Approximate Maximum Likelihood Estimation for 1D Diffusions Observed on a Fine Grid

Kevin W. Lu

University of Washington, USA

Phillip J. Paine

BCAA Insurance, Vancouver, British Columbia, Canada

Simon P. Preston

University of Nottingham, UK

Andrew T. A. Wood

Australian National University, Australia

Abstract

We consider a one-dimensional stochastic differential equation that is observed on a fine grid of equally-spaced time points. A novel approach for approximating the transition density of the stochastic differential equation is presented, which is based on an Itô-Taylor expansion of the sample path, combined with an application of the so-called ϵ-expansion; the resulting approximation is highly economical with respect to the number of terms needed to achieve a given level of accuracy in a high-frequency sampling framework. This method of density approximation leads to a closed-form approximate likelihood function from which an approximate maximum likelihood estimator may be calculated numerically. A detailed theoretical analysis of the proposed estimator is provided and it is shown that it compares favourably to the Gaussian likelihood-based estimator and does an excellent job of approximating the exact, but usually intractable, maximum likelihood estimator. Further, numerical simulations indicate that the exact and our approximate maximum likelihood estimator tend to be close, and the latter performs very well relative to other approximate methods in the literature in terms of speed, accuracy and ease of implementation.

Keywords: *ϵ-expansion; Itô-Taylor expansion; mixing; stochastic differential equation; transition density approximation.*

1 Introduction

1.1 Preliminary remarks

In diverse fields, including finance (e.g. Shreve (2004)), biology (e.g. Wilkinson (2006)) and physics and chemistry (e.g. van Kampen (1992) and Gardiner (2003)), one can model certain phenomena using a continuous-time time-homogeneous diffusion process. Mathematically, this takes the form of a stochastic differential equation given by

$$\mathrm{d}X_t = \mu(X_t, \alpha)\mathrm{d}t + \sigma(X_t, \beta)\mathrm{d}B_t, \tag{1.1}$$

where $X_t \equiv X(t)$ is the state variable, μ is the drift function, σ the diffusion function, $\theta = (\alpha^\top, \beta^\top)^\top$ is an unknown vector of parameters, and $B_t \equiv B(t)$ is a standard Brownian motion.

The question that motivates this work is how to perform statistical inference for the unknown parameter vector θ. It is assumed throughout the paper that the observed data consists of observations $x(t_0), x(t_1), \ldots, x(t_n)$ of the state variable X_t at equally-spaced discrete time points $t_0 < \cdots < t_n$, where $t_i - t_{i-1} = \Delta$, and Δ is small. For simplicity, we shall always condition on the initial observation, $x(t_0)$. The asymptotic regime under consideration is the following: $n \to \infty$, $\Delta \to 0$ and $n\Delta \to \infty$. This corresponds to high-frequency sampling, a type of sampling of central importance in finance in particular.

This paper makes the following contributions. An approximation to the transition density is derived using the so-called ϵ-expansion (Cox and Reid (1987); Barndorff-Nielsen and Cox (1989)), applied to the terms in an Itô-Taylor approximation (Kloeden and Platen (1992)). See §3 and, in particular, Theorem 3.1. The ϵ-expansion is, in a certain sense, a generalisation of the classical Edgeworth expansion; see Remark 3.1 below.

We provide theoretical results which describe the asymptotic properties of the approximate maximum likelihood estimator, written $\hat{\theta}^{(\epsilon)}$, of the parameter vector $\theta = (\alpha^\top, \beta^\top)^\top$ in (1.1), based on the approximate transition density mentioned above. In Theorem 4.2, it is shown that, in a theoretical sense, the discrepancy between $\hat{\theta}^{(\epsilon)}$ and the exact maximum likelihood estimator, $\hat{\theta}_{MLE}$, is never of larger order than the discrepancy between $\hat{\theta}_{MLE}$ and the estimator based on a Gaussian likelihood considered by Kessler (1997); and in certain well-defined circumstances, essentially when Δ does not go to zero too quickly as $n \to \infty$, the former discrepancy will be of strictly smaller order than the latter discrepancy. From a computational point of view the estimator $\hat{\theta}^{(\epsilon)}$ proposed here performs extremely well in terms of speed, accuracy and ease of implementation, and is competitive with the best of the alternative methods; see the numerical results in §5.

In Supplementary Material we have provided additional numerical results, proofs of all theoretical results in the paper and Matlab programs for performing all the calculations in the paper.

1.2 Review of related work

Slightly abusing notation, define $f(x_1, \ldots, x_n | x_0; \theta)$ to be the joint density of $x_1 \equiv x_{t_1}, \ldots, x_n \equiv x_{t_n}$ conditional on $x_0 \equiv x_{t_0}$, and let $p(x_i | x_{i-1}; \theta)$ denote the transition density of X_t between times t_{i-1} and t_i. Due to the Markov structure of (1.1), the log-likelihood for θ based on the observation vector $x = (x_0, \ldots, x_n)^\top$, conditional on $X_0 = x_0$, simplifies to

$$\ell(\theta; x) = \log f(x_1, \ldots, x_n | x_0; \theta) = \sum_{i=1}^{n} \log p(x_i | x_{i-1}; \theta). \tag{1.2}$$

However, the practical challenge posed by the log-likelihood (1.2) as a basis for inference for θ is that the transition density $p(x_i | x_{i-1}; \theta)$ is only known in closed form for very few models of the form (1.1).

1.2.1 The closed-form likelihood approach.

We now mention the work of other authors who have considered the same estimation problem, i.e. approximation and maximisation of the log-likelihood (1.2). Dacunha-Castelle and Florens-Zmirou (1986) prove under certain conditions that $\hat{\theta}_{MLE}$ in model (1.1) is consistent and asymptotically normal under certain conditions and they give an expression for the information matrix under conditions which ensure that an invariant distribution exists. These authors make use of an exact representation of the transition density obtained via Girsanov's theorem, as we do; see (2.5) below. This representation is not convenient for numerical calculation but it does play an important theoretical role. Some other important contributions on essentially the same problem were give in by Prakasa Rao (1983, 1099), Yoshida (1992), Shoji and Ozaki (1998) and Gobet (2002); and Kessler (1997), who presented an approach based on a Gaussian approximation to the transition density, in which higher-order approximations are used for the incremental means and variances. The recent paper by Li (2013) makes use of a Malliavin-type approach, based on the work of Watanabe (1987), for obtaining an expansion of the transition density. See also Li (2014) and Li et al. (2016) who provide efficient methods for calculation of the expansions developed in Li (2013), using a somewhat different expansion to that considered here. However, in Li (2013) and the subsequent papers, there is no discussion of joint asymptotics as $\Delta \to 0$ and $n \to \infty$, nor are any asymptotic normality results for approximate maximum likelihood estimators stated or proved. Nevertheless, the approach of Li (2013) for transition density approximation is a powerful one that deserves further study.

