-\kappa}$where $\ell_T$ is neither exploding nor vanishing on its support. For these two things to happen together, we expect the support of $\ell_T$ to expand at rate $T^{1-\kappa}$ as $T \to \infty$. On its support, $\ell_T^q(x) \sim 1$ and we may therefore expect to have

$$\int_{-\infty}^{\infty} x^p L^q(T,x)dx \sim T^{q\kappa} \int_{-T^{1-\kappa}}^{T^{1-\kappa}} x^p \ell_T^q(x)dx \sim T^{(p+1)(1-\kappa)+q\kappa}$$

and for a bounded function $b$ :

$$\int_{-\infty}^{\infty} b(x) L^q(T,x)dx \sim T^{q\kappa} \int_{-T^{1-\kappa}}^{T^{1-\kappa}} b(x) \ell_T^q(x)dx \sim T^{q\kappa}.$$

The first set of results concern the consistency of the test. So let us assume that the model is misspecified even if the diffusion is stationary. Aït-Sahalia and Park (2005) show that the test is consistent if $\Delta = o(T^{(9\kappa-5)/4-\delta})$ for some $\delta > 0$. So the test is consistent if $\Delta$ is sufficiently small relative to $T$. Intuitively, as $\Delta$ decreases we collect more information and as $T$ increases the underlying diffusions exhibit more nonstationary characteristics. For stationary diffusions, the result implies that the misspecified models are rejected asymptotically with probability one, under no conditions other than $T \to \infty$. Indeed, in this case we have $\kappa = 1$, and the condition requires that $\Delta = o(T^{1-\delta})$ for some $\delta > 0$. But as $\kappa$ decreases down from 1, we need more stringent conditions for the test to be consistent. For Brownian motion, we have $\kappa = 1/2$. The required condition for the test consistency becomes $\Delta = o(T^{-1/8-\delta})$. Now, in most practical applications, $\Delta$

is much smaller than the reciprocal of any fractional power of the time span $T$. If we use daily observations, then $\Delta \approx 1/252 = 0.004$. Even for $T = 100$ years, we have $T^{-1/8} \approx 0.5623$, which is much bigger. So for the usual values of $T$ and $\Delta$, the test is likely to reject.

We can use this finding to predict the performance of the test for stationary diffusions which are nearly nonstationary. For instance, consider a highly persistent Ornstein-Uhlenbeck process (i.e., with mean-reversion parameter positive but close to 0). In that situation, the test is expected to over-reject the null in small samples against any parametric specification, and this is what is found empirically by Pritzker (1998).

Next, we derive the asymptotic distribution of the test statistic $\hat{M}$ when the data generating process is nearly Brownian motion, and later investigate the ability of that limiting distribution to approximate the small behavior of the test when in fact the data generating process is stationary but exhibits very low speeds of mean reversion – as is the case for US interest rates, for instance. We let $X$ be generated as

$$dX_t = -\frac{\delta}{T} X_t dt + \sigma dW_t, \tag{7.2}$$

i.e., an Ornstein-Uhlenbeck process with the mean reversion parameter $\beta = \delta/T$ for some $\delta \geq 0$ and $X_0 = 0$. Clearly, $X$ reduces to a Brownian motion, if $\delta = 0$. We have that

$$\begin{aligned} X_t &= \sigma \int_0^t \exp\left[-(\delta/T)(t-s)\right] dW_s \\ &=_d \sigma\sqrt{T} \int_0^{t/T} \exp\left[-\delta(t/T - s)\right] dW_s \\ &= \sigma\sqrt{T} V^\delta_{t/T}, \end{aligned} \tag{7.3}$$

where $V^\delta$ is the Ornstein-Uhlenbeck process with the mean reversion parameter $\delta > 0$, unit variance and $V^\delta_0 = 0$ a.s. The result in (7.3) follows in particular from the scale invariant property of the Brownian motion $W$, i.e., $W_{Ts} =_d \sqrt{T} W_s$.

For the nearly Brownian motion in (7.2), we may obtain the limit distribution of the test statistic $\hat{M}$ more explicitly. We let $h \to 0$, $T \to \infty$ and $\Delta/h^2 \to 0$. Aït-Sahalia and Park (2005) show that for the near Brownian motion in (7.2), we have

$$\frac{\Delta}{h\sqrt{T}} \hat{M} \to_d \frac{L_\delta(1,0)}{\sigma} \min_{\theta \in \Theta} \int_{-\infty}^{\infty} \pi^2(x,\theta) w(x)\, dx, \tag{7.4}$$

where $L_\delta$ is the local time of the Ornstein-Uhlenbeck process $V^\delta$. The distribution of $L_0(1,0)$ is given in (7.1). Of course, the constant term in the limiting distribution of $\hat{M}$ can be computed once the parametric family $\pi(\cdot,\theta)$ is specified. For Brownian motion, the variance of the marginal distribution increases and explodes without a bound as $T \to \infty$. Therefore, the density that most closely approximates the (non-existing) limiting marginal density of Brownian motion is naturally given by the most diffuse distribution in the family, i.e., the distribution with the largest variance in case of the normal family. The mean becomes unimportant in this case. Given this observation, it is intuitive that the distribution of $\hat{M}$ involves the maximal variance and no mean parameter.

# 8 Relaxing the Model Specification: Allowing for Discontinuous Dynamics

The fact that jumps play an important role in many variables in economics and finance, such as asset returns, interest rates or currencies, as well as a sense of diminishing marginal returns in studies of the "simple" diffusive case, has led to a flurry of recent activity dealing with jump processes. I now examine some of these questions, thereby relaxing Assumption 2 to allow for jumps. Assumptions 1 and 3 are maintained.

## 8.1 Testing for Jumps Using the Transition Density

In Aït-Sahalia (1996a) and Aït-Sahalia (2002c), I investigated whether discretely sampled financial data can help us decide which continuous-time models are sensible. Diffusion processes are characterized by the continuity of their sample paths. This cannot be verified from the discrete sample path: even if the underlying path were continuous, data sampled at discrete times will always appear as a succession of jumps. Instead, I relied on a necessary and sufficient characterization of the transition density to determine whether the discontinuities observed are the result of the discreteness of sampling, or rather evidence of genuine jump dynamics for the underlying continuous-time process.

So, what can be said if the process is only observed at a finite observation interval $\Delta$? Consider a family of probability distributions for the Markov process $X$ with state space $\mathbb{R}$, and indexed by the time interval $\Delta$: $P(\cdot|x,\Delta)$. Could this family of densities have come from a scalar diffusion process, i.e., a process with continuous sample paths, or must jumps be included? As discussed in these two papers, this question can be addressed in light of the total positivity characterization of Karlin and McGregor (1959b). The relevant result here is the fact that $P_\theta$ represents the family of transition densities for a (univariate) diffusion if and only if the transition function of any diffusion process must obey the *total positivity* inequality. While total positivity has a more general representation and probabilistic interpretation, it implies

$$P\left(B|x,\Delta\right)P\left(\tilde{B}|\tilde{x},\Delta\right)-P\left(B|\tilde{x},\Delta\right)P\left(\tilde{B}|x,\Delta\right)>0 \tag{8.1}$$

whenever, $x<\tilde{x}$ and $B<\tilde{B}$ (where $B<\tilde{B}$ is interpreted to mean that every element of $B$ is less than every element of $\tilde{B}$). Since this must hold for any choice of $\tilde{x}$ and $\tilde{B}$, there is a local (in the state) counterpart that we express using the logarithm of the density:

$$\frac{\partial^2}{\partial x\partial y}\ln p(y|x,\Delta)>0 \tag{8.2}$$

for all $x$ and $y$ and interval $\Delta$. This cross derivative restriction for each choice of $x$, $y$ and $\Delta$ is a necessary condition for transition distributions to be those implied by a scalar diffusion. A partial converse is also available. Suppose that the family of distribution functions of a Markov process on $\mathbb{R}$ satisfies (8.1) for any positive $\Delta$. Then under a side condition, there exists a realization of the process such that almost all sample paths are continuous. Aït-Sahalia (2002c) shows how to build and justify formally a statistical test, in the parametric case, of this cross-derivative restriction for data sampled at a given sampling interval $\Delta$.

The following example shows how criterion (8.2) can be used to eliminate some transition densities as coming from a model of a scalar diffusion. Suppose that $p(y|x,\Delta)$ depends on the state $(y,x)$ only through $y-x$. Using the criterion (8.2), it can be shown that the only admissible solutions is

$$p\left(y|x,\Delta\right) = (2\pi\beta^2\Delta)^{-1/2}\exp\left\{-\frac{(y-x-\alpha\Delta)^2}{2\beta^2\Delta}\right\} \tag{8.3}$$

where $\theta = (\alpha,\beta)$, that is the transition density of an arithmetic Brownian motion. Consider the generalized Cauchy density

$$\ln p\left(y|x,\Delta\right) = -\ln\pi + \ln a(\Delta) - \ln\left[a(\Delta)^2 + (y-x)^2\right]$$

where $a(\Delta)$ is positive. The criterion (8.2) will fail for large $y-x$. Aït-Sahalia (2002c) contains other examples.

More generally, total positivity implies restrictions on processes defined on state spaces other than $\mathbb{R}$. Consider a continuous-time, stationary, Markov chain that can only take countable discrete values, say, $\{\ldots,-1,0,1,\ldots\}$. When does such a process have continuous sample paths? Obviously, the notion of continuity of a sample path depends on the state space: in $\mathbb{R}$, this is the usual definition of a continuous function. More generally, by continuity one means continuity with respect to the order topology of the state space of the process. In a discrete state space, the appropriate notion of continuity of the chain's sample paths is the following intuitive one: it constrains the chain to never jump by more than one state at a time, either up or down. It turns out that the restriction on the chain's transition probabilities analogous to (8.1) characterizes precisely this form of continuity: total positivity across all intervals restricts the process to be a so called birth-and-death process (see Karlin and McGregor (1959a)). In this sense, a birth-and-death process is the discrete-state analog to a scalar diffusion. See Aït-Sahalia (2002c) for further discussion and implications for derivative pricing methods, such as binomial trees.

For a fixed $\Delta$, total positivity is a necessary restriction on the transition distribution but not a sufficient one. Given a candidate transition distribution over an interval $\Delta$, we did not construct a diffusion with that transition density. Frydman and Singer (1979) study the analogous question for a finite state birth and death process. In their study they show that to embed a single transition matrix (over an interval $\Delta$) satisfying total positivity in a continuous-time Markov process it is sometimes necessary that the continuous-time process be time-inhomogeneous. They show that the total positivity function is a weaker restriction than embeddability for a continuous-time process that is restricted to be time-homogeneous.

But what if we are not willing to take a stand on a parametric specification for the transition density $p$? Aït-Sahalia and Fan (2005a) propose to design and implement a test for restriction (8.2) based on nonparametric estimators of $p$. There are several methods to testing the constraints (8.2). They can be classified as local and global approaches. The local test is to examine whether constraints (8.2) are violated at various places, while global approaches consist in testing whether the minimum of (8.2) for $(x,y)$ over a compact set, which contains for instance 80% of data, is positive. In both cases, we face the challenge of the estimation of the second partial derivatives. With such estimates, the nonparametric test statistics can easily be formulated.

There are three possible approaches to estimating $\partial^2 \ln p(y|x,\Delta)/\partial x\partial y$. The naive one is the substitution estimator, using one of the estimators discussed in Section 6.2.3. The second approach is to extend the local likelihood estimation of the density function (see Hjort and Jones (1996) and Loader (1996)) to the estimation of the conditional density. Combining this with the local linear modeling of Fan (1992), a direct estimator of $\partial^2 \ln p(y|x,\Delta)/\partial x \partial y$ can be obtained. The third approach is to extend the 'parametric-start' idea of Hjort and Glad (1995), Efron and Tibshirani (1996) and Glad (1998) to estimating nonparametrically the transition density. Take a parametric family of transition density such as those in Aït-Sahalia (2002c) as a starting point, correct the possible biases in the parametric estimation by using a nonparametric estimation to the difference. This kind of idea originates from the prewhitening technique of Press and Turkey (1956) on the estimation of spectral density. The advantage is that it allows us to incorporate the prior knowledge on the shape of the transition density. When the bandwidths of the nonparametric estimates are very large, the nonparametric estimates reduce to parametric estimates. This provides a smooth family of estimates indexed by the bandwidths, starting from parametric estimates of the transition density to full nonparametric estimate of transition density as bandwidths vary. This is particularly useful for our problem, as we need to estimate the difficult quantity $\partial^2 \ln p(y|x,\Delta)/\partial x \partial y$. On one hand, nonparametric estimates give a very slow rate of convergence while on the other the parametric estimates enjoy the usual root-$n$ consistency. Our approach bridges the gaps between these two important approaches, allowing us to incorporate prior knowledge and making practical implementation feasible.

Other approaches to testing for jumps are based on the behavior of short dated options (Carr and Wu (2003)) or the different limiting behavior of the quadratic variation and related quantities, in the presence of jumps (see Barndorff-Nielsen and Shephard (2003), Andersen et al. (2003) and Huang and Tauchen (2006).)

## 8.2 Estimating the Volatility Parameter in the Presence of Jumps

A different issue related to the presence of jumps is whether they have an impact on our ability to estimate the other parameters of the process. In Aït-Sahalia (2004), I asked whether the presence of jumps impact our ability to estimate the diffusion parameter $\sigma^2$. Despite intuition that seems to suggest that the identification of $\sigma^2$ is hampered by the presence of the jumps, I showed that maximum-likelihood can actually *perfectly* disentangle Brownian noise from jumps provided one samples frequently enough. For instance, suppose that, instead of Assumption 4 we are under Assumption 10. For simplicity, let us start by further assuming that $(\mu, \sigma, \lambda)$ are constant and that $J_t$ is normally distributed in

$$dX_t = \mu dt + \sigma dW_t + J_t dN_t. \tag{8.4}$$

I first show this result in the context of a compound Poisson process, i.e., the jump-diffusion model in (8.4).

The first result there showed that it is still possible, using maximum likelihood, to identify $\sigma^2$ with the same degree of precision as if there were no jumps, namely that when the Brownian motion is contaminated by compound Poisson jumps, it remains the case that

$$\text{AVAR}_{\text{MLE}}\left(\sigma^2\right) = 2\sigma^4 \Delta + o(\Delta) \tag{8.5}$$

so that in the limit where sampling occurs infinitely often ($\Delta \rightarrow 0$), the MLE estimator of $\sigma^2$ has the same asymptotic distribution as if no jumps were present. These arguments are asymptotic in small $\Delta$, that is, take the form of a Taylor expansion in $\Delta$ around $\Delta = 0$.

Note also that this result states that the presence of the jumps imposes no cost on our ability to estimate $\sigma^2$ : the variance which is squared in the leading term is only the diffusive variance $\sigma^2$, not the total variance $\sigma^2 + (\beta^2 + \eta)\lambda$. This can be contrasted with what would happen if, say, we contaminated the Brownian motion with another independent Brownian motion with known variance $v$. In that case, we could also estimate $\sigma^2$, but the asymptotic variance of the MLE would be $2\left(\sigma^2 + v\right)^2 \Delta$.

What is happening here is that, as $\Delta$ gets smaller, our ability to identify price discontinuities improves. This is because these Poisson discontinuities are, by construction, discrete, and there are few of them relative to the diffusive moves. Then if we can see them, we can exclude them, and do as if they did not happen in the first place. More challenging therefore will be the case where the jumps are both infinitely frequent and infinitely small.

One may indeed wonder whether this result is driven by the fact that Poisson jumps share the dual characteristic of being large and infrequent. Is it possible to perturb the Brownian noise by a Lévy pure jump process other than Poisson, and still recover the parameter $\sigma^2$ as if no jumps were present? The reason one might expect this not to be possible is the fact that, among Lévy pure jump processes, the Poisson process is the only one with a finite number of jumps in a finite time interval. All other pure jump processes exhibit an *infinite number of small jumps* in any finite time interval. Intuitively, these tiny jumps ought to be harder to distinguish from Brownian noise, which is itself made up of many small moves. Perhaps more surprisingly then, I will show that maximum likelihood can still perfectly discriminate between Brownian noise and a Cauchy process, a canonical example of such processes.

I then examine whether the perfect distinction afforded by MLE is specific to the fact that the jump process considered so far was a compound Poisson process, or whether it extends to other types of jump processes. Among the class of continuous-time Markov processes, it is natural to look at Lévy processes. Poisson jumps are a unique case in the Lévy universe. Yet, it is possible to find examples of other pure jump processes for which the same result continues to hold, which cannot be explained away as easily as in the Poisson case.

A Lévy process can be decomposed as the sum of three independent components: a linear drift, a Brownian motion and a pure jump process. Correspondingly, the log-characteristic function of a sum of independent random variables being the sum of their individual characteristic functions, the characteristic function of a Lévy process given in Assumption 11 is given by the Lévy-Khintchine formula, which states that there exist constants $b \in \mathbb{R}$, $c \geq 0$ and a positive sigma-finite measure $\nu(\cdot)$ on $\mathbb{R}\backslash\{0\}$ (extended to $\mathbb{R}$ by setting $v(\{0\}) = 0$) satisfying $\int_{-\infty}^{+\infty} \min\left(1, z^2\right) \nu(dz) < \infty$ such that the log-characteristic function $\psi(u)$ has the form for $u \in \mathbb{R}$ :

$$\psi(u) = ibu - \frac{c^2}{2}u^2 + \int_{-\infty}^{+\infty} \left(e^{iuz} - 1 - iuzh(z)\right) \nu(dz). \tag{8.6}$$

The three quantities $(\gamma_c, \sigma, \nu(\cdot))$, called the characteristics of the Lévy process, completely describe the probabilistic behavior of the process. $\gamma_c$ is the drift rate of the process, $\sigma$ its volatility from the Brownian component and the measure $\nu(\cdot)$ describes the pure jump component. It is known as the Lévy measure and has the interpretation that $\nu(E)$ for any subset $E \subset \mathbb{R}$ is the rate at which the process takes jumps of size $x \in E$, i.e., the number of jumps of size falling in $E$ per unit of time. Sample paths of the process are continuous if and only if $\nu \equiv 0$. Note that $\nu(\cdot)$ is not necessarily a probability measure, in that $\nu(\mathbb{R})$ may be finite or infinite. The function $h(z)$ is a weighting function whose role is to make the integrand in (8.6) integrable.

Examples of Lévy processes include the Brownian motion ($h = 0$, $b = 0$, $c = 1$, $\nu = 0$), the Poisson process ($h = 0$, $b = 0$, $c = 0$, $\nu(dx) = \lambda\delta_1(dx)$ where $\delta_1$ is a Dirac point mass at $x = 1$) and the Poisson jump diffusion I considered above in (8.4), corresponding to $h = 0$, $b = \mu$, $c > 0$, $\nu(dx) = \lambda n(x; \beta, \eta)dx$ where $n(x; \beta, \eta)$ is the Normal density with mean $\beta$ and variance $\eta$.

The question I will address is whether it is possible to perturb the Brownian noise by a Lévy pure jump process other than Poisson, and still recover the parameter $\sigma^2$ as if no jumps were present. The reason one might expect this not to be possible is the fact that, among Lévy pure jump processes, the Poisson process is the only one with a finite $\nu(\mathbb{R})$, i.e., a finite number of jumps in a finite time interval (and the sample paths are piecewise constant). In that case, define $\lambda = \nu(\mathbb{R})$ and the distribution of the jumps has measure $n(dx) = v(dx)/\lambda$. All other pure jump processes are such that $\nu([-\varepsilon, +\varepsilon]) = \infty$ for any $\varepsilon > 0$, so that the process exhibits an infinite number of small jumps in any finite time interval. Intuitively, these tiny jumps ought to be harder to distinguish from Brownian noise, which is itself made up of many small moves. Can the likelihood still tell them perfectly apart from Brownian noise?

I considered in Aït-Sahalia (2004) as an example the Cauchy process, which is the pure jump process ($c = 0$) with Lévy measure $\nu(dx) = \alpha dx/x^2$ and, with weight function $h(z) = 1/(1 + z^2)$, $\gamma_c = 0$. This is an example of a symmetric stable distribution of index $0 < \xi < 2$ and rate $\alpha > 0$, with log characteristic function proportional to $\psi(u) = -(\alpha\,|u|)^\xi$, and Lévy measure $\nu(dx) = \alpha^\xi \xi dx/\,|x|^{1+\xi}$. The Cauchy process corresponds to $\xi = 1$, while the limit $\xi \to 2$ (from below) produces a Gaussian distribution.

So I next look at the situation where $dX_t = \mu dt + \sigma dW_t + dC_t$, where $C_t$ is a Cauchy process independent of the Brownian motion $W_t$. The answer is, surprisingly, yes: when the Brownian motion is contaminated by Cauchy jumps, it still remains the case that

$$\text{AVAR}_{MLE}\left(\sigma^2\right) = 2\sigma^4\Delta + o(\Delta). \tag{8.7}$$

Intuitively, while there is an infinite number of small jumps in a Cauchy process, this "infinity" remains relatively small (just like the cardinal of the set of integers is smaller than the cardinal of the set of reals) and while the jumps are infinitesimally small, they remain relatively bigger than the increments of a Brownian motion during the same time interval $\Delta$. In other words, they are harder to pick up from inspection of the sample path than Poisson jumps are, but with a fine enough microscope, still possible. And the likelihood is the best microscope there is, in light of the Cramer-Rao lower bound.

## 8.3 Jumps vs. Volatility: Questions of Uniformity

Whether these results continue to hold for general Lévy jump processes is investigated in Aït-Sahalia and Jacod (2004) and Aït-Sahalia and Jacod (2006). The idea is to characterize precisely the class of jump processes which can still perfectly be distinguished from Brownian volatility. Let the Fisher information at stage $n$ for

$\sigma^2$ be $nI(\sigma^2, \Delta, G)$, where $G$ denotes the law of the pure jump process.

If we are interested in $\sigma^2$ only, it is natural to consider the law of the jump process, that is $G$, as a nuisance parameter. Hence the idea of proving a convergence like (8.5) and (8.7) –for AVAR$_{MLE}$ which is $\Delta$ times the inverse of $I$, or equivalently for $I$ itself– which is uniform in $G$. Here, $G$ is arbitrary in the set $\mathcal{G}$ of all infinitely divisible law with vanishing Gaussian part. The closure of $\mathcal{G}$ (for the weak convergence) contains all Gaussian laws: so if the convergence were uniform in $G \in \mathcal{G}$ it would hold as well when the Lévy process is also a Wiener process with variance, say, $v$: then the best one can do is to estimate $\sigma^2 + v$, and as noted above one cannot have even consistent estimators for $\sigma^2$ when $v$ is altogether unknown.

So the idea is to restrict the set $\mathcal{G}$ to a subset which lies at a positive distance of all Gaussian laws. For this, we recall that $G \in \mathcal{G}$ is characterized by its drift $b \in I\!R$ and its Lévy measure $F$, through the Lévy–Khintchine representation of infinitely divisible distributions, given in (8.6). For any constant $K$ and index $\alpha \in [0, 2]$ we denote by $\mathcal{G}(K, \alpha)$ the family of all infinitely divisible laws of the form (8.6) with

$$|b| \leq K, \qquad F([-x, x]^c) \leq K\left(1 \vee \frac{1}{x^\alpha}\right) \quad \forall\, x > 0. \tag{8.8}$$

A stable law of index $\alpha < 2$, which has $F(dx)$ is proportional to $|x|^{-\alpha-1}\, dx$, belongs to $\mathcal{G}(K, \alpha)$ for some $K$. Any infinitely divisible law without Gaussian part belongs to $\mathcal{G}(K, 2)$ for some $K$ . If $G \in \mathcal{G}(K, 0)$ then $Y$ is a compound Poisson process plus a drift. The closure of $\mathcal{G}(K, 2)$ contains Gaussian laws, but if $\alpha < 2$ the set $\mathcal{G}(K, \alpha)$ is closed and does not contain any non–trivial Gaussian law.

We then prove results of the following type, giving both the uniformity of the convergence on the set $\mathcal{G}(K, \alpha)$, and the lack of uniformity otherwise: For all $K > 0$ and $\alpha \in [0, 2)$ and $\sigma^2 > 0$ we have

$$\sup_{G \in \mathcal{G}(K,\alpha)} \left(\frac{1}{2\sigma^4} - I(\sigma^2, \Delta, G)\right) \to 0 \qquad \text{as} \quad \Delta \to 0. \tag{8.9}$$

For each $n$ let $G^n$ be the symmetric stable law with index $\alpha_n \in (0, 2)$ and scale parameter $v/2$ (i.e., its characteristic function id $u \mapsto \exp\left(-\frac{v}{2}|u|^{\alpha_n}\right)$. Then if $\alpha_n \to 2$, for all sequences $\Delta_n \to 0$ satisfying $(2 - \alpha_n) \log \Delta_n \to 0$ we have

$$I(\sigma^2, \Delta_n, G^n) \to \frac{1}{2(\sigma^2 + v)^2}. \tag{8.10}$$

## 8.4 GMM Estimation in the Presence of Jumps

Can the identification of $\sigma^2$ achieved by the likelihood, despite the presence of jumps, be reproduced by conditional moments of the process of integer or non-integer type, and which moments or combinations of moments come closest to achieving maximum likelihood efficiency .While it is clear that MLE is the preferred method, and as discussed above has been used extensively in that context, it is nevertheless instructive to determine which specific choices of moment functions do best in terms of approximating its efficiency.

In Aït-Sahalia (2004), I studied GMM moment conditions to estimate the parameters in the presence of jumps, with the objective of studying their ability to reproduce the efficiency of MLE. I consider in particular absolute moments of order $r$ (i.e., the plims of the power variations). To form unbiased moment conditions, I need an exact expression for these moments, $M_a(\delta, \theta, r)$, which I derive in closed form. I form moment functions of the type $h(y, \delta, \theta) = y^r - M(\delta, \theta, r)$ and/or $h(y, \delta, \theta) = |y|^r - M_a(\delta, \theta, r)$ for various values of $r$. By construction, these moment functions are unbiased and all the GMM estimators considered will be consistent. The question becomes one of comparing their asymptotic variances among themselves, and to that of MLE. I refer to different GMM estimators of $\theta$ by listing the moments $M(\Delta, \theta, r)$ and/or $M_a(\Delta, \theta, r)$ that are used for that particular estimator. For example, the estimator of $\sigma^2$ obtained by using the single moment $M(\Delta, \theta, 2)$ corresponds to the discrete approximation to the quadratic variation of the process. Estimators based on the single moment $M_a(\delta, \theta, r)$ correspond to the power variation, etc. By using Taylor expansions in $\Delta$, I characterize in closed form the properties of these different GMM estimators. The end result is a direct comparison of the different types of moment conditions, and the selection of optimal combinations of moment conditions for the purpose of estimating the parameters of a jump-diffusion.

Aït-Sahalia and Jacod (2006) consider GMM-type estimators for the model

$$dX_t = \sigma dW_t + dY_t,$$

where $W$ is a standard Brownian motion or, more generally, a symmetric stable process of index $\beta$ and the process $Y$ is another Lévy process without Brownian (or continuous) part and with jumps dominated by those of $W$. The aim is to construct estimators for $\sigma$ which behave under the model $X_t = \sigma W_t + Y_t$ as well as under the model $X_t = \sigma W_t$ asymptotically as $\Delta_n \to 0$ and $n \to \infty$. In some applications, the jump perturbation $Y$ may represent frictions that are due to the mechanics of the trading process. Or in the case of compound Poisson jumps it may represent the infrequent arrival of relevant information related to the asset, in which case the law of $Y$ may be difficult to pin down due to the peso problem.