The Hermite polynomial method of Ait-Sahalia (1999, 2002) is concerned with the same problem, that of estimating the transition density of the stochastic differential equation (1.1) and, arguably, this approach provides the best numerical performance in the literature to date of the so-called closed-form approximations. In the Hermite polynomial method the transition density is approximated directly by a Hermite polynomial expansion. This is rather different to the perspective taken here, where the transition density approximation is derived through a combination of the Itô-Taylor expansion and the ϵ-expansion. A further factor which complicates the comparison between our expansion and Ait-Sahalia's expansion is that two distinct asymptotic regimes are relevant in Ait-Sahalia's approach: (i) $\Delta \to 0$, where Δ is the sampling interval; and (ii) allowing $J \to \infty$, where J is the number of terms in the expansion; see Chang and Chen (2011) for extensive discussion. Only asymptotic regime (i) is relevant to the present paper and we focus on this case below.

A key point is that our expansion and Ait-Sahalia's expansion for the transition density are essentially equivalent, in the sense that, when the comparison is done carefully, the two methods give the same coefficient for a term of given order, though this may not be immediately obvious due to the differences in the derivations. An important implication is that differences in the two types of the expansion are due largely to which higher-order terms are retained and which are omitted. We do not analyse Ait-Sahalia's estimators within our asymptotic framework, as a comprehensive analysis under various asymptotic scenarios has already been undertaken by Chang and Chen (2011), though we do compare the resulting approximations numerically in §5; see also Remark 4.3. However, an important consequence of this equivalence is that, in the numerical comparisons given below in Section 5 and in part A of the Supplementary Material, the approximation proposed in this paper, referred to as *epsilon*, is a representative of approximations which are correct up to and including all terms of order Δ^2 (see (3.15) while the Ait-Sahalia approximation *AS* is a representative of approximations which are correct up to and including all terms of order Δ^3). So to a considerable extent the comparison is about comparing an expansion correct to order Δ^2 to one correct to order Δ^3. This point is discussed further in Remark 3.6 and in Section B6 of the Supplementary Material.

The closed-form approximation approach inspired by Ait-Sahalia (1999, 2002) continues to be an active area of research. Recent work includes Yang et al. (2019) who, as in this paper, use the Itô-Taylor expansion approach but provide a completely different justification, based on certain estimates from partial differential equation theory. One difference in outcome with this paper is that they prove consistency of parameter estimates but not asymptotic normality whereas our results include both; but, on the other hand, they consider multidimensional diffusions whereas we just consider the univariate case. An important new direction for the closed-form approximation approach is the extension from continuous diffusions to jump diffusions; see Li and Chen (2016) and Wang and Yang (2021). Further important recent work includes Ait-Sahalia et al. (2021a), who consider parametric and nonparametric estimation in implied stochastic volatility models; and Ait-Sahalia et al. (2021b), where the extension to implied volatility surfaces for stochastic volatility models with jumps is considered.

1.2.2 Saddlepoint approximation.

Preston and Wood (2012) developed an alternative method based on obtaining the moment generating function of the first few terms of the Itô-Taylor expansion, and then using the saddlepoint method to approximate the transition density and hence the likelihood. The resulting approximation of the maximum likelihood estimator performs quite well numerically, though generally not as well as the approximation proposed in this paper, as indicated by results not shown here. The Preston and Wood (2012) approximation also has a number of further drawbacks relative to the new approximation: its calculation is relatively slow and cumbersome when compared with the transition density approxmation proposed here; it is only possible to include one non-Gaussian term if existence of the moment generating function off the imaginary axis is to be retained; and its properties as an approximation are less amenable to theoretical study. In contrast, the new approximation is very fast to compute and there is no theoretical or practical restriction on the number of non-Gaussian terms that can be included.

1.2.3 Simulation-based methods.

We also briefly mention simulation-based approaches to inference in model (1.1). In Beskos et al. (2006), and in subsequent work by these and other authors, an exact simulation-based approach to Bayesian inference is developed, while simulation-based approaches to frequentist likelihood-based inference were developed in Pedersen (1995), Durham and Gallant (2002) and Beskos et al. (2009). Two key differences between Beskos et al. (2006), Pedersen (1995), Durham and Gallant (2002) and Beskos (2009) on the one hand, and the present paper and papers such as Dacunha-Castelle and Florens-Zmirou (1986), Prakasa Rao (1983, 1986), Yoshida (1992), Kessler (1997), Ait-Sahalia (1999, 2002), Gobet (2002), Shoji and Ozaki (1998) on the other, are the following: in contrast to the latter set of papers, the former papers do not require Δ to be small; and, in contrast to the former set of papers, the latter set of papers typically provide closed-form approximations to the likelihood.

One further point: fast, accurate and easy-to-implement closed-form small-Δ approximations to the log-likelihood of the type developed here can be used to advantage in, for example, the importance-sampling approach of Durham and Gallant (2002).

1.2.4 Outline of paper.

The contents of the paper are as follows. In §2 we review the exact representation of the transition density. In §3 we introduce the ϵ-expansion and the Itô-Taylor expansion and show how they can be used to construct an explicit transition density approximation whose theoretical properties are summarised in Theorem 3.1. In §4 the key result, Theorem 4.2, compares the proposed approximation with a variant of the Kessler approximation and identifies the asymptotic regimes in which the former has superior performance. Extensive numerical results are described in §5 and in part A of the Supplementary Material. Auxiliary results are stated in §6 and all theoretical results are proved in part B of the Supplementary Material and the Matlab programs used for all the computations in the paper are given in part C of the Supplementary Material.