Let us distinguish between a parametric case, where the law of $Y$ is known, and a semiparametric case, where it is not. In the parametric case, we construct estimators which are asymptotically efficient. In the semiparametric case, obtaining asymptotically efficient estimators requires $\Delta_n$ to go fast enough to 0. We can then construct estimators that are efficient uniformly when the law of $Y$ stays in a set sufficiently separated from the law of $W$. The estimators are based on a variety of moment conditions, such as the empirical characteristic function or power and truncated power functions.

## 8.5 Option-Based Transition Densities in the Presence of Jumps

Recall the discussion of option-based transition densities in Section 6.2.4. Aït-Sahalia et al. (2001) infer information from different estimates of $p^*$. The estimator of $p^*$ constructed from (6.29) is a cross-sectional estimator because it uses information on a set of option prices with different strikes at one point in time. But recalling that under Assumption 9, $p^*$ corresponds to the dynamics of $X$ with drift set to $(r-q)x$ and local volatility function $\sigma(x)$ unchanged, it is possible to construct a different estimate of $p^*$ that uses time series observations on $X$. Since $\sigma(x)$ is the same under both the risk neutral and actual probability measures, the function $\sigma(x)$ can be estimated from the observed dynamics of $X$. Then, since $r-q$ is observable from the price of a forward contract written on $X$, an estimate of $p^*$ can be computed to be the transition function corresponding to the SDE with (now known) drift $(r-q)x$ and volatility $\sigma(x)$ : one could use the method described in Section 4.1 to construct this estimator in closed form. One can then test the overidentifying restriction that the cross-sectional and time-series $p^*$ are identical.

Empirical results based on S&P 500 options suggest that a peso problem is at play: the cross-sectional $p^*$ prices options as if $X$ were susceptible to large (downward) jumps, even though those jumps are generally absent from the time series data. Consequently, the time series $p^*$ will not show any evidence of jumps whereas the cross-sectional one will. When the actual dynamics of $X$ are given by a nonparametric version of Assumption 10, the risk-neutral dynamics of $X$ become

$$dX_t = (r - q - \lambda^* \kappa^*)X_t dt + \sigma\left(X_t\right) dW_t + J_t X_t dN_t$$

where $\kappa^*$ is the risk neutral expected value of the jump size $J_t$ and $\lambda^*$ is the risk neutral intensity of the Poisson process $N_t$. Under a peso problem situation, the occurrence of jumps in the time-series of actual observations on $X$ is infrequent, so we can use the same estimator of $\sigma\left(\cdot\right)$ as if no jumps had been observed during the period of interest. However, when we simulate the risk-neutral dynamics, we can draw from a process that incorporates the jump term. Now, $\lambda^*$ and $\kappa^*$ are determined by investors' preferences, so without assumptions on preferences, the equality between the cross-sectional and time series $p^*$ is no longer an over-identifying restriction. Instead, it allows us to restore the exact identification of the system and we can infer the arrival rate of jumps required to make the two $p^*$ equal.

## 8.6 Likelihood Estimation for Jump-Diffusions

It is possible to extend the basic closed form likelihood expansion described in Section 4 for diffusions to cover the situation where Assumptions 4 and 8 are generalized to processes driven by a Brownian motion and a compound Poisson process, i.e., Assumption 10.

The expression, due to Yu (2003), is of the form:

$$\begin{aligned} p_X^{(K)}\left(\Delta, x|x_0;\theta\right) &= \exp\left(-\frac{m}{2}\ln\left(2\pi\Delta\right) - D_v\left(x;\theta\right) + \frac{c_X^{(-1)}\left(x|x_0;\theta\right)}{\Delta}\right)\sum\nolimits_{k=0}^{K} c_X^{(k)}\left(x|x_0;\theta\right)\frac{\Delta^k}{k!} \\ &+ \sum\nolimits_{k=1}^{K} d_X^{(k)}\left(x|x_0;\theta\right)\frac{\Delta^k}{k!} \end{aligned} \qquad (8.11)$$

Again, the series can be calculated up to arbitrary order $K$ and the unknowns are the coefficients $c_X^{(k)}$ and $d_X^{(k)}$. The difference between the coefficients $c_X^{(k)}$ in (8.11) and $C_X^{(k)}$ in (4.12) is due to the fact that the former is written for $\ln p_X$ while the latter is for $p_X$ itself; the two coefficients families match once the terms of the Taylor series of $\ln(p_X^{(K)})$ in $\Delta$ are matched to the coefficients $C_X^{(k)}$ of the direct Taylor series $\ln p_X^{(K)}$. The coefficients $d_X^{(k)}$ are the new terms needed to capture the presence of the jumps in the transition function. The latter terms are needed to capture the different behavior of the tails of the transition density when jumps are present. (These tails are not exponential in $x$, hence the absence of a the factor $\exp(c_X^{(-1)}\Delta^{-1})$ in front of the sum of $d_X^{(k)}$ coefficients.) The coefficients can be computed analogously to the pure diffusive case.

An alternative extension of Aït-Sahalia (2002b) which applies to processes driven by more general Lévy jump processes than compound Poisson is due to Schaumburg (2001). There are cases where that expansion is not fully computable in closed form, as it requires computation of the orthonormal polynomials associated with the Lévy jump measure (just like the Hermite polynomials which form the basis for the method in the basic diffusive case are the natural family to use when the driving process is Brownian motion, i.e., Gaussian.)

# 9 Relaxing the Model Specification: Allowing for Non-Markov Dynamics

Doing inference without Assumption 1 is asking for too much at this point. However, it is possible to test that hypothesis. The specification analysis described in Section 6 assumes that the process is Markovian. In Aït-Sahalia (1996a) and Aït-Sahalia (2002a), I describe a set of observable implications which follow from the Markov property. A necessary condition for the process $X$ to be Markovian is that its transition function satisfy the Chapman-Kolmogorov equation in the form

$$p\left(y,t_3|x,t_1\right)=\int_{z\in S}p\left(y,t_3|z,t_2\right)p\left(z,t_2|x,t_1\right)dz \tag{9.1}$$

for every $t_3>t_2>t_1\geq 0$, $x$ and $y$ in $S$.

Under time-homogeneity, the Markov hypothesis can then be tested in the form $H_0$ against $H_1$, where

$$\begin{cases} H_0: & p\left(y|x,2\Delta\right)-r\left(y|x,2\Delta\right)=0 \quad \text{for all } (x,y)\in S^2 \\ H_1: & p\left(y|x,2\Delta\right)-r\left(y|x,2\Delta\right)\neq 0 \quad \text{for some } (x,y)\in S^2 \end{cases} \tag{9.2}$$

with

$$r\left(y|x,2\Delta\right)\equiv\int_{z\in S}p\left(y|z,\Delta\right)p\left(z|x,\Delta\right)dz. \tag{9.3}$$

Both $p\left(y|x,\Delta\right)$ and $p\left(y|x,2\Delta\right)$ can be estimated from data sampled at interval $\Delta$, thanks to time homogeneity. The successive pairs of observed data $(x_0,x_\Delta)$, $(x_\Delta,x_{2\Delta})$, $(x_{2\Delta},x_{3\Delta})$, etc., form a sample from the distribution with density $p\left(y|x,\Delta\right)$, from which the estimator $\hat{p}\left(y|x,\Delta\right)$ can be constructed and then $\hat{r}\left(y|x,2\Delta\right)$ as indicated in equation (9.3). Meanwhile, the successive pairs $(x_0,x_{2\Delta})$, $(x_\Delta,x_{3\Delta})$, $\ldots$, form a sample from the distribution with density $p\left(y|x,2\Delta\right)$ which can be used to form the direct estimator $\hat{p}\left(y|x,2\Delta\right)$.

In other words, the test compares a direct estimator of the $2\Delta$-interval conditional density, $\hat{p}\left(y|x,2\Delta\right)$, to an indirect estimator of the $2\Delta$-interval conditional density, $\hat{r}\left(y|x,2\Delta\right)$, obtained by iterating a direct $\Delta$-interval estimator of $\hat{p}\left(y|x,\Delta\right)$ according to (9.3). If the process is actually Markovian, then the two estimates should be close (for some distance measure) in a sense made precise by the use of the statistical distribution of these estimators.

If instead of $2\Delta$ transitions we test the replicability of $j\Delta$ transitions, where $j$ is an integer greater or equal to 2, there is no need to explore all the possible combinations of these $j\Delta$ transitions in terms of shorter ones $(1,j-1)$, $(2,j-2)$, $\ldots$: verifying equation (9.1) for one combination is sufficient as can be seen by a recursion argument. In the event of a rejection of $H_0$ in (9.2), there is no need to consider transitions of order $j$. In general, a vector of "transition equalities" can be tested in a single pass in a GMM framework with as many moment conditions as transition intervals.

In Aït-Sahalia and Fan (2005b), we propose two classes of tests for restriction (9.2) based on nonparametric estimation of the transition densities and distributions. To be more specific, observe that

$$r(y|x, 2\Delta) = E\{p(y|X_\Delta, \Delta)|X_0 = x\}, \tag{9.4}$$

Hence, the function $r(y|x, 2\Delta)$ can be estimated by regressing nonparametrically $\hat{p}(y|X_{j\Delta}, \Delta)$ on $X_{(j-1)\Delta}$. This avoids integration in (9.3) and makes implementations and theoretical studies easier. Local linear estimator will be applied, resulting in an estimator $\hat{r}(y|x, 2\Delta)$. A nonparametric test statistic for problem (9.2) is naturally

$$T_4 = \sum_{i=1}^{n} \{\hat{p}(X_{(i+1)\Delta}|X_{i\Delta}, 2\Delta) - \hat{r}(X_{(i+1)\Delta}|X_{i\Delta}, 2\Delta)\}^2 w(X_{i\Delta}, X_{(i+1)\Delta}). \tag{9.5}$$

The transition distribution-based tests can be formulated too. Let $\hat{P}(y|x, 2\Delta)$ be the direct estimator for the $2\Delta$-conditional distribution. Let $R(y|x, 2\Delta)$ be the cumulated version of $r(y|x, 2\Delta)$, which can be estimated by regression transition distribution $\hat{P}(y|X_{j\Delta}, \Delta)$ on $X_{(j-1)\Delta}$, yielding $\hat{R}(y|x, 2\Delta)$. The transition distribution based test will naturally be

$$T_5 = \sum_{i=1}^{n} \{\hat{P}(X_{(i+1)\Delta}|X_{i\Delta}, 2\Delta) - \hat{R}(X_{(i+1)\Delta}|X_{i\Delta}, 2\Delta)\}^2 w(X_{i\Delta}, X_{(i+1)\Delta}). \tag{9.6}$$

Note that this test statistic involves only one-dimensional smoothing. Hence, it is expected to be more stable than $T_4$. In addition, the null distribution can be better approximated by the asymptotic null distribution.

# 10 Relaxing the Sampling Assumptions: Allowing for Randomly Spaced Sampling Intervals

Transaction-level data are not only discretely sampled in time, they are also sampled at random time intervals. In discrete time, models allowing for random times have been studied by Engle and Russell (1998) and Engle (2000).

In the context of continuous-time, this calls for relaxing Assumption 5, at least to Assumption 22. For example, if the $\Delta_i$'s are random and i.i.d., then $E[\Delta]$ has the usual meaning, but even if this is not the case, by $E[\Delta]$ we mean the limit (in probability, or just the limit if the $\Delta_i$'s are non-random) of $\sum_{i=1}^{n} \Delta_i/n$ as $n$ tends to infinity. This permits the inclusion of the random non-i.i.d. and the nonrandom (but possibly irregularly spaced) cases for the $\Delta_i$'s. At the cost of further complications, the theory allows for dependence in the sampling intervals, whereby $\Delta_n$ is drawn conditionally on $(Y_{n-1}, \Delta_{n-1})$, which is the assumption under which Aït-Sahalia and Mykland (2003) work. Renault and Werker (2003) argue for more general specification of the randomness driving the sampling intervals than allowed under Assumption 22, which would give rise to more general version of the likelihood function than just discussed (see the related discussion on causality p. 491-494 in Aït-Sahalia and Mykland (2003)).

In order to concentrate on the main ideas, I will focus here on Assumption 22 only.

## 10.1 Likelihood Inference with Random Sampling Intervals

In Aït-Sahalia and Mykland (2003), we describe the additional effect that the randomness of the sampling intervals might have when estimating a continuous-time model with discrete data, as would be the case with transaction-level returns data. We disentangle the effect of the sampling randomness from the effect of the sampling discreteness, and compare their relative magnitudes. We also examine the effect of simply ignoring the sampling randomness. We achieve this by comparing the properties of three likelihood-based estimators, which make different use of the observations on the state process and the times at which these observations have been recorded. We design these estimators in such a way that each one of them is subject to a specific subset of the effects we wish to measure. As a result, the differences in their properties allow us to zero in and isolate these different effects.

I will now describe the effect that the very existence of such a distribution (as opposed to having non-random times between trades) would have on the behavior of estimators of continuous-time models for the price process, as well as with the interesting issues arising at the interface of the continuous and discrete time domains. We assume that the discrete data we observe have been generated by a time-homogeneous stationary diffusion on the real line $dX_t = \mu(X_t;\kappa)dt + \sigma dW_t$. We will show that the properties of estimators vary widely depending upon whether only the drift or the diffusion parameters, or both together, are estimated. Hence we consider the three cases of estimating $\theta = (\kappa, \sigma^2)$ jointly, estimating $\theta = \kappa$ with $\sigma^2 = \sigma_0^2$ known or estimating $\theta = \sigma^2$ with $\kappa = \kappa_0$ known. The parameter vector $\theta$ is to be estimated at time $T$ on the basis of $N_T + 1$ discrete observations, the $Y_n$'s given by $Y_n = X_{\tau_n}$. We let the sampling intervals $\Delta_n = \tau_n - \tau_{n-1}$ be random variables.

The first estimator of $\theta$ we consider is the Full Information Maximum Likelihood (FIML) estimator, using the bivariate observations $(Y_n, \Delta_n)$; the second is the partial information maximum likelihood estimator using only the state observations $Y_n$, with the sampling intervals integrated out (IOML for Integrated Out Maximum Likelihood); the third is the pseudo maximum likelihood estimator pretending that the observations times are fixed (PFML for Pretend Fixed Maximum Likelihood). Not surprisingly, the first estimator, FIML, is asymptotically efficient, making the best possible use of the joint data $(Y_n, \Delta_n)$. The second estimator, IOML, corresponds to the asymptotically optimal choice if one recognizes that the sampling intervals $\Delta_n$'s are random but does not observe them. The third estimator, PFML, corresponds to the "head-in-the-sand" policy consisting of acting as if the sampling intervals were all identical (pretending that $\Delta_n = \bar{\Delta}$ for all $n$) when in fact they are random.

Both FIML and IOML confront the randomness issue head-on. FIML uses the recorded sampling times, IOML does not, but still recognizes their relevance by integrating them out in the absence of observations on them. Because the data are always discretely sampled, each estimator is subject to the "cost of discreteness," which we define to be the additional variance relative to the variance of an asymptotically efficient estimator based on the full continuous-time sample path. It also represents the error that one would make if one were to use continuous-time asymptotics when the data are in fact discretely sampled. However, FIML is only subject to the cost of discreteness, while IOML is penalized by both the fact that the data are discrete (the continuous-time sample path is not observed) and randomly spaced in time (the sampling intervals are not observed). The additional variance of IOML over that of FIML will therefore be identified as the "cost of randomness," or the cost of not observing the randomly-spaced sampling intervals. But if in fact one had recorded the observations times but chosen not to use them in the empirical estimation phase, then what we call the cost of randomness can be interpreted as the cost of throwing away, or not using, these data.

By contrast, PFML does as if the sampling times were simply not randomly spaced. Comparing it to FIML gives rise to the cost imputable to *ignoring* the randomness of the sampling intervals, as opposed to the what we call the cost of randomness, which is the cost due to *not observing* the randomly-spaced sampling intervals. In the former case, one (mistakenly) uses PFML, while in the latter case one realizes that the intervals are informative but, in their absence, IOML is the best one can do. Different types of estimation strategies in empirical market microstructure that do not use the sampling intervals can be viewed as versions of either IOML or PFML, depending upon their treatment of the sampling intervals: throw them away, or ignore their randomness. They will often be suboptimal versions of these estimators because they are subject to an additional efficiency loss if they do not use maximum-likelihood.

All three estimators rely on maximizing a version of the likelihood function of the observations. Let $p(y_1|y_0, \delta, \theta)$ denote the transition function of the process $X$. Because of the time homogeneity of the model, the transition function $p$ depends only on $\delta$ and not on $(t, t+\delta)$ separately. All three estimators make use of some functional of the density $p$: namely, $p(Y_n|Y_{n-1}, \Delta_n, \theta)$ for FIML; the expectation $\tilde{p}(Y_n|Y_{n-1}, \theta)$ of $p(Y_n|Y_{n-1}, \Delta_n, \theta)$ over the law of $\Delta_n|Y_{n-1}$ for IOML; and $p(Y_n|Y_{n-1}, \bar{\Delta}, \theta)$ for PFML (i.e., like FIML except that $\bar{\Delta}$ is used in place of the actual $\Delta_n$). In practice, even though most diffusion models do not admit closed-form transition densities, all three estimators can be calculated for any diffusion $X$ using arbitrarily accurate closed-form approximations of the transition function $p$ (see Aït-Sahalia (2002b), described above). We also show that $\tilde{p}$ can be obtained in closed form in an important special case.

Under Assumption 12, we have that

$$T^{1/2}(\hat{\theta} - \bar{\theta}) \rightarrow N(0, \Omega_\theta)$$

with

$$\Omega_\theta = \begin{pmatrix} \omega_\kappa & \omega_{\kappa\sigma^2} \\ \omega_{\kappa\sigma^2} & \omega_{\sigma^2} \end{pmatrix} \tag{10.1}$$

For FIML and IOML, $\bar{\theta} = \theta_0$, but PFML is going to be asymptotically biased.

We then derive Taylor expansions of the asymptotic variance and bias of these estimators. We denote a random variable from the common distribution of the sampling intervals as

$$\Delta = \varepsilon \Delta_0, \tag{10.2}$$

where $\varepsilon$ is deterministic and $\Delta_0$ has a given finite distribution conditional on $Y_0$, whose density we write as $d(\Delta|Y_0)$. We compute Taylor expansions in $\varepsilon$ of the expectations of interest, around $\varepsilon = 0$ (the limiting case were the full continuous-time sample path is observable), leading to results of the type:

$$\begin{aligned}\Omega_\theta &= \Omega_\theta^{(0)} + \varepsilon\,\Omega_\theta^{(1)} + \varepsilon^2\,\Omega_\theta^{(2)} + O\left(\varepsilon^3\right)\\ \bar{\theta} - \theta_0 &= \varepsilon\, b_\theta^{(1)} + \varepsilon^2\, b_\theta^{(2)} + O\left(\varepsilon^3\right)\end{aligned}$$

where the higher order terms in $\varepsilon$ correct the leading term for the discreteness of the sampling. Differences between estimation methods and data use show up in the functions $\Omega_\theta^{(i)}$ and $b_\theta^{(i)}$, $i = 0, 1, ....$

These calculations are based on a new operator, the *generalized infinitesimal generator* $\Gamma$ of the diffusion $X$. We show that

$$E\left[f(X, X_0, \Delta, \theta, \varepsilon)|X_0, \Delta\right] = \sum_{j=0}^{J} \frac{\Delta^j}{j!}\Gamma^j \cdot f + O\left(\Delta^{J+1}\right)$$

where $\Gamma$ is the operator that returns

$$\Gamma \cdot f \;=\; \Delta_0\,\mathcal{A}\cdot f + \frac{\partial f}{\partial \varepsilon} + \frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial \varepsilon} \tag{10.3}$$

with $\mathcal{A}$ denoting the standard infinitesimal generator, which for a diffusion yields:

$$\mathcal{A}\cdot f = \frac{\partial f}{\partial \delta} + \mu\frac{\partial f}{\partial y_1} + \frac{1}{2}\sigma_0^2\frac{\partial^2 f}{\partial y_1^2}$$

The operator $\Gamma$ is random in that it depends on $\Delta_0$.

Specifically, the log-likelihood function using all information is

$$\sum_{n=1}^{N-1} \ln p(X_{\tau_n}|X_{\tau_{n-1}}, \Delta_n, \theta) + \sum_{n=1}^{N-1} \ln d(\Delta_n|X_{\tau_{n-1}})$$

where $d$ is the density of the sampling interval given the most recent price observation (recall Assumption 22). Since we only care about $\theta$ (not the parameters, if any, entering the density $d$) we maximize

$$l_T(\theta) = \sum_{n=1}^{N-1} \ln p(X_{\tau_n}|X_{\tau_{n-1}}, \Delta_n, \theta).$$

Suppose now that the observation times are either not observable or discarded prior to conducting inference on $\theta$. They can be integrated out to obtain a proper likelihood. We would then base inference on the density

$$\tilde{p}(x|x_0, \theta) = E_\Delta\left[p(x|X_0, \Delta, \theta)\,|X_0 = x_0\right]$$

The part of the log-likelihood function dependent on $\theta$ is then

$$\lambda_T(\theta) = \sum_{n=1}^{N-1} \ln \tilde{p}(X_{\tau_n}|X_{\tau_{n-1}}, \theta).$$

Using the Hermite-based closed-form expansion of the transition function $p(x|x_0, \Delta, \theta)$ described above, the corresponding expansion for the density $\tilde{p}$ can also be obtained explicitly in the important special case where the density $d$ of $\Delta$ given $X_0$ is exponential with mean $E[\Delta|X_0]$, :

$$\tilde{p}^{(J)}(x|x_0,\theta) = \exp\left(\int_{x_0}^{x}\frac{\mu(w,\kappa)}{\sigma^2}dw\right)\times\sum_{j=0}^{J}\Bigg\{c_j\left(\frac{x}{\sigma}|\frac{x_0}{\sigma},\theta\right)\frac{2^{(1-2j)/4}E[\Delta|X_0]^{(2j-3)/4}}{j!\pi^{1/2}\sigma^{j+(3/2)}}$$
$$B_{j+(1/2)}\left(\frac{2^{1/2}\left|x-x_0\right|}{E[\Delta|X_0]^{1/2}\sigma}\right)\left|x-x_0\right|^{(2j+1)/2}\Bigg\}$$

where $B_{j+(1/2)}\left(z\right)$ is the Bessel $K$ function of half-integer order $j+(1/2)$, which is in closed-form for any $j$.

The cost of discreteness is the cost attributable to not observing the full continuous-time sample path. It is the coefficient at the first order $i$ in $\varepsilon$ for which the FIML variance differs from its continuous-time limit $\Omega_\theta^{(0)}$. It is also the error that one would make if one were to use continuous-time asymptotics ($\Omega_\theta^{(0)}$) instead of the full $\Omega_\theta$ when the data are in fact discretely sampled. The cost of randomness is the extra variance due to not using the sampling intervals: it is the first order $i$ in $\varepsilon$ at which the coefficient $\Omega_\theta^{(i)}$ for IOML differs from the corresponding coefficient $\Omega_\theta^{(i)}$ for FIML, and how much bigger the IOML coefficient at that order is compared to the FIML coefficient. It turns out that the cost of randomness is at least as great, and often substantially greater than the cost of discreteness.

Specifically, we have that

$$\Omega_\kappa^{\text{FIML}} = \Omega_\kappa^{(\text{FIML},0)} + O\left(\varepsilon^2\right) \tag{10.4}$$
$$\Omega_\kappa^{\text{IOML}} = \Omega_\kappa^{(\text{IOML},0)} + \varepsilon\,\Omega_\kappa^{(\text{IOML},1)} + O\left(\varepsilon^2\right), \tag{10.5}$$

where

$$\Omega_\kappa^{(\text{FIML},0)} = \Omega_\kappa^{(\text{IOML},0)} = \left(E_{Y_0}\left[\sigma^{-2}(Y_0,\gamma_0)\left(\partial\mu(Y_0,\kappa_0)/\partial\kappa\right)^2\right]\right)^{-1} \tag{10.6}$$

which is the the leading term in $\Omega_\kappa$ corresponding to efficient estimation of $\kappa$ with a continuous record of observations.

And the price of ignoring the sampling times $\tau_0,\tau_1,...$ when estimating $\kappa$ is, to first order, represented by

$$\Omega_\kappa^{(\text{IOML},1)} = \frac{E[Var[\Delta_0|\chi_1^2\Delta_0]]}{E\left[\Delta_0\right]}\;V,$$

and "$\chi_1^2$" is a $\chi_1^2$ distributed random variable independent of $\Delta_0$, and

$$V = \frac{\left(E_{Y_0}\left[\sigma_0^4\left(\frac{\partial^2\mu(Y_0,\beta_0)}{\partial y\partial\kappa}\right)^2\right] - 2E_{Y_0}\left[\sigma_0^2\frac{\partial\mu(Y_0,\kappa_0)}{\partial y}\left(\frac{\partial\mu(Y_0,\kappa_0)}{\partial\kappa}\right)^2\right]\right)}{4\;E_{Y_0}\left[\left(\frac{\partial\mu(Y_0,\kappa_0)}{\partial\kappa}\right)^2\right]^2}. \tag{10.7}$$

Note that $V\geq 0$ by the asymptotic efficiency of FIML.

And the leading term in $\Omega_\gamma$ corresponding to efficient estimation of $\gamma$ is

$$\Omega_\gamma^{\text{FIML}} = \varepsilon\;\Omega_\gamma^{(\text{FIML},1)} + O\left(\varepsilon^2\right)$$
$$\Omega_\gamma^{\text{IOML}} = \varepsilon\,\Omega_\gamma^{(\text{IOML},1)} + O\left(\varepsilon^2\right),$$

where

$$\Omega_{\gamma}^{(\mathrm{FIML},1)} = \Omega_{\gamma}^{(\mathrm{IOML},1)} = E\left[\Delta_0\right]\left(2E_{Y_0}\left[\left(\partial\sigma(Y_0,\gamma_0)/\partial\gamma\right)^2\sigma(Y_0,\gamma_0)^{-2}\right]\right)^{-1}.$$

In the special case where $\sigma^2$ is constant ($\gamma = \sigma^2$), this becomes the standard AVAR of MLE from iid Gaussian observations, i.e., $\Omega_{\gamma}^{(1)} = 2\sigma_0^4 E\left[\Delta_0\right]$.

These leading terms are achieved in particular when $h$ is the likelihood score for $\kappa$ and $\gamma$ respectively, as analyzed in Aït-Sahalia and Mykland (2003), but also by other estimating functions that are able to mimic the behavior of the likelihood score at the leading order.

In other words, when estimating $\gamma$, the costs incurred due to the discreteness and the randomness of time both appear at the same order $\varepsilon^1$, while the limiting variance (the term of order $\varepsilon^0$) is zero if the full continuous time sample path is observed. The loss due to not using sampling times can be an arbitrarily large *multiple* of the loss due to the discreteness of the data. When estimating $\kappa$, the cost of randomness is of order $\varepsilon^1$: it is therefore an order of magnitude in $\varepsilon$ bigger than the loss from observing the process discretely rather than continuously (which is of order $\varepsilon^2$). However, both are small relative to the sampling variable that is present even if the full continuous time sample path is observed (order $\varepsilon^0$).

The cost of ignoring the randomness is the cost imputable to *ignoring* the randomness of the sampling intervals, by doing *as if* the $\Delta_n$'s were not random, as opposed to the cost of randomness, which is the cost due to *not observing* the randomly-spaced sampling intervals but still accounting for their randomness. In terms of RMSE, any biased estimator such as PFML will always do worse than an unbiased estimator since its variance is $O(T^{-1})$ whereas the squared bias is $O(1)$.