2 Exact representation of the transition density

It is useful to transform the original diffusion process $X_\Delta \equiv X(\Delta)$ in (1.1) to a related process $Y_\Delta \equiv Y(\Delta)$ that has a unit diffusion coefficient, i.e. $\sigma \equiv 1$ for the Y_Δ-process. The particular motivation for employing this transformation in the present context is that it leads to a more efficient expansion, in the sense that fewer terms are needed to attain a given order of accuracy. For fixed parameter vector β, transformation from x_Δ to y_Δ for given $X_0 = x_0$ is defined by

$$y_\Delta - y_0 = \int_{x_0}^{x_\Delta} \frac{1}{\sigma(u,\beta)} \mathrm{d}u = \gamma(x_\Delta, \beta) - \gamma(x_0, \beta). \tag{2.1}$$

If we apply Itô's lemma to this integral then we find the the transformed SDE has unit diffusion coefficient, i.e.

$$\mathrm{d}Y_t = \mu_Y(y_t;\theta)\mathrm{d}t + \mathrm{d}B_t, \tag{2.2}$$

where

$$\mu_Y(y_t,\theta) = \mu(\gamma^{-1}(y_t,\beta);\alpha)\frac{1}{\sigma(\gamma^{-1}(y_t,\beta),\beta)} - \frac{1}{2}\frac{\partial}{\partial y}\sigma(\gamma^{-1}(y_t,\beta),\beta). \tag{2.3}$$

In (2.3), $\gamma^{-1}(.,\beta)$ is the inverse function of $\gamma(.,\beta)$ with β fixed. This transformation improves the accuracy of simulation methods; and further, even if a numerical approximation to γ is required, the improvement in accuracy is usually worth the extra computing time; see Stramer and Yan (2007) for a full treatment of this transformation. As we shall see shortly, this transformation also enables us to write down high-order exapansions in more compact form.

Using the transformation of variables given by (2.1) we can write the transition density of the X-process in terms of the transition density of the Y-process:

$$p_X(x_\Delta|x_0;\theta) = \frac{1}{\sigma(x_\Delta,\beta)} p_Y(\gamma(x_\Delta,\beta)|\gamma(x_0,\beta);\theta), \tag{2.4}$$

where we have used the fact that $dy_\Delta/dx_\Delta = 1/\sigma(x_\Delta,\beta)$, which is a consequence of (2.1).

Dacunha-Castelle and Florens-Zmirou (1986, lemmas 1 and 2) use Girsanov's theorem (see e.g. Øksendal, 1998) to show that, under mild conditions, the transition density of a stochatic differential equation with unit diffusion coefficient, i.e. $\sigma(X_t,\beta) \equiv 1$ in (1.1), is

$$\begin{aligned} p_Y(y_\Delta|y_0;\theta) = {} & \frac{1}{(2\pi\Delta)^{1/2}} \exp\left\{ -\frac{(y_\Delta - y_0)^2}{2\Delta} + \int_{y_0}^{y_\Delta} \mu_Y(u,\theta)\mathrm{d}u \right\} \\ & \times E\left[\exp\left(\Delta \int_0^1 g((1-u)y_0 + uy_\Delta + \Delta^{1/2}B_u)du \right) \right], \end{aligned} \tag{2.5}$$

where μ_Y, defined in (2.3), is the drift of the unit diffusion Y_t, the expectation in (2.5) is with respect to the standard Brownian bridge $(B_u)_{0\leq u\leq 1}$ in (2.5) with $B_0 = 0 = B_1$, and

$$g(y) \equiv g(y,\theta) = -\frac{1}{2}\left(\mu_Y(y,\theta)^2 + \partial_y \mu_Y(y,\theta)\right), \tag{2.6}$$

where ∂_y indicates the partial derivative with respect to y.

3 Small-Δ expansion of the transition density

Our approach is based on deriving an asymptotic approximation for the transition density for small Δ. There are two quite different but mathematically equivalent ways of deriving such expansions: (i) to expand the transition density (2.5) for Y_t and then use the change of variable formula (2.4) to obtain the transition density for the X_t process; and (ii) to derive an Itô-Taylor expansion (Kloeden and Platen, 1992) of $Y_\Delta - y_0$ and then use the so-called ϵ-expansion (Cox and Reid, 1987; Barndorff-Nielsen and Cox, 1989) to derive an approximation to the Y_t process, and then use (2.4) to obtain the approximation of the transition density of the X_t process. It turns out that approach (ii) is more convenient for the purpose of actually calculating the terms in the expansion, while approach (i) provides a more convenient method for justifying the expansion rigorously.

3.1 The ϵ-expansion

A direct and convenient method for approximating a finite sum of leading terms in an Itô-Taylor expansion of the form (3.12) is to use the ϵ-expansion, defined in this subsection. For some fixed m, consider a random variable $X \equiv X^{(n)}$ of the form

$$X = X_0 + \sum_{a=1}^{m} \epsilon_a c_a X_a, \tag{3.1}$$

where the $\epsilon_a \equiv \epsilon_a^{(n)}$ are non-zero numbers which converge to 0; $c_1, \ldots, c_m$ are constants; and $X_0, \ldots, X_m$ are fixed, continuous random variables with arbitrary dependence structure. We denote the density function of X by $g(x)$, the density function of X_0 by $g_0(x)$, and it is assumed that the joint density of $X_0, \ldots, X_m$ is sufficiently smooth to allow the interchange of differentiation and integration; see Cox and Reid (1987) and also Barndorff-Nielsen and Cox (1989) for details. Define, for $a, b = 1, \ldots, m$, $\kappa_a(x) = E[X_a | X_0 = x]$ and $\kappa_{ab}(x) = E[X_a X_b | X_0 = x]$. Also, define

$$q_j(x) = \frac{1}{g_0(x)} \partial_x^j g_0(x), \tag{3.2}$$

where ∂_x denotes the partial differentiation operator with respect to x, so that $q_0(x) = 1$, $q_1(x) = \{\partial_x g_0(x)\}/g_0(x)$ and so on. Then the ϵ-expansion up to and including all terms linear and quadratic in the $\epsilon_1, \ldots, \epsilon_m$ is given by

$$\hat{g}^{\epsilon}_{[2]}(x) = g_0(x) \Bigg[1 - \sum_{a=1}^{m} \epsilon_a c_a \left\{ q_1(x)\kappa_a(x) + \partial_x \kappa_a(x) \right\} \\ + \frac{1}{2} \sum_{a,b=1}^{m} \epsilon_a \epsilon_b c_a c_b \left\{ q_2(x)\kappa_{ab}(x) + 2q_1(x)\partial_x \kappa_{ab}(x) + \partial_x^2 \kappa_{ab}(x) \right\} \Bigg]. \tag{3.3}$$