## 10.2 GMM Under Random Sampling Intervals

Aït-Sahalia and Mykland (2004) study the asymptotic properties of estimators based on general moment conditions, of the type described in Section 4.2.1, when the data are sampled not only discretely but also randomly (Assumption 22). Of course, the results also apply to the situation of Assumption 5 as a special case (where $\Delta_n = \bar{\Delta}$ for all $n$, corresponding to the distribution of $\Delta_n$ being a Dirac mass at $\bar{\Delta}$) so this Section covers the asymptotics for all the estimators described in Section 4.2.1 under Assumption 5.

The moment conditions are $h(y_1, y_0, \delta, \theta, \varepsilon)$. Compared to the situation discussed in Section 4.2.1, where Assumption 5 held, the moment conditions under Assumption 22 could conceivably depend on $\varepsilon$ in addition or instead of $\delta$ ($\varepsilon$ is defined in (10.2)). As before, $y_1$ plays the role of the forward state variable, $y_0$ the role of the backward state variable and $\delta$ the sampling interval.

Then let

$$m_T(\theta) \equiv N_T^{-1}\sum_{n=1}^{N_T-1} h(Y_n, Y_{n-1}, \Delta_n, \theta, \varepsilon) \tag{10.8}$$

where $Y_n = X_{\tau_n}$ and obtain $\hat{\theta}$ by minimizing the quadratic form

$$Q_T(\theta) \equiv m_T(\theta)'\, W_T\, m_T(\theta) \tag{10.9}$$

where $W_T$ is an $r \times r$ positive definite weight matrix assumed to converge in probability to a positive definite limit $W_\theta$. If the system is exactly identified, $r = d$, the choice of $W_T$ is irrelevant and minimizing (10.9) amounts to setting $m_T(\theta)$ to 0. The function $h$ is known in the econometric literature as a "moment function" (see Hansen (1982)) and in the statistical literature as an "estimating equation" (see e.g., Godambe (1960) and Heyde (1997).)

Consistency of $\hat{\theta}$ is guaranteed as long as $h$ is such that

$$E_{\Delta,Y_1,Y_0}\left[h(Y_1, Y_0, \Delta, \theta_0, \varepsilon)\right] = 0. \tag{10.10}$$

where $E_{\Delta,Y_1,Y_0}$ denotes expectations taken with respect to the joint law of $(\Delta, Y_1, Y_0)$ at the true parameter $\theta_0$, and write $E_{\Delta,Y_1}$, etc., for expectations taken from the appropriate marginal laws of $(\Delta, Y_1)$, etc.

Some otherwise fairly natural estimating strategies lead to inconsistent estimators. To allow for this, we do not assume that (10.10) is necessarily satisfied. Rather, we simply assume that the equation

$$E_{\Delta,Y_1,Y_0}\left[h(Y_1, Y_0, \Delta, \theta, \varepsilon)\right] = 0 \tag{10.11}$$

admits a unique root in $\theta$, which we define as $\bar{\theta} = \bar{\theta}(\theta_0, \varepsilon)$.

For the estimator to be consistent, it must be that $\bar{\theta} \equiv \theta_0$ but, again, this will not be the case for every estimation method. However, in all the cases we consider, and one may argue for *any* reasonable estimation method, the bias will disappear in the limit where $\varepsilon \to 0$, i.e., $\bar{\theta}(\theta_0, 0) = \theta_0$ (so that there is no bias in the limiting case of continuous sampling) and we have an expansion of the form

$$\bar{\theta} = \bar{\theta}(\theta_0, \varepsilon) = \theta_0 + b^{(1)}\varepsilon + b^{(2)}\varepsilon^2 + O\left(\varepsilon^3\right). \tag{10.12}$$

With $N_T/T$ converging in probability to $(E\left[\Delta\right])^{-1}$, it follows from standard arguments (see e.g., Hansen (1982)) that $\sqrt{T}(\hat{\theta} - \bar{\theta})$ converges in law to $N(0, \Omega_\theta)$, with

$$\Omega_\theta^{-1} = (E\left[\Delta\right])^{-1} D_\theta' S_\theta^{-1} D_\theta. \tag{10.13}$$

where

$$D_\theta \equiv E_{\Delta,Y_1,Y_0}\left[\dot{h}(Y_1, Y_0, \Delta, \bar{\theta}, \varepsilon)\right], \; S_{\theta,j} \equiv E_{\Delta,Y_1,Y_0}\left[h(Y_{1+j}, Y_j, \Delta, \bar{\theta}, \varepsilon) h(Y_1, Y_0, \Delta, \bar{\theta}, \varepsilon)'\right]$$

and $S_\theta \equiv \sum_{j=-\infty}^{+\infty} S_{\theta,j}$. If $r > d$, the weight matrix $W_T$ is assumed to be any consistent estimator of $S_\theta^{-1}$; otherwise its choice is irrelevant. A consistent first-step estimator of $\theta$, needed to compute the optimal weight matrix, can be obtained by minimizing (10.9) with $W_T = Id$.

The simplest case arises when the moment function is a martingale, i.e.,

$$E_{\Delta,Y_1}\left[h(Y_1, Y_0, \Delta, \theta_0, \varepsilon)|Y_0\right] = 0. \tag{10.14}$$

since then $S_{\theta,j} = 0$ for all $j \neq 0$. However, this property will not be satisfied by many otherwise relevant examples, and so Aït-Sahalia and Mykland (2004) also allow for near-martingale moment conditions. To define the distance from a moment function to a martingale, denote by $h_i$ the $i$th component of $h$, and define $q_i$ and $\alpha_i$ by

$$E_{\Delta,Y_1}\left[h_i(Y_1,Y_0,\Delta,\bar{\theta},\varepsilon)|Y_0\right]=\varepsilon^{\alpha_i}q_i(Y_0,\beta_0,0)+O(\varepsilon^{\alpha_i+1}) \tag{10.15}$$

where $\alpha_i$ is an integer greater than or equal to zero for each moment function $h_i$. $\alpha_i$ is an index of the order at which the moment component $h_i$ deviates from a martingale. A martingale moment function corresponds to the limiting case where $\alpha_i=+\infty$. When the moment functions are not martingales, the difference $T_\theta\equiv S_\theta-S_{\theta,0}$ is a matrix whose element $(i,j)$ has a leading term of order $O\left(\varepsilon^{\min(\alpha_i,\alpha_j)}\right)$. Intuitively, the closer $h$ will be to a martingale, the smaller $T_\theta$.

This is all nice and well, but, for given $h$ function(s), how does one compute the matrices $D_\theta$ and $S_\theta$ in (10.13), and the coefficients of the bias expansion (10.12)? Using the same tools as in the special case of the likelihood, i.e., the generalized infinitesimal general defined in (10.3), Aït-Sahalia and Mykland (2004) derive expressions of the form

$$\Omega_\theta=\Omega_\theta^{(0)}+\varepsilon\,\Omega_\theta^{(1)}+\varepsilon^2\,\Omega_\theta^{(2)}+O\left(\varepsilon^3\right) \tag{10.16}$$

and detail the effects of the choice of inference strategy and/or distribution of the sampling intervals, as they show up in the functions $\Omega_\theta^{(i)}$, $i=0,1,...$ and in the bias coefficients $b^{(i)}$, $i=0,1,...$ of (10.12), since these coefficients are derived in closed form for arbitrary $h$ moment conditions.

These general results can be applied to exact likelihood inference as in Aït-Sahalia and Mykland (2003), but also to study the properties of estimators of the drift and diffusion coefficients obtained by replacing the true likelihood function $l(y_1|y_0,\delta,\theta)$ with its discrete Euler approximation

$$l^E(y_1|y_0,\delta,\theta)=-\frac{1}{2}\ln(2\pi\sigma^2(y_0;\gamma)\delta)-\frac{(y_1-y_0-\mu(y_0;\kappa)\delta)^2}{2\sigma^2(y_0;\gamma)\delta}. \tag{10.17}$$

This estimator is commonly used in empirical work in finance, where researchers often write a theoretical model set in continuous-time but then switch gear in their empirical work, in effect estimating the parameters of the discrete time series model

$$X_{t+\Delta}-X_t=\mu(X_t;\kappa)\Delta+\sigma(X_t;\gamma)\sqrt{\Delta}\eta_{t+\Delta} \tag{10.18}$$

where the disturbance $\eta$ is $N(0,1)$. The properties of this estimator have been studied in the case where $\Delta$ is not random by Florens-Zmirou (1989). Our results apply to this particular situation as a special case.

In the terminology of the general case, our vector of moment functions is the Euler score

$$h(y_1,y_0,\delta,\theta,\varepsilon)=\begin{bmatrix}\dot{l}_\theta^E(y_1|y_0,\delta,\theta)\\ \dot{l}_{\sigma^2}^E(y_1|y_0,\delta,\theta)\end{bmatrix} \tag{10.19}$$

We find that the asymptotic variance is, to first order in $\varepsilon$, the same as for MLE inference. The impact of using the approximation is to second order in variances (and of course is responsible for bias in the estimator). When estimating one of the two parameters with the other known, the impact of the discretization approximation on the variance (which MLE avoids) is one order of magnitude higher than the effect of the discreteness of the data (which MLE is also subject to).

In Aït-Sahalia and Mykland (2005), we apply this general theory to determine the asymptotic properties of estimators using the C1 and C2 moment conditions of Hansen and Scheinkman (1995) discussed in Section 4.2.1; as a special case, our results give these properties in the form (10.16) – these estimators are unbiased, so all the coefficients in (10.12) are identically zero – when the sampling intervals are fixed and deterministic, but also when they are random. Since we have closed form expressions for the coefficients $\Omega_\theta^{(i)}$, we then derive the optimal choice of test functions by minimizing the asymptotic variance of the estimator and compare the results to the efficient choice represented by likelihood inference (where $h$ is given by the log-likelihood score function).

# 11 Relaxing the Sampling Assumptions: Allowing for Market Microstructure Noise

Over the past few years, price data sampled at very high frequency have become increasingly available, in the form of the Olsen dataset of currency exchange rates or the TAQ database of NYSE stocks. In earlier work, I informally warned about the potential dangers of using high frequency financial data without accounting for their inherent noise (see page 529 of Aït-Sahalia (1996b)), and I now discuss a formal modelization of that phenomenon.

So, inference will now be conducted when Assumption 6 is relaxed and instead explicitly allow for the presence of observation errors. Under Assumption 15, the observed transaction or quoted log-price will be $Y_t$, and take the form of the unobservable efficient log-price $X_t$ plus some noise component due to the imperfections of the trading process, $\varepsilon_t$, collectively known as market microstructure noise.

So consider now the implications of Assumption 15 for the estimation of the volatility of the efficient log-price process, $X_t$, assumed to follow a SDE of the type studied in the previous Sections:

$$dX_t = \mu_t dt + \sigma_t dW_t,$$

using discretely sampled data on the transaction price process at times 0, $\Delta$,..., $N\Delta = T$.

With either quote or transaction data, we are in a situation where $\Delta$ will be measured in seconds rather than minutes or hours. Under these circumstances, the drift is of course irrelevant, both economically and statistically, and so we shall focus on functionals of the $\sigma_t$ process and set $\mu_t = 0$. It is also the case that transactions and quotes data series in finance are often observed at random time intervals (see Section 10 for inference under these circumstances). We make essentially no assumptions on the $\sigma_t$ process: its driving process can of course be correlated with the Brownian motion $W_t$ driving the asset price process, and it need not even have continuous sample paths.

The noise term $\varepsilon$ summarizes a diverse array of market microstructure effects, which can be roughly divided into three groups. First, $\varepsilon$ represents the frictions inherent in the trading process: bid-ask bounces, discreteness of price changes and rounding, trades occurring on different markets or networks, etc. Second, $\varepsilon$ captures informational effects: differences in trade sizes or informational content of price changes, gradual response of prices to a block trade, the strategic component of the order flow, inventory control effects, etc. Third, $\varepsilon$ encompasses measurement or data recording errors such as prices entered as zero, misplaced decimal points, etc., which are surprisingly prevalent in these types of data. As is clear from the laundry list of potential sources of noise, the data generating process for $\varepsilon$ is likely to be quite involved. Therefore, robustness to departures from any assumptions on $\varepsilon$ is desirable.

Different solutions have been proposed for this problem. In the constant $\sigma$ case, Zhou (1996) who proposes a bias correcting approach based on autocovariances. The behavior of this estimator has been studied by Zumbach et al. (2002). In the constant $\sigma$ case, efficient likelihood estimation of $\sigma$ is studied Aït-Sahalia et al. (2005b), showing that incorporating $\varepsilon$ explicitly in the likelihood function of the observed log-returns $Y$ provides consistent, asymptotically normal and efficient estimators of the parameters.

Hansen and Lunde (2006) study the Zhou estimator and extensions in the case where volatility is time varying but conditionally nonrandom. Related contributions have been made by Oomen (2006) and Bandi and Russell (2003). The Zhou estimator and its extensions, however, are inconsistent despite being unbiased. This means in this particular case that, as the frequency of observation increases, the estimator diverges instead of converging to the object of interest.

In the parametric case, Aït-Sahalia et al. (2005b) show that modelling $\varepsilon$ explicitly through the likelihood restores the first order statistical effect that sampling as often as possible is optimal. But, more surprisingly, this is true even if one misspecifies the distribution of $\varepsilon$. This robustness result argues for incorporating $\varepsilon$ when estimating continuous time models with high frequency financial data, even if one is unsure about the true distribution of the noise term. We also study the same questions when the observations are sampled at random time intervals $\Delta$, which are an essential empirical feature of transaction-level data.

In the nonparametric or stochastic volatility case, the first consistent estimator is due to Zhang et al. (2005). Ignoring the noise is worse than in the parametric case, in that the quadratic variation no longer estimates a mixture of the price volatility and the noise, but now estimates exclusively the variance of the noise. What makes this situation nonstandard among the class of measurement error problems is the fact that the contribution of the measurement error to the quantity being measured is the dominating one (typically, the effect of the measurement error is of the same magnitude, or smaller, than that of the signal). We propose a solution based on subsampling and averaging, which again makes use of the full data; for reasons that will become clear below, we call this estimator Two Scales Realized Volatility (TSRV).