Under suitable conditions, when $\epsilon_a \to 0$ for $a = 1, \ldots, m$, we have

$$|\hat{g}_{[2]}(x) - g(x)| = O\left(\sum_{a=1}^{m} \epsilon_a^3 \right),$$

and the stronger result

$$\sup_{x \in \mathbf{R}} |\hat{g}_{[2]}(x) - g_0(x)| = O(\epsilon_m^3)$$

often holds too; see Cox and Reid (1987) for the types of conditions needed.

The particular case of relevance here is where X_0 is Gaussian. Suppose $X_0 \sim \mathcal{N}(\xi, 1)$ and write $\bar{x} = x - \xi$. Then in this case (3.3) has the form

$$\hat{g}_{[2]}(x) = g_0(\bar{x}) \Bigg[1 + \sum_{a=1}^{m} \epsilon_a c_a \left\{ H_1(\bar{x})\kappa_a(\bar{x}) - \partial_x \kappa_a(\bar{x}) \right\} \\ \frac{1}{2} \sum_{a,b=1}^{m} \epsilon_a \epsilon_b c_a c_b \left\{ H_2(\bar{x})\kappa_{ab}(\bar{x}) - 2H_1(\bar{x})\partial_x \kappa_{ab}(\bar{x}) + \partial_x^2 \kappa_{ab}(\bar{x}) \right\} \Bigg], \tag{3.4}$$

where $H_j(x)$ is the Hermite polynomial of degree j, defined by $(-1)^j \{\partial_x^j \phi(x)\}/\phi(x)$ where ϕ is the standard Gaussian density $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$. Consequently, it follows from (3.2) that if $g_0(x) = \phi(x)$, then $q_j(x) = (-1)^j H_j(x)$.

Remark 3.1. The ϵ-expansion is a generalisation of the Edgeworth expansion. To see this, take $m = 1$ and consider the representation of Edgeworth-type expansions originally given by Davis (1976); see also McCullagh (1987, p.145) where it is assumed that the random variables Y and Z are independent, which results in an Edgeworth-type expansion. If we generalise by allowing Y and Z to have arbitrary dependence, this leads to the ϵ-expansion in the case $m = 1$.

Remark 3.2. The right side of (3.4) gives the two leading terms in the ϵ-expansion, i.e. the term which is linear in the ϵ_a and the term which is quadratic in the ϵ_a. We note that the general term of r-th order may be written

$$\frac{(-1)^r}{r!} \sum_{a_1=1}^{m} \cdots \sum_{a_r=1}^{m} \left(\prod_{i=1}^{r} \epsilon_{a_i} c_{a_i} \right) \frac{1}{g_0(x)} \partial_x^r \left\{ g_0(x) \kappa_{a_1 \ldots a_r}(x) \right\} \\ = \frac{(-1)^r}{r!} \sum_{a_1, \ldots, a_r = 1}^{m} \left(\prod_{i=1}^{r} \epsilon_{a_i} c_{a_i} \right) \sum_{j=0}^{r} \binom{r}{j} q_j(x) \partial_x^{r-j} \kappa_{a_1 \ldots a_r}(x), \tag{3.5}$$

where $\kappa_{a_1 \ldots a_r}(x) = E[X_{a_1} \cdots X_{a_r} | X_0 = x]$, and we have used (3.2) and the Leibniz formula to obtain the right side of (3.5). In the Gaussian case where $g_0 = \phi$, (3.5) becomes

$$\frac{1}{r!} \sum_{a_1, \ldots, a_r = 1}^{m} \left(\prod_{i=1}^{r} \epsilon_{a_i} c_{a_i} \right) \sum_{j=0}^{r} (-1)^{r-j} \binom{r}{j} H_j(x) \partial_x^{r-j} \kappa_{a_1 \ldots a_r}(x), \tag{3.6}$$

where the H_j are Hermite polynomials. In all our numerical calculations we have used (3.4). However, in the proof of two of the parts of Theorem 4.2, we need to know something about the structure of (3.6) when $r > 2$.

3.2 The Itô-Taylor expansion

Kloeden and Platen (1992) provide a detailed discussion of Itô-Taylor expansions of solutions of stochastic differential equations. These are small-Δ expansions which are derived by repeated application of Itô's lemma, applied to the drift and diffusion fuction μ and σ, respectively, in (1.1). The particular Itô-Taylor expansion we focus on here is for a unit diffusion stochastic differential equation and is given by

$$Y_\Delta = y_0 + C_0 I_{(0)} + C_{00} I_{(0,0)} + C_{000} I_{(0,0,0)} + I_1 + C_{10} I_{(1,0)} \\ + C_{110} I_{(1,1,0)} + C_{100} I_{(1,0,0)} + C_{010} I_{(0,1,0)} + C_{1110} I_{(1,1,1,0)}. \tag{3.7}$$

In the above, the quantities $C_0, C_{00}, C_{10}, \ldots$, which are all of order 1, are functions of derivatives of μ_Y evaluated at y_0 and θ, and are defined by

$$\begin{aligned}
&C_0 = \mu_Y(y_0, \theta),\ C_{00} = \mu_Y \partial_y \mu_Y + \frac{1}{2}\partial_y^2 \mu_Y,\\
&C_{000} = \mu_Y \left(\mu_Y \partial_y^2 \mu_Y + (\partial_y \mu_Y)^2 + \frac{1}{2}\partial_y^3 \mu_Y \right) + \frac{1}{2}\left(\mu_Y \partial_y^3 \mu_Y + 3\partial_y \mu_Y \partial_y^2 \mu_Y + \right) + \frac{1}{4}\partial_y^4 \mu_Y,\\
&C_{10} = \partial_y \mu_Y,\ C_{100} = \mu_Y \partial_y^2 \mu_Y + \left(\partial_y \mu_Y\right)^2 + \frac{1}{2}\partial_y^3 \mu_Y,\\
&C_{010} = \mu_Y \partial_y^2 \mu_Y + \frac{1}{2}\partial_y^3 \mu_Y,\ C_{110} = \partial_y^2 \mu_Y \quad \text{and} \quad C_{1110} = \partial_y^3 \mu_Y. \qquad (3.8)
\end{aligned}$$

The quantities $I_{(0)}, I_{(1)}, \ldots$ are defined by (Kloeden and Platen, 1992)

$$\begin{aligned}
&I_{(0)} = \Delta,\ I_{(0,0)} = \Delta^2/2,\ I_{(0,0,0)} = \Delta^3/6,\ I_{(1)} = B_\Delta,\ I_{(1,0)} = \int_0^\Delta B_u \mathrm{d}u,\\
&I_{(1,0,0)} = \int_{s=0}^\Delta \int_{u=0}^s B_u \mathrm{d}u \mathrm{d}s,\ I_{(0,1,0)} = \int_{s=0}^\Delta \int_{u=0}^s u \mathrm{d}B_u \mathrm{d}s,\\
&I_{(1,1,0)} = \int_{s=0}^\Delta \int_{u=0}^s B_u \mathrm{d}B_u \mathrm{d}s \quad \text{and} \quad I_{(1,1,1,0)} = \int_{s=0}^\Delta \int_{u=0}^s \int_{v=0}^u B_v \mathrm{d}B_v \mathrm{d}B_u \mathrm{d}s. \qquad (3.9)
\end{aligned}$$

The order of a typical $I_{(\nu)}$ where ν is a finite sequence of zeros and ones is $O(\Delta^{u+\psi/2})$, where u is the number of zeros and ψ is the number of ones in ν.

The expression (3.7) and formulae in (3.8) and (3.9) follow easily from the expression in Kloeden and Platen [20, p. 182], noting that in formula (5.1) (op. cit.), $m = 1$ and, because we are considering a unit diffusion, i.e. $b = 1$ in their notation, so that many terms in the resulting expression disappear.

By rescaling the time variables in the stochastic integrals in (3.9), and using the scaling property of Brownian motion, it is easily checked that

$$\begin{aligned}
&I_{(1,0)} = \Delta^{3/2}\check{I}_{(1,0)},\quad I_{(1,1,0)} = \Delta^2 \check{I}_{(1,1,0)},\quad I_{(1,0,0)} = \Delta^{5/2}\check{I}_{(1,0,0)}\\
&I_{(0,1,0)} = \Delta^{5/2}\check{I}_{(0,1,0)},\quad I_{(1,1,1,0)} = \Delta^{5/2}\check{I}_{(1,1,1,0)}, \qquad (3.10)
\end{aligned}$$

where in all cases in (3.10), $\check{I}_{(\nu)}$ indicates evaluation at $\Delta = 1$, and the $\check{I}_{(\nu)}$ are fixed random variables which do not depend on Δ or the sample size n. Using (3.9) and (3.10), we may re-write (3.7) as

$$\begin{aligned}
Y_\Delta = y_0 + \Delta C_0 + \frac{1}{2}\Delta^2 C_{00} + \frac{1}{6}\Delta^3 C_{000} + \Delta^{1/2}\check{I}_{(1)} + \Delta^{3/2} C_{10}\check{I}_{(1,0)}\\
+ \Delta^2 C_{110}\check{I}_{(1,1,0)} + \Delta^{5/2}\left\{ C_{100}\check{I}_{(1,0,0)} + C_{010}\check{I}_{(0,1,0)} + C_{1110}\check{I}_{(1,1,1,0)} \right\}. \qquad (3.11)
\end{aligned}$$

Note that the right side of (3.11) contains all terms in the Itô-Taylor expansion of Y_Δ up to and including those of order $\Delta^{5/2}$.

Since the $\check{I}_{(\nu)}$ in (3.11) are fixed random variables, we can apply the ϵ-expansion directly to (3.11) for the purpose of approximating the transition density.

3.3 Application of the ϵ-expansion to the Itô-Taylor expansion

To apply the ϵ-expansion to (3.11), and noting (3.1), we define

$$\begin{aligned}
X &= \Delta^{-1/2}\left(Y_\Delta - y_0 - \Delta C_0 - \frac{1}{2}\Delta^2 C_{00} - \frac{1}{6}\Delta^3 C_{000}\right) \\
&= \check{I}_{(1)} + \Delta C_{10}\check{I}_{(1,0)} + \Delta^{3/2}C_{110}\check{I}_{(1,1,0)} + \Delta^2\left(C_{100}\check{I}_{(1,0,0)} + C_{010}\check{I}_{(0,1,0)} + C_{1110}\check{I}_{(1,1,1,0)}\right) \\
&= X_0 + \sum_{a=1}^{5}\epsilon_a c_a X_a, \qquad (3.12)
\end{aligned}$$

where $X_0 = \check{I}_{(1)}$, $X_1 = \check{I}_{(1,0)}$, $X_2 = \check{I}_{(1,1,0)}$, $X_3 = \check{I}_{(1,0,0)}$, $X_4 = \check{I}_{(0,1,0)}$ and $X_5 = \check{I}_{(1,1,1,0)}$; $\epsilon_1 = \Delta$, $\epsilon_2 = \Delta^{3/2}$ and $\epsilon_3 = \epsilon_4 = \epsilon_5 = \Delta^2$; and $c_1 = C_{10}$, $c_2 = C_{110}$, $c_3 = C_{100}$, $c_4 = C_{010}$ and $c_5 = C_{1110}$. Note that X_0 is a standard normal random variable and that $X_1, \ldots, X_5$ are not independent of X_0, as they are based on the same Brownian motion.