## 11.1 Parametric Modeling: Likelihood Corrections and Robustness

Suppose for now that market microstructure noise is present, under Assumptions 15 and 17, but that the presence of the $\varepsilon's$ (iid, mean 0, variance $a^2$) is ignored when estimating $\sigma^2$. Assume also that we are under Assumption 4 (with $\mu = 0$ since the drift is essentially impossible to measure accurately at these frequencies) and Assumption 5.

In other words, we use the same $N(0, \sigma^2\Delta)$ likelihood as under Section 2 even though the true structure of the observed log-returns $Y$ is given by an MA(1) process since

$$Y_i = \sigma\left(W_{\tau_i} - W_{\tau_{i-1}}\right) + \varepsilon_{\tau_i} - \varepsilon_{\tau_{i-1}}. \tag{11.1}$$

The variance and first order correlation coefficient of the log-returns are $(\gamma^2, \eta)$ where $\gamma^2(1+\eta^2) = Var[Y_i] = \sigma^2\Delta + 2a^2$ and $\gamma^2\eta = cov(Y_i, Y_{i-1}) = -a^2$.

Theorem 1 of Aït-Sahalia et al. (2005b) gives an exact small sample (finite $T$) result for the bias and variance of the estimator $\hat{\sigma}^2$. Its RMSE has a unique minimum in $\Delta$ which is reached at the optimal sampling interval:

$$\begin{aligned}\Delta^* &= \left(\frac{2a^4T}{\sigma^4}\right)^{1/3}\left(\left(1-\left(1-\frac{2\left(3a^4+\mathrm{Cum}_4\left[\varepsilon\right]\right)^3}{27\sigma^4a^8T^2}\right)^{1/2}\right)^{1/3}\right. \\ &\qquad \left.+\left(1+\left(1-\frac{2\left(3a^4+\mathrm{Cum}_4\left[\varepsilon\right]\right)^3}{27\sigma^4a^8T^2}\right)^{1/2}\right)^{1/3}\right)\end{aligned} \tag{11.2}$$

where $\mathrm{Cum}_4\left[\varepsilon\right]$ denotes the fourth cumulant of the random variable $\varepsilon$.

But this solution provides at best a partial answer to the problem. Indeed, the presence of the noise is acknowledged by reducing the sampling frequency, but this cannot be the optimal answer. Such an answer must proceed by fully specifying the likelihood function of the observations, including the noise. This immediately raises the question of what distribution to assume for the noise. To start, let us suppose that $\varepsilon \sim N(0, a^2)$. Then the likelihood function for the $Y's$ is then given by

$$\begin{aligned} l(\sigma^2, a^2) &= -\ln\det(V)/2 - N\ln(2\pi\gamma^2)/2 - (2\gamma^2)^{-1}Y'V^{-1}Y \\ V = [v_{ij}] &= \begin{pmatrix} 1+\eta^2 & \eta & \cdots & 0 \\ \eta & 1+\eta^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \eta \\ 0 & \cdots & \eta & 1+\eta^2 \end{pmatrix} \end{aligned} \tag{11.3}$$

We can see that the MLE $(\hat{\sigma}^2, \hat{a}^2)$ is consistent and its asymptotic variance is given by

$$\mathrm{AVAR}_{\mathrm{normal}}(\hat{\sigma}^2, \hat{a}^2) = \begin{pmatrix} 4\left(\sigma^6\Delta\left(4a^2+\sigma^2\Delta\right)\right)^{1/2} + 2\sigma^4\Delta & -\sigma^2\Delta h \\ \bullet & \frac{\Delta}{2}\left(2a^2+\sigma^2\Delta\right)h \end{pmatrix} \tag{11.4}$$

with $h \equiv 2a^2 + \left(\sigma^2\Delta\left(4a^2+\sigma^2\Delta\right)\right)^{1/2} + \sigma^2\Delta$. Since $\mathrm{AVAR}_{\mathrm{normal}}(\hat{\sigma}^2)$ is increasing in $\Delta$, it is optimal to sample as often as possible. Further, since

$$\mathrm{AVAR}_{\mathrm{normal}}(\hat{\sigma}^2) = 8\sigma^3a\Delta^{1/2} + 2\sigma^4\Delta + o(\Delta), \tag{11.5}$$

the loss of efficiency relative to the case where no market microstructure noise is present (and $\mathrm{AVAR}(\hat{\sigma}^2) = 2\sigma^4\Delta$ as given in (11.5) if $a^2 = 0$ is not estimated, or $\mathrm{AVAR}(\hat{\sigma}^2) = 6\sigma^4\Delta$ if $a^2 = 0$ is estimated) is at order

$\Delta^{1/2}$.

Of course, we can only compute this MLE if we assume a model for the noise distribution, in this case $\varepsilon \sim N(0, a^2)$. Given the complexity and diversity of the mechanisms giving rise to market microstructure noise, it is somewhat likely that any parametric assumption on the noise distribution would be misspecified. What, then, is the effect of misspecifying the distribution of the microstructure noise? Specifically, we consider the case where the $\varepsilon' s$ are assumed by the econometrician to be normally distributed when they are not. We still suppose that the $\varepsilon' s$ are iid with mean zero and variance $a^2$.Since the econometrician assumes $\varepsilon \sim N$, inference is still done with the Gaussian log-likelihood $l(\sigma^2, a^2)$, using the scores $\dot{l}_{\sigma^2}$ and $\dot{l}_{a^2}$ as moment functions. But since the expected values of $\dot{l}_{\sigma^2}$ and $\dot{l}_{a^2}$ only depend on the second order moment structure of the log-returns $Y$, which is unchanged by the absence of normality, the moment functions are unbiased $E_{\text{true}}[\dot{l}_{\sigma^2}] = E_{\text{true}}[\dot{l}_{a^2}]$ $= 0$ where "true" denotes the true distribution of the $Y's$. Hence the estimator $(\hat{\sigma}^2, \hat{a}^2)$ based on these moment functions remains consistent and the effect of misspecification therefore lies in the AVAR. By using the cumulants of the distribution of $\varepsilon$, we express the AVAR in terms of deviations from normality:

Theorem 2 shows that the estimator $(\hat{\sigma}^2, \hat{a}^2)$ is consistent and its asymptotic variance is given by

$$\text{AVAR}_{\text{true}}(\hat{\sigma}^2, \hat{a}^2) = \text{AVAR}_{\text{normal}}(\hat{\sigma}^2, \hat{a}^2) + \text{Cum}_4[\varepsilon] \begin{pmatrix} 0 & 0 \\ 0 & \Delta \end{pmatrix} \tag{11.6}$$

where $\text{AVAR}_{\text{normal}}(\hat{\sigma}^2, \hat{a}^2)$ is the asymptotic variance in the case where the distribution of $\varepsilon$ is Normal. We have therefore shown that $\text{AVAR}_{\text{normal}}(\hat{\sigma}^2, \hat{a}^2)$ coincides with $\text{AVAR}_{\text{true}}(\hat{\sigma}^2, \hat{a}^2)$ for all but the $(a^2, a^2)$ term.

We show in the paper how to interpret this in terms of the profile likelihood and the second Bartlett identity. Next, we relax Assumption 5 by considering the case where the $\Delta_i' s$ are random (for simplicity iid, independent of the $W$ process), which is a special case of Assumption 22. We Taylor-expand around $\bar{\Delta} = E[\Delta] : \Delta_i = \bar{\Delta}(1 + \epsilon \xi_i)$, where $\epsilon$ and $\bar{\Delta}$ are nonrandom and the $\xi_i' s$ are iid random variables with mean zero. we show that in that situation the MLE $(\hat{\sigma}^2, \hat{a}^2)$ is again consistent, this time with asymptotic variance obtained in closed form as

$$\text{AVAR}(\hat{\sigma}^2, \hat{a}^2) = A^{(0)} + \epsilon^2 A^{(2)} + O(\epsilon^3) \tag{11.7}$$

where $A^{(0)}$ is the asymptotic variance matrix already present in the deterministic sampling case except that it is evaluated at $\bar{\Delta}$, and the second order correction term $A^{(2)}$ is proportional to $Var[\xi]$ and is therefore zero in the absence of sampling randomness.

Further extensions considered in the paper include: the presence of a drift coefficient, which, because of the block-diagonality of the asymptotic variance matrix, does not affect our earlier conclusions; serially correlated noise which, instead of being iid (Assumption 17) follows a mean reverting process (a special case of Assumption 18). This type of noise could capture the gradual adjustment of prices in response to a shock such as a large trade. Now, the variance contributed by the noise is of order $O(\Delta)$, that is of the same order as the variance of the efficient price process $\sigma^2\Delta$, instead of being of order $O(1)$ under Assumption 17. We show that this type of noise is not nearly as bad as iid noise for the purpose of inferring $\sigma^2$. The final extension covers the case where the noise is correlated with the price process, thereby relaxing Assumption 16: microstructure noise attributable to informational effects could be correlated with the efficient price process, since it is generated by the response of market participants to information signals. We show how the form of the variance matrix of the observed log-returns $Y$ must be altered in that case.

## 11.2 The Nonparametric Stochastic Volatility Case: The TSRV Estimator

I now examine the situation where we relax Assumption 4, and allow for stochastic volatility. We also replace Assumption 5 with Assumption 21, since there is no other way in the presence of unobserved stochastic volatility. With $dX_t = \sigma_t dW_t$, the object of interest is now the quadratic variation

$$\langle X, X \rangle_T = \int_0^T \sigma_t^2 dt \tag{11.8}$$

over a fixed time period $[0, T]$, typically one day in empirical applications. This quantity can then be used to hedge a derivatives' portfolio, forecast the next day's integrated volatility, etc. Without noise, $Y = X$ and the realized volatility (RV) estimator $[Y, Y]_T^{(\text{all})} = \sum_{i=1}^n (Y_{t_{i+1}} - Y_{t_i})^2$ provides an estimate of the quantity $\langle X, X \rangle_T$, and asymptotic theory would lead one to sample as often as possible, or use all the data available, hence the "all" superscript. The sum $[Y, Y]_T^{(\text{all})}$ converges to the integral $\langle X, X \rangle_T$, with a known distribution, a result which dates back to Jacod (1994) and Jacod and Protter (1998) in great generality (i.e., for continuous semi-martingales):

$$[Y, Y]_T^{(\text{all})} \quad \overset{\mathcal{L}}{\approx} \quad \langle X, X \rangle_T \quad + \quad \left[ \frac{2T}{n} \int_0^T \sigma_t^4 dt \right]^{1/2} \quad Z \tag{11.9}$$

conditionally on the $X$ process, where $Z$ denotes a standard normal variable; Barndorff-Nielsen and Shephard (2002) looked at the special case of Brownian motion. Mykland and Zhang (2002) studied general sampling schemes. This expression naturally reduces to the simple (2.3) when volatility is not stochastic, $\sigma_t = \sigma$.

In the context where volatility is stochastic and sampling is increasingly frequent, Zhang et al. (2005) showed that ignoring market microstructure noise (of the type described in Assumptions 15, 16 and 17) leads to an even more dangerous situation than the one described above, where $\sigma$ is constant and $T \to \infty$. After suitable scaling, the realized volatility is a consistent and asymptotically normal estimator − but of the quantity $2nE[\varepsilon^2]$. In general, this quantity has nothing to do with the object of interest, the quadratic variation $\langle X, X \rangle_T$. Said differently, market microstructure noise totally swamps the variance of the price signal at the level of the realized volatility.

Indeed, if one uses all the data (say sampled every second), we showed that

$$\begin{aligned} [Y, Y]_T^{(\text{all})} \quad \overset{\mathcal{L}}{\approx} \quad & \underbrace{\langle X, X \rangle_T}_{\text{object of interest}} \quad + \quad \underbrace{2nE[\varepsilon^2]}_{\text{bias due to noise}} \\ & + \quad \underbrace{[\ \underbrace{4nE[\varepsilon^4]}_{\text{due to noise}} + \underbrace{\frac{2T}{n} \int_0^T \sigma_t^4 dt}_{\text{due to discretization}}\ ]}_{\text{total variance}}{}^{1/2} Z_{\text{total}}. \end{aligned} \tag{11.10}$$

conditionally on the $X$ process. Therefore, the realized volatility $[Y,Y]_T^{(\text{all})}$ has a positive bias whose magnitude increases linearly with the sample size $n$, as in the parametric case.

Of course, completely ignoring the noise and sampling as prescribed by $[Y,Y]_T^{(\text{all})}$ is not what the empirical literature does in practice (see e.g., Andersen et al. (2001)). There, one uses the estimator $[Y,Y]_T^{(\text{sparse})}$, constructed as above but using sparse sampling once every, say, 5 minutes. For example, if $T = 1$ NYSE day and we start with stock returns data sampled every $\delta = 1$ second, then for the full dataset the sample size is $n = T/\delta = 23,400$. But sampling sparsely once every 5 minutes means throwing out 299 out of every 300 observations, and the sample size is now only $n_{\text{sparse}} = 78$. There is a large literature devoted to this estimator: see the survey Andersen et al. (2002). As in the parametric case, if one insists upon sampling sparsely, we showed in Zhang et al. (2005) how to determine optimally the sparse sampling frequency:

$$n^*_{\text{sparse}} = \left(\frac{T}{4\,E[\varepsilon^2]^2}\int_0^T \sigma_t^4 dt\right)^{1/3}. \tag{11.11}$$

So, one could benefit from using infrequently sampled data. And yet, one of the most basic lessons of statistics is that one should not do this. Zhang et al. (2005) present a method to tackle the problem. We partition the original grid of observation times, $G = \{t_0, ..., t_n\}$ into subsamples, $G^{(k)}$, $k = 1, ..., K$ where $n/K \to \infty$ as $n \to \infty$. For example, for $G^{(1)}$ start at the first observation and take an observation every 5 minutes; for $G^{(2)}$, start at the second observation and take an observation every 5 minutes, etc. Then we average the estimators obtained on the subsamples. To the extent that there is a benefit to subsampling, this benefit can now be retained, while the variation of the estimator can be lessened by the averaging.