To calculate (3.3) in the case of (3.12) we need to find the conditional moments $\kappa_a(x) = E[X_a|X_0 = x]$ for $a = 1, \ldots, 5$, and $\kappa_{11}(x) = E[X_1^2|X_0 = x]$. The easiest way to calculate these conditional moments is to use the representation of Brownian motion considered by Shepp (1982), which is equivalent to the Karhunen–Loève expansion. Relevant calculations are performed by Preston and Wood (2012), who calculate the moment generating function of $X_0 + \sum_{a=1}^{4}\epsilon_a c_a X_a$; the calculations needed here are much simpler than in Preston and Wood (2012) because only low-order moments are needed. These turn out to be

$$\begin{aligned}
&\kappa_1(x) = \frac{1}{2}H_1(x), \quad \kappa_2(x) = \frac{1}{6}H_2(x), \quad \kappa_3(x) = \frac{1}{6}H_1(x), \quad \kappa_4(x) = \frac{1}{6}H_1(x), \\
&\kappa_5(x) = \frac{1}{24}H_3(x) \quad \text{and} \quad \kappa_{11}(x) = \frac{1}{4}x^2 + \frac{1}{12}, \qquad (3.13)
\end{aligned}$$

where H_j is Hermite polynomial of degree j; details of the calculations leading to (3.13) are given in part B5 of the Supplementary Material.

In the present context it is not feasible to check the sufficient conditions for the validity of the ϵ-expansion stated by Cox and Reid (1987) and Barndorff-Nielsen and Cox (1989), so we have adopted a different approach for proving validity. In brief, the remainder term in the ϵ-expansion is bounded by comparison with the remainder term in the expansion of the exact representation of the transition density (2.5). See the statement and proof of Theorem 3.1 below for further details.

3.4 Assumptions

We now state the assumptions on $\mu(x, \alpha)$ and $\sigma(x, \beta)$ in (1.1) which are needed in subsequent results. Below, we assume that $\theta = (\alpha^\top, \beta^\top)^\top$ denotes the full parameter vector and that $\theta \in \Theta$, where Θ is the parameter space. Define also

$$\mathcal{A} = \{\alpha : (\alpha^\top, \beta^\top)^\top \in \Theta \text{ for some } \beta\} \quad \text{and} \quad \mathcal{B} = \{\beta : (\alpha^\top, \beta^\top)^\top \in \Theta \text{ for some } \alpha\}.$$

We have already defined ∂_x for real x; if α is a p-vector, we define ∂_α to be the gradient with respect to α, and for integer $h > 1$ we interpret ∂_α^h to be an array with p^h elements, where each element is a partial derivative of order h.

(A0) The parameter space Θ is a bounded open set. Moreover, this model is identifiable in the sense that if $\theta_1, \theta_2 \in \Theta$ and $\theta_1 \neq \theta_2$, then the resulting SDEs are distinct.

(A1) The diffusion function σ is bounded away from zero, i.e.

$$\inf_{x\in\mathbf{R}}\inf_{\beta\in\mathcal{B}}\sigma(x,\beta)>0.$$

$(\mathrm{A2})_k$ For $j=0,1,\ldots,4$ and $h=0,\ldots,k$, the functions $\partial_x^j\partial_\alpha^h\mu(x,\alpha)$ and $\partial_x^j\partial_\beta^h\sigma(x,\beta)$ are continuous in their arguments and satisfy

$$\sup_{x\in\mathbf{R}}\sup_{\alpha\in\mathcal{A}}||\partial_x^j\partial_\alpha^h\mu(x,\alpha)||<\infty \quad\text{and}\quad \sup_{x\in\mathbf{R}}\sup_{\beta\in\mathcal{B}}||\partial_x^j\partial_\beta^h\sigma(x,\beta)||<\infty,$$

where $||.||$ denotes the usual Euclidean norm, and we use the obvious definition if the argument is an array.

(A3) For some $r>0$ and $\gamma>0$, and all x such that $|x|\geq r$,

$$\mu(x,\theta_0)\mathrm{sgn}(x)\leq-\gamma,$$

where θ_0 is the true value of θ and $\mathrm{sgn}(x)=-1,0,1$ depending on whether x is negative, zero or positive, respectively.

Remark 3.3. Note that the subscript k in $(\mathrm{A2})_k$ indicates how many derivatives with respect to components of θ should be considered.

3.5 Approximation of the transition density

We are now in a position to state our main result concerning the density approximation derived by applying the ϵ-expansion (3.4) to the Itô-Taylor expansion given in (3.7). The proof is given in part B1 of the Supplementary Material. It will be slightly easier to express the transition density approximation in terms of an approximation to the density of $Z_\Delta=(X_\Delta-x_0)/\Delta^{1/2}$ rather than X_Δ.

Theorem 3.1. *Define* $z_\Delta=(x_\Delta-x_0)/\Delta^{1/2}$ *and*

$$\bar{y}_\Delta\equiv\bar{y}_\Delta(z_\Delta)=\Delta^{-1/2}\left(\gamma(x_0+\Delta^{1/2}z_\Delta,\beta)-\gamma(x_0,\beta)-\Delta C_0-\frac{1}{2}\Delta^2C_{00}-\frac{1}{6}\Delta^3C_{000}\right), \tag{3.14}$$

and assume that assumptions (A0), (A1) and $(A2)_0$ *hold.*

(i) *The ϵ-expansion approximation for the density of the random variable* $Z_\Delta=(X_\Delta-x_0)/\Delta^{1/2}$ *given* $X_0=x_0$ *based on (3.4) and (3.7) up to and including all terms of order* Δ^2 *or larger, is given by*

$$\begin{aligned}\hat{p}_Z^{(\epsilon)}(z_\Delta|x_0;\theta)=\frac{1}{\sigma(x_0+\Delta^{1/2}z_\Delta,\beta)}\phi(\bar{y}_\Delta)\Bigg[1+\frac{1}{2}\Delta C_{10}H_2(\bar{y}_\Delta)+\frac{1}{6}\Delta^{3/2}C_{110}H_3(\bar{y}_\Delta)\\ +\Delta^2\left\{\frac{1}{8}C_{10}^2H_4(\bar{y}_\Delta)+\frac{1}{6}(C_{100}+C_{010}+C_{10}^2)H_2(\bar{y}_\Delta)+\frac{1}{24}C_{1110}H_4(\bar{y}_\Delta)\right\}\Bigg],\end{aligned} \tag{3.15}$$

where, as above, the H_j *are Hermite polynomials and the* C*'s are defined in (3.8).*