Subsampling and averaging together gives rise to the estimator

$$[Y,Y]_T^{(\text{avg})} = \frac{1}{K}\sum_{k=1}^{K}[Y,Y]_T^{(\text{sparse},k)} \tag{11.12}$$

constructed by averaging the estimators $[Y,Y]_T^{(\text{sparse},k)}$ obtained by sampling sparsely on each of the $K$ grids of average size $\bar{n}$. We show that:

$$\begin{aligned}[Y,Y]_T^{(\text{avg})} \;\overset{\mathcal{L}}{\approx}\;& \underbrace{\langle X,X\rangle_T}_{\text{object of interest}} + \underbrace{2\bar{n}E[\varepsilon^2]}_{\text{bias due to noise}} \\ &+ \underbrace{[\underbrace{4\frac{\bar{n}}{K}E[\varepsilon^4]}_{\text{due to noise}} + \underbrace{\frac{4T}{3\bar{n}}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}}_{\text{total variance}}]^{1/2} Z_{\text{total}}\end{aligned} \tag{11.13}$$

So, $[Y,Y]_T^{(\text{avg})}$ remains a biased estimator of the quadratic variation $\langle X,X\rangle_T$ of the true return process. But the bias $2\bar{n}E[\varepsilon^2]$ now increases with the average size of the subsamples, and $\bar{n} \leq n$. Thus, $[Y,Y]_T^{(\text{avg})}$ is a better estimator than $[Y,Y]_T^{(\text{all})}$, bust still biased.

But the lower bias can now be removed. Recall indeed that $E[\varepsilon^2]$ can be consistently approximated by

$$\widehat{E[\varepsilon^2]} = \frac{1}{2n}[Y,Y]_T^{(\text{all})}. \tag{11.14}$$

Hence a bias-adjusted estimator for $\langle X, X\rangle$ can thus be constructed as

$$\widehat{\langle X, X\rangle}_T^{(\mathrm{tsrv})} = \underbrace{[Y, Y]_T^{(\mathrm{avg})}}_{\text{slow time scale}} - \frac{\bar{n}}{n} \underbrace{[Y, Y]_T^{(\mathrm{all})}}_{\text{fast time scale}} \tag{11.15}$$

and we call this, now for obvious reasons, the Two Scales Realized Volatility estimator.

If the number of subsamples is optimally selected as $K^* = cn^{2/3}$, then TSRV has the following distribution:

$$\widehat{\langle X, X\rangle}_T^{(\mathrm{tsrv})} \overset{\mathcal{L}}{\approx} \underbrace{\langle X, X\rangle_T}_{\text{object of interest}} + \frac{1}{n^{1/6}} \underbrace{[\underbrace{\frac{8}{c^2} E[\epsilon^2]^2}_{\text{due to noise}} + \underbrace{c\frac{4T}{3}\int_0^T \sigma_t^4 dt}_{\text{due to discretization}}}_{\text{total variance}} ]^{1/2} Z_{\mathrm{total}} \tag{11.16}$$

and the constant $c$ can be set to minimize the total asymptotic variance above.

Unlike all the previously considered ones, this estimator is now correctly centered, and to the best of our knowledge is the first consistent estimator for $\langle X, X\rangle_T$ in the presence of market microstructure noise. Unbiased, but inconsistent, estimators have been studied by Zhou (1996), who, in the parametric case of constant $\sigma$, proposed a bias correcting approach based on autocovariances. The behavior of this estimator has been studied by Zumbach et al. (2002). Related studies are Hansen and Lunde (2004) and Bandi and Russell (2003); both papers consider time varying $\sigma$ in the conditionally nonrandom case, and by Oomen (2006). These estimators are unfortunately also inconsistent. Under different assumptions (a pure jump process), the corresponding estimation problem is studied by Large (2005).

A small sample refinement to $\widehat{\langle X, X\rangle}_T^{(\mathrm{tsrv})}$ can be constructed as follows

$$\widehat{\langle X, X\rangle}_T^{(\mathrm{tsrv,adj})} = \left(1 - \frac{\bar{n}}{n}\right)^{-1} \widehat{\langle X, X\rangle}_T^{(\mathrm{tsrv})}. \tag{11.17}$$

The difference from the estimator in (11.15) is of order $O_p(\bar{n}/n) = O_p(K^{-1})$, and thus the two estimators behave identically to the asymptotic order that we consider. The estimator (11.17), however, has the appeal of being unbiased to higher order.

Just like the marginal distribution of the noise is likely to be unknown, its degree of dependence is also likely to be unknown. This calls for relaxing Assumption 17, and we developed in Aït-Sahalia et al. (2005c) a serial-dependence-robust TSRV estimator under Assumption 18. In a nutshell, we continue combining two different time scales, but rather than starting with the fastest possible time scale as our starting point, one now needs to be adjust how fast the fast time scale is. We also analyze there the impact of serial dependence in the noise on the distribution of the RV estimators, $[Y, Y]_T^{(\mathrm{all})}$ and $[Y, Y]_T^{(\mathrm{sparse})}$, and on a further refinement to this approach, called Multiple Scales Realized Volatility (MSRV), which achieves further asymptotic efficiency gains over TSRV (see Zhang (2004)).

Following our work, Barndorff-Nielsen et al. (2006) have shown that our TSRV estimator can be viewed as a form of kernel based estimator. However, all kernel-based estimators are inconsistent estimators of $\langle X, X\rangle_T$ under the presence of market microstructure noise. When viewed as a kernel-based estimator, TSRV owes its consistency to its automatic selection of end effects which must be added “manually” to a kernel estimator to make it match TSRV. Optimizing over the kernel weights leads to an estimator with the same properties as

MSRV, although the optimal kernel weights will have to be found numerically, whereas the optimal weights for MSRV will be explicit (see Zhang (2004)). With optimal weights, the rate of convergence can be improved from $n^{-1/6}$ for TSRV to $n^{-1/4}$ for MSRV as the cost of the higher complexity involved in combining $O(n^{1/2})$ time scales instead of just two as in (11.15). In the fully parametric case we studied in Aït-Sahalia et al. (2005b), we showed that when $\sigma_t = \sigma$ is constant, the MLE for $\sigma^2$ converges for $T$ fixed and $\Delta \rightarrow 0$ at rate $\Delta^{1/4}/T^{1/4} = n^{-1/4}$ (see equation (31) p. 369 in Aït-Sahalia et al. (2005b)). This establishes $n^{-1/4}$ as the best possible asymptotic rate improvement over (11.16).

To conclude, in the parametric case of constant volatility, we showed that in the presence of market microstructure noise that is unaccounted for, it is optimal to sample less often than would otherwise be the case: we derive the optimal sampling frequency. A better solution, however, is to model the noise term explicitly, for example by likelihood methods, which restores the first order statistical effect that sampling as often as possible is optimal. But, more surprisingly, we also demonstrate that the likelihood correction is robust to misspecification of the assumed distribution of the noise term. In the nonparametric case of stochastic volatility, it is possible to correct for the noise by subsampling and averaging and obtain well behaved estimators that make use of all the data. These results collectively suggest that attempts to incorporate market microstructure noise when estimating continuous-time models based on high frequency data should have beneficial effects. And one final important message of these two papers: any time one has an impulse to sample sparsely, one can always do better: for example, using likelihood corrections in the parametric case or subsampling and averaging in the nonparametric case.

An alternative to the additive noise model of Assumption 15 is that described by Assumption 19, which captures the rounding that takes place in many financial markets, where prices are often quoted as a multiple of a given tick size. This form of measurement error has been studied by Jacod (1996), Delattre and Jacod (1997) and Gloter and Jacod (2000).

In the multivariate case of Assumption 7, an additional source of error is represented by the nonsynchronous trading of the different assets making up the $X$ vector. Hayashi and Yoshida (2005) study the estimation of the quadratic covariation $\left\langle X^{(i)}, X^{(j)} \right\rangle_T$, defined analogously to (11.8) under Assumption 20, without market microstructure noise, i.e., under Assumption 6. Zhang (2005) generalizes this to the situation where both Assumptions 15 and 20 hold.

# References

Aït-Sahalia, Y. (1996a): "Do Interest Rate Really Follow Continuous-Time Markov Diffusions?" Tech. rep., University of Chicago Working Paper.

——— (1996b): "Nonparametric Pricing of Interest Rate Derivative Securities," *Econometrica*, 64, 527–560.

——— (1996c): "Testing Continuous-Time Models of the Spot Interest Rate," *Review of Financial Studies*, 9, 385–426.

——— (1999): "Transition Densities for Interest Rate and Other Nonlinear Diffusions," *Journal of Finance*, 54, 1361–1395.

——— (2001): "Closed-Form Likelihood Expansions for Multivariate Diffusions," Tech. rep., Princeton University.

——— (2002a): "Empirical Option Pricing and the Markov Property," Tech. rep., Princeton University.

——— (2002b): "Maximum-Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approximation Approach," *Econometrica*, 70, 223–262.

——— (2002c): "Telling from Discrete Data Whether the Underlying Continuous-Time Model is a Diffusion," *Journal of Finance*, 57, 2075–2112.

——— (2004): "Disentangling Diffusion from Jumps," *Journal of Financial Economics*, 74, 487–528.

Aït-Sahalia, Y. and J. Duarte (2003): "Nonparametric Option Pricing Under Shape Restrictions," *Journal of Econometrics*, 116, 9–47.

Aït-Sahalia, Y. and J. Fan (2005a): "Nonparametric Transition Density-Based Tests for Jumps," Tech. rep., Princeton University.

——— (2005b): "Nonparametric Transition Density-Based Tests of the Markov Hypothesis," Tech. rep., Princeton University.

Aït-Sahalia, Y., J. Fan, and H. Peng (2005a): "Nonparametric Transition-Based Tests for Diffusions," Tech. rep., Princeton University.

Aït-Sahalia, Y., L. P. Hansen, and J. A. Scheinkman (2002): "Operator Methods for Continuous-Time Markov Processes," in *Handbook of Financial Econometrics, forthcoming*, ed. by Y. Aït-Sahalia and L. P. Hansen, Amsterdam, The Netherlands: North Holland.

Aït-Sahalia, Y. and J. Jacod (2004): "Fisher's Information for Discretely Sampled Lévy Processes," Tech. rep., Princeton University and Université de Paris 6.

——— (2006): "Volatility Estimators for Discretely Sampled Lévy Processes," *Annals of Statistics, forthcoming.*

Aït-Sahalia, Y. and R. Kimmel (2002): "Estimating Affine Multifactor Term Structure Models Using Closed-Form Likelihood Expansions," Tech. rep., Princeton University.

——— (2004): "Maximum Likelihood Estimation of Stochastic Volatility Models," *Journal of Financial Economics, forthcoming.*

Aït-Sahalia, Y. and A. Lo (1998): "Nonparametric Estimation of State-Price-Densities Implicit in Financial Asset Prices," *Journal of Finance*, 53, 499–547.

——— (2000): "Nonparametric Risk Management and Implied Risk Aversion," *Journal of Econometrics*, 94, 9–51.

Aït-Sahalia, Y. and P. A. Mykland (2003): “The Effects of Random and Discrete Sampling When Estimating Continuous-Time Diffusions,” *Econometrica*, 71, 483–549.

——— (2004): “Estimators of Diffusions with Randomly Spaced Discrete Observations: A General Theory,” *The Annals of Statistics*, 32, 2186–2222.

——— (2005): “An Analysis of Hansen-Scheinkman Moment Estimators for Discretely and Randomly Sampled Diffusions,” Tech. rep., Princeton University.

Aït-Sahalia, Y., P. A. Mykland, and L. Zhang (2005b): “How Often to Sample a Continuous-Time Process in the Presence of Market Microstructure Noise,” *Review of Financial Studies*, 18, 351–416.

——— (2005c): “Ultra High Frequency Volatility Estimation with Dependent Microstructure Noise,” Tech. rep., Princeton University.

Aït-Sahalia, Y. and J. Park (2005): “Specification Testing for Nonstationary Diffusions,” Tech. rep., Princeton University and Rice University.

Aït-Sahalia, Y., Y. Wang, and F. Yared (2001): “Do Option Markets Correctly Price the Probabilities of Movement of the Underlying Asset?” *Journal of Econometrics*, 102, 67–110.

Aït-Sahalia, Y. and J. Yu (2005): “Saddlepoint Approximations for Continuous-Time Markov Processes,” *Journal of Econometrics*, forthcoming.

Akonom, J. (1993): “Comportement asymptotique du temps d’occupation du processus des sommes partielles,” *Annales de l’Institut Henri Poincaré*, 29, 57–81.

Andersen, T., T. Bollerslev, and F. X. Diebold (2003): “Some Like it Smooth, and Some Like it Rough,” Tech. rep., Northwestern University.

Andersen, T. G., T. Bollerslev, and F. X. Diebold (2002): “Parametric and Nonparametric Measurements of Volatility,” in *Handbook of Financial Econometrics, forthcoming*, ed. by Y. Aït-Sahalia and L. P. Hansen, Amsterdam, The Netherlands: North Holland.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and P. Labys (2001): “The Distribution of Exchange Rate Realized Volatility,” *Journal of the American Statistical Association*, 96, 42–55.

Bakshi, G. S. and N. Yu (2005): “A Refinement to Aït-Sahalia’s (2002) "Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-form Approximation Approach",” *Journal of Business*, 78, 2037–2052.

Bandi, F. M. and P. C. B. Phillips (2002): “Nonstationary Continuous-Time Processes,” in *Handbook of Financial Econometrics, forthcoming*, ed. by Y. Aït-Sahalia and L. P. Hansen, Amsterdam, The Netherlands: North Holland.

——— (2003): “Fully Nonparametric Estimation of Scalar Diffusion Models,” *Econometrica*, 71, 241–283.

Bandi, F. M. and J. R. Russell (2003): “Microstructure Noise, Realized Volatility and Optimal Sampling,” Tech. rep., University of Chicago Graduate School of Business.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2006): “Regular and Modified Kernel-Based Estimators of Integrated Variance: The Case with Independent Noise,” Tech. rep., Department of Mathematical Sciences, University of Aarhus.