(ii) *As* $\Delta\to0$ *we have*

$$\sup_{z_\Delta\in\mathbf{R}}\sup_{x_0\in\mathbf{R}}\sup_{\theta\in\Theta}|\hat{p}_Z^{(\epsilon)}(z_\Delta|x_0;\theta)-p_Z(z_\Delta|x_0;\theta)|=O(\Delta^{5/2}), \tag{3.16}$$

where $p_Z(z_\Delta|x_0;\theta)$ is the exact density of Z_Δ (under model (1.1) with parameter vector θ) obtained from (2.4), (2.5) and the transformation $z_\Delta = (x_\Delta - x_0)/\Delta^{1/2}$.

Remark 3.4. By strengthening assumption $(\mathrm{A2})_0$ to $(\mathrm{A2})_k$ for $k > 0$, the result (3.16) can be extended to uniform bounds on the discrepancy between θ-derivatives of $\hat{p}_Z^{(\epsilon)}$ and θ-derivatives of p_Z.

Remark 3.5. In (3.15) we have chosen to construct an approximation so that the square bracket contains all terms up to and including order Δ^2. There is no obstacle to calculating expansions of higher order. This would involve calculating an Itô-Taylor expansion of higher order (see Kloeden and Platen, 1992) and then applying an ϵ-expansion of higher-order, using formula (3.6), with a larger m in (3.1). To derive higher-order versions of (3.15) one needs to calculate conditional moments of the form $\kappa_{a_1\cdots a_r}(x)$ and their derivatives with respect to x; see Remark 3.2. This can be done using calculations similar to those in Section B5 of the Supplementary Material. These calculations are elementary but become more cumbersome the higher the order of the conditional moment.

Remark 3.6. As was mentioned in the Introduction, showing the essential equivalence of the expansion underlying the *AS* approximation and the expansion underlying the *epsilon* approximation is a non-trivial matter. In Section B6 of the Supplementary Material we offer some comments which indicate how this can be done, but without going into details.

4 Asymptotic Properties of the Estimators

4.1 Asymptotic normality of the exact MLE

Let $\hat{\theta}_{MLE} = (\hat{\alpha}_{MLE}^\top, \hat{\beta}_{MLE}^\top)^\top$ denote the maximum likelihood estimator of θ_0 based on the exact transition density $p(x_\Delta|x_0;\theta)$ of the model (1.1), with log-likelihood specified in (1.2). In Dacunha-Castelle and Florens-Zmirou (1986) it was proved under certain conditions that $\hat{\theta}_{MLE}$ is asymptotically normal and efficient. Let us denote the distribution function of the invariant distribution at the true parameter vector $\theta_0 = (\alpha_0^\top, \beta_0^\top)^\top$ by $F_0(x)$; the invariant distribution F_0 exists under the conditions stated below; see for example the proof due to Gobet (2002). A slightly more general version of the result in Dacunha-Castelle and Florens-Zmirou [11], in which α and β may be vectors and a general $\sigma(x,\beta)$ is permitted, is now stated. The details of the proof are essentially the same and so it is omitted.

Theorem 4.1. *Suppose that $\Delta \to 0$ and $n\Delta \to \infty$. Assume that model (1.1) is correct and let $\theta_0 = (\alpha_0^\top, \beta_0^\top)^\top$ denote the true parameter vector. Under assumptions (A0), (A1), $(A2)_4$ and (A3) in Section 3, the following results hold:*

(i) the estimator $\hat{\theta}_{MLE}$ is consistent, i.e $\hat{\theta}_{MLE} \xrightarrow{p} \theta_0$ as $n \to \infty$;

(ii) the estimator $\hat{\theta}_{MLE} = (\hat{\alpha}_{MLE}^\top, \hat{\beta}_{MLE}^\top)^\top$ is asymptotically normal and efficient, in the sense that

$$\Xi_{n,\Delta}\begin{bmatrix} \hat{\alpha}_{MLE} - \alpha_0 \\ \hat{\beta}_{MLE} - \beta_0 \end{bmatrix} \xrightarrow{d} \mathcal{N}_{p+q}\left(\begin{pmatrix} 0_p \\ 0_q \end{pmatrix}, \begin{pmatrix} \mathcal{I}_{\alpha,0}^{-1} & 0_{q,p} \\ 0_{p,q} & \mathcal{I}_{\beta,0}^{-1} \end{pmatrix}\right), \tag{4.1}$$

where $\Xi_{n,\Delta} = diag\left\{(n\Delta)^{1/2}I_p, n^{1/2}I_q\right\}$, I_p is the $p \times p$ identity matrix, 0_p is the p-vector of zeros, $0_{p,q}$

is the $p \times q$ matrix of zeros, and

$$\mathcal{I}_{\alpha,0} = \int_{x\in\mathbb{R}} \frac{1}{\sigma^2(x,\beta_0)} \frac{\partial\mu}{\partial\alpha}(x,\alpha_0)\frac{\partial\mu}{\partial\alpha^\top}(x,\alpha_0)dF_0(x) \tag{4.2}$$

and

$$\mathcal{I}_{\beta,0} = 2\int_{x\in\mathbb{R}} \frac{1}{\sigma^2(x,\beta_0)} \frac{\partial\sigma}{\partial\beta}(x,\beta_0)\frac{\partial\sigma}{\partial\beta^\top}(x,\beta_0)dF_0(x) \tag{4.3}$$

are the $p \times p$ and $q \times q$ information matrices for α and β respectively, evaluated at $\alpha = \alpha_0$ and $\beta = \beta_0$, and $\partial/\partial\alpha$ and $\partial/\partial\beta$ produce column vectors while $\partial/\partial\alpha^\top$ and $\partial/\partial\beta^\top$ produce row vectors.