Barndorff-Nielsen, O. E. and N. Shephard (2002): “Econometric Analysis of Realized Volatility and Its Use in Estimating Stochastic Volatility Models,” *Journal of the Royal Statistical Society, B*, 64, 253–280.

——— (2003): “Power Variation with Stochastic Volatility and Jumps,” Tech. rep., University of Aarhus.

Bibby, B. M., M. Jacobsen, and M. Sørensen (2002): "Estimating Functions for Discretely Sampled Diffusion-Type Models," in *Handbook of Financial Econometrics*, ed. by Y. Aït-Sahalia and L. P. Hansen, Amsterdam, The Netherlands: North Holland.

Bibby, B. M. and M. S. Sørensen (1995): "Estimation Functions for Discretely Observed Diffusion Processes," *Bernoulli*, 1, 17–39.

Billingsley, P. (1961): *Statistical Inference for Markov Processes*, Chicago: The University of Chicago Press.

Black, F. and M. Scholes (1973): "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, 81, 637–654.

Breeden, D. and R. H. Litzenberger (1978): "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business*, 51, 621–651.

Carr, P. and L. Wu (2003): "What Type of Process Underlies Options? A Simple Robust Test," *Journal of Finance*, 58, 2581–2610.

Chapman, D. A. and N. Pearson (2000): "Is the Short Rate Drift Actually Nonlinear?" *Journal of Finance*, 55, 355–388.

Chen, S. X. and J. Gao (2004a): "An Adaptive Empirical Likelihood Test For Time Series Models," Tech. rep., Iowa State University.

——— (2004b): "On the Use of the Kernel Method for Specification Tests of Diffusion Models," Tech. rep., Iowa State University.

Cheridito, P., D. Filipović, and R. L. Kimmel (2005): "Market Price of Risk Specifications for Affine Models: Theory and Evidence," *Journal of Financial Economics*, forthcoming.

Corradi, V. and N. R. Swanson (2005): "A Bootstrap Specification Test for Diffusion Processes," *Journal of Econometrics, forthcoming.*

Cox, J. C., J. E. Ingersoll, and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, 385–408.

Dai, Q. and K. J. Singleton (2000): "Specification Analysis of Affine Term Structure Models," *Journal of Finance*, 55, 1943–1978.

Daniels, H. (1954): "Saddlepoint Approximations in Statistics," *Annals of Mathematical Statistics*, 25, 631–650.

Delattre, S. and J. Jacod (1997): "A Central Limit Theorem for Normalized Functions of the Increments of a Diffusion Process, in the Presence of Round-Off Errors," *Bernoulli*, 3, 1–28.

DiPietro, M. (2001): "Bayesian Inference for Discretely Sampled Diffusion Processes with Financial Applications," Ph.D. thesis, Department of Statistics, Carnegie-Mellon University.

Duarte, J. (2004): "Evaluating an Alternative Risk Preference in Affine Term Structure Models," *Review of Financial Studies*, 17, 379–404.

Duffee, G. R. (2002): "Term Premia and Interest Rate Forecasts in Affine Models," *Journal of Finance*, 57, 405–443.

Duffie, D. and R. Kan (1996): "A Yield-Factor Model of Interest Rates," *Mathematical Finance*, 6, 379–406.

Efron, B. and R. Tibshirani (1996): "Using specially designed exponential families for density estimation," *The Annals of Statistics*, 24, 2431–2461.

Egorov, A. V., H. Li, and Y. Xu (2003): "Maximum Likelihood Estimation of Time Inhomogeneous Diffusions," *Journal of Econometrics*, 114, 107–139.

Engle, R. F. (2000): "The Econometrics of Ultra-High Frequency Data," *Econometrica*, 68, 1–22.

Engle, R. F. and J. R. Russell (1998): "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data," *Econometrica*, 66, 1127–1162.

Fan, J. (1992): "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, 87, 998–1004.

Fan, J. and Q. Yao (1998): "Efficient estimation of conditional variance functions in stochastic regression," *Biometrika*, 85, 645–660.

Fan, J., Q. Yao, and H. Tong (1996): "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*, 83, 189–206.

Florens-Zmirou, D. (1989): "Approximate Discrete-Time Schemes for Statistics of Diffusion Processes," *Statistics*, 20, 547–557.

Frydman, H. and B. Singer (1979): "Total Positivity and the Embedding Problem for Markov Chains," *Mathematical Proceedings of the Cambridge Philosophical Society*, 86, 339–344.

Gallant, A. R. and G. T. Tauchen (1996): "Which Moments to Match?" *Econometric Theory*, 12, 657–681.

——— (2002): "Simulated Score Methods and Indirect Inference for Continuous-time Models," in *Handbook of Financial Econometrics, forthcoming*, ed. by Y. Aït-Sahalia and L. P. Hansen, Amsterdam, The Netherlands: North Holland.

Genon-Catalot, V., T. Jeantheau, and C. Larédo (1999): "Parameter Estimation for Discretely Observed Stochastic Volatility Models," *Bernoulli*, 5, 855–872.

Glad, I. K. (1998): "Parametrically guided non-parametric regression," *Scandinavian Journal of Statistics*, 25, 649–668.

Gloter, A. and J. Jacod (2000): "Diffusions with Measurement Errors: I - Local Asymptotic Normality and II - Optimal Estimators," Tech. rep., Université de Paris VI.

Godambe, V. P. (1960): "An Optimum Property of Regular Maximum Likelihood Estimation," *Annals of Mathematical Statistics*, 31, 1208–1211.

Gouriéroux, C., A. Monfort, and E. Renault (1993): "Indirect Inference," *Journal of Applied Econometrics*, 8, S85–S118.

Hall, P. and C. C. Heyde (1980): *Martingale Limit Theory and Its Application*, Boston: Academic Press.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P. and T. J. Sargent (1983): "The Dimensionality of the Aliasing Problem in Models with Rational Spectral Densities," *Econometrica*, 51, 377–387.

Hansen, L. P. and J. A. Scheinkman (1995): "Back to the Future: Generating Moment Implications for Continuous-Time Markov Processes," *Econometrica*, 63, 767–804.

Hansen, P. R. and A. Lunde (2004): "Realized Variance and IID Market Microstructure Noise," Tech. rep., Stanford University, Department of Economics.

——— (2006): "Realized Variance and Market Microstructure Noise," *Journal of Business and Economic Statistics, forthcoming.*

Harrison, M. and D. Kreps (1979): "Martingales and Arbitrage in Multiperiod Securities Markets," *Journal of Economic Theory*, 20, 381–408.

Hayashi, T. and N. Yoshida (2005): "On Covariance Estimation of Nonsynchronously Observed Diffusion Processes," Tech. rep., Columbia University.

Heston, S. (1993): "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bonds and Currency Options," *Review of Financial Studies*, 6, 327–343.

Heyde, C. C. (1997): *Quasi-Likelihood and Its Application*, New York: Springer-Verlag.

Hjort, N. L. and I. K. Glad (1995): "Nonparametric density estimation with a parametric start," *The Annals of Statistics*, 23, 882–904.

Hjort, N. L. and M. Jones (1996): "Locally parametric nonparametric density estimation," *The Annals of Statistics*, 24, 1619–1647.

Hong, Y. and H. Li (2005): "Nonparametric Specification Testing for Continuous-Time Models with Applications to Term Structure of Interest Rates," *Review of Financial Studies*, 18, 37–84.

Huang, X. and G. Tauchen (2006): "The relative contribution of jumps to total price variance," *Journal of Financial Econometrics*, 4, 456–499.

Hull, J. and A. White (1987): "The Pricing of Options on Assets with Stochastic Volatilities," *Journal of Finance*, 42, 281–300.

Hurn, A. S., J. Jeisman, and K. Lindsay (2005): "Seeing the Wood for the Trees: A Critical Evaluation of Methods to Estimate the Parameters of Stochastic Differential Equations," Tech. rep., School of Economics and Finance, Queensland University of Technology.

Jacod, J. (1994): "Limit of Random Measures Associated with the Increments of a Brownian Semimartingale," Tech. rep., Université de Paris VI.

——— (1996): "La Variation Quadratique du Brownien en Présence d'Erreurs d'Arrondi," *Astérisque*, 236, 155–162.

Jacod, J. and P. Protter (1998): "Asymptotic Error Distributions for the Euler Method for Stochastic Differential Equations," *Annals of Probability*, 26, 267–307.

Jeganathan, P. (1995): "Some Aspects of Asymptotic Theory with Applications to Time Series Models," *Econometric Theory*, 11, 818–887.

Jensen, B. and R. Poulsen (2002): "Transition Densities of Diffusion Processes: Numerical Comparison of Approximation Techniques," *Journal of Derivatives*, 9, 1–15.

Jiang, G. J. and J. Knight (1997): "A Nonparametric Approach to the Estimation of Diffusion Processes - With an Application to a Short-Term Interest Rate Model," *Econometric Theory*, 13, 615–645.

Karlin, S. and J. McGregor (1959a): "Coincidence Probabilities," *Pacific journal of Mathematics*, 9, 1141–1164.

——— (1959b): "Coincidence Properties of Birth-and-Death Processes," *Pacific journal of Mathematics*, 9, 1109–1140.

Kent, J. (1978): "Time Reversible Diffusions," *Advanced Applied Probability*, 10, 819–835.

Kessler, M. and M. Sørensen (1999): "Estimating Equations Based on Eigenfunctions for a Discretely Observed Diffusion," *Bernoulli*, 5, 299–314.

Kristensen, D. (2004): "Estimation in Two Classes of Semiparametric Diffusion Models," Tech. rep., University of Wisconsin-Madison.

Large, J. (2005): "Quadratic Variation When Quotes Change One Tick at a Time," Tech. rep., Oxford University.

Lewis, A. L. (2000): *Option Valuation under Stochastic Volatility*, Newport Beach, CA: Finance Press.

Lo, A. W. (1988): "Maximum Likelihood Estimation of Generalized Itô Processes with Discretely Sampled Data," *Econometric Theory*, 4, 231–247.

Loader, C. R. (1996): "Local likelihood density estimation," *The Annals of Statistics*, 24, 1602–1618.

Mykland, P. A. and L. Zhang (2002): "ANOVA for Diffusions," *The Annals of Statistics, forthcoming*, –, –.

Oomen, R. C. (2006): "Properties of Realized Variance under Alternative Sampling Schemes," *Journal of Business and Economic Statistics*, 24, 219–237.

Pedersen, A. R. (1995): "A New Approach to Maximum-Likelihood Estimation for Stochastic Differential Equations Based on Discrete Observations," *Scandinavian journal of Statistics*, 22, 55–71.

Philips, P. C. B. (1973): "The Problem of Identification in Finite Parameter Continuous Time Models," *Journal of Econometrics*, 1, 351–362.

Phillips, P. C. and J. Yu (2005): "A Two-Stage Realized Volatility Approach to the Estimation of Diffusion Processes from Discrete Observations," Tech. rep., Singapore Management University.

Press, H. and J. W. Turkey (1956): *Power spectral methods of analysis and their application to problems in airplane dynamics*, Bell Telephone System Monograph 2606.

Pritzker, M. (1998): "Nonparametric Density Estimation and Tests of Continuous Time Interest Rate Models," *Review of Financial Studies*, 11, 449–487.

Renault, E. and B. J. Werker (2003): "Stochastic Volatility Models with Transaction Time Risk," Tech. rep., Tilburg University.

Romano, M. and N. Touzi (1997): "Contingent Claims and Market Completeness in a Stochastic Volatility Model," *Mathematical Finance*, 7, 399–412.

Schaumburg, E. (2001): "Maximum Likelihood Estimation of Jump Processes with Applications to Finance," Ph.D. thesis, Princeton University.

Smith, A. A. (1993): "Estimating Nonlinear Time Series Models Using Simulated Vector Autoregressions," *Journal of Applied Econometrics*, 8, S63–S84.

Sørensen, H. (2001): "Discretely Observed Diffusions: Approximation of the Continuous-Time Score Function," *Scandinavian Journal of Statistics*, 28, 113–121.

Stanton, R. (1997): "A Nonparametric Model of Term Structure Dynamics and the Market Price of Interest Rate Risk," *Journal of Finance*, 52, 1973–2002.

Stramer, O. and J. Yan (2005): "On Simulated Likelihood of Discretely Observed Diffusion Processes and Comparison to Closed-Form Approximation," Tech. rep., University of Iowa.

Thompson, S. (2004): "Identifying Term Structure Volatility from the LIBOR-Swap Curve," Tech. rep., Harvard University.

Vasicek, O. (1977): "An Equilibrium Characterization of the Term Structure," *Journal of Financial Economics*, 5, 177–188.

Wong, E. (1964): "The Construction of a Class of Stationary Markoff Processes," in *Sixteenth Symposium in Applied Mathematics - Stochastic Processes in Mathematical Physics and Engineering*, ed. by R. Bellman, Providence, RI: American Mathematical Society, 264 – 276.

Yu, J. (2003): "Closed-Form Likelihood Estimation of Jump-Diffusions with an Application to the Realignment Risk Premium of the Chinese Yuan," Ph.D. thesis, Princeton University.

Zhang, L. (2004): "Efficient Estimation of Stochastic Volatility Using Noisy Observations: A Multi-Scale Approach," Tech. rep., Carnegie-Mellon University.

——— (2005): "Estimating Covariation: Epps Effect and Microstructure Noise," Tech. rep., Carnegie-Mellon University.

Zhang, L., P. A. Mykland, and Y. Aït-Sahalia (2005): "A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High-Frequency Data," *Journal of the American Statistical Association*, 100, 1394–1411.

Zhou, B. (1996): "High-Frequency Data and Volatility in Foreign-Exchange Rates," *Journal of Business & Economic Statistics*, 14, 45–52.

Zumbach, G., F. Corsi, and A. Trapletti (2002): "Efficient Estimation of Volatility using High Frequency Data," Tech. rep., Olsen & Associates.