4.2 Asymptotic proximity of the estimators

We now study the asymptotic closeness of two estimators, the Kessler (1997) estimator, and the estimator based on the approximate likelihood (3.15) and (4.10) proposed in this paper, to the exact maximum likelihood estimator. Recall that the estimator in Kessler (1997) is based on a suitably high-order approximation of the mean and variance of the observed process, but it uses a Gaussian approximation to the transition density in (1.2) rather than the exact transition density. In order to specify these likelihoods explicitly, for $i = 1, \dots, n$ define $z_i = (x_i - x_{i-1})/\Delta^{1/2}$ and define $\bar{y}_i$ as $\bar{y}_\Delta$ in (3.14), but with z_i replacing z_Δ and x_{i-1} replacing x_0 on the right hand side of (3.14). Let

$$\begin{aligned} p_Z(z_i|x_{i-1};\theta) = {} & \frac{1}{\{2\pi\}^{1/2}\sigma(x_i,\beta)} \exp\left[-\frac{\{\gamma(x_i,\beta)-\gamma(x_{i-1},\beta)\}^2}{2\Delta} + \int_{\gamma(x_{i-1},\beta)}^{\gamma(x_i,\beta)} \mu(u,\theta)du\right] \\ & \times E\left[\exp\left\{\Delta\int_0^1 g\left\{(1-u)\gamma(x_{i-1},\beta) + u\gamma(x_i,\beta) + \Delta^{1/2}B_u\right\}du\right\}\right] \end{aligned} \tag{4.4}$$

denote the exact density of $Z_i = (X_i - X_{i-1})/\Delta^{1/2}$ given $X_{i-1} = x_{i-1}$, where in (4.4), $x_i = x_i(z_i) = x_{i-1} + \Delta^{1/2}z_i$. The corresponding exact likelihood for θ is given by

$$\ell_0(\theta) = \sum_{i=1}^n \log\{p_Z(z_i|x_{i-1};\theta)\}, \tag{4.5}$$

and the exact maximum likelihood estimator is defined as

$$\hat{\theta}_{MLE} = (\hat{\alpha}_{MLE}^\top, \hat{\beta}_{MLE}^\top)^\top = \underset{\theta=(\alpha^\top,\beta^\top)^\top}{\operatorname{argmax}} \sum_{i=1}^n \log\{p_Z(z_i|x_{i-1};\theta)\}. \tag{4.6}$$

We now consider Kessler's (1997, formulae (1.7) and (1.8)) estimator. To match the notion in (1.1) of our paper and (1.1) of Kessler's (1997) paper, we put $b(x,\theta) = \mu(x,\theta)$ and $c(x,\theta) = \sigma(x,\theta)^2$. Then, using results in Kessler (1997), it is found that

$$E[X_i|X_{i-1} = x_{i-1}] = m(x_{i-1},\theta) + O(\Delta^3)$$

and

$$\text{Var}[X_i|X_{i-1} = x_{i-1}] = m_2(x_{i-1},\theta) + O(\Delta^3),$$

where the expectations are taken under the "correct" value of θ,

$$m(x,\theta) = x + \Delta b + \frac{\Delta^2}{2}\left(bb' + \frac{1}{2}cb''\right)$$

and

$$m_2(x,\theta) = \Delta c + \Delta^2\left(\frac{1}{2}bc' + b'c + \frac{1}{4}cc''\right),$$

$b' = \partial_x b(x,\theta)$ and $c'' = \partial_x^2 c(x,\theta)$. The implied approximate conditional mean and conditional variance of $z_i = \Delta^{-1/2}(x_i - x_{i-1})|X_{i-1} = x_{i-1}$ are then given by

$$\tilde{m}(x_{i-1},\theta) = \Delta^{1/2}b + \frac{\Delta^{3/2}}{2}\left(bb' + \frac{1}{2}cb''\right)$$

and

$$\tilde{m}_2(x_{i-1},\theta) = c + \Delta\left(\frac{1}{2}bc' + b'c + \frac{1}{4}cc''\right).$$

Kessler's (1997) Gaussian approximation to the conditional density of $z_i|x_{i-1}$ is then given by

$$\hat{p}_K(z_i|x_{i-1};\theta) = \frac{1}{\sqrt{2\pi\tilde{m}_2(x_{i-1},\theta)}}\exp\left[-\frac{\{z_i - \tilde{m}(x_{i-1},\theta)\}^2}{2\tilde{m}_2(x_{i-1},\theta)}\right], \tag{4.7}$$

the corresponding approximate log-likelihood is

$$\ell_K(\theta) = \sum_{i=1}^{n}\log\{\hat{p}_K(z_i|x_{i-1};\theta)\}, \tag{4.8}$$

and the Gaussian likelihood estimator is defined by

$$\hat{\theta}_K = (\hat{\alpha}_K^\top, \hat{\beta}_K^\top)^\top = \underset{\theta=(\alpha^\top,\beta^\top)^\top}{\operatorname{argmax}} \sum_{i=1}^{n}\log\{\hat{p}_K(z_i|x_{i-1};\theta)\}. \tag{4.9}$$

However, the Gaussian approximation on which (4.7)–(4.9) are based does not account for any skewness or kurtosis in the transition density, whereas the approximation likelihood (3.15) and (4.10) proposed here includes two terms which account for skewness and kurtosis, namely

$$\frac{1}{6}\Delta^{3/2}C_{110}H_3(\bar{y}_\Delta) \quad \text{and} \quad \frac{1}{24}\Delta^2 C_{1110}H_4(\bar{y}_\Delta).$$

The new estimator $\hat{\theta}^{(\epsilon)}$ is based on the density $\hat{p}^{(\epsilon)}(\cdot|x_{i-1},\theta)$ defined in (3.15); that is,

$$\ell^{(\epsilon)}(\theta) = \sum_{i=1}^{n}\log\left\{\hat{p}^{(\epsilon)}(z_i|x_{i-1};\theta)\right\},$$

where $z_i = (x_i - x_{i-1})/\Delta^{1/2}$, $i = 1,\ldots,n$. Then we define the new estimator $\hat{\theta}^{(\epsilon)}$ by

$$\hat{\theta}^{(\epsilon)} = \underset{\theta=(\alpha^\top,\beta^\top)^\top}{\operatorname{argmax}} \ \ell^{(\epsilon)}(\theta). \tag{4.10}$$

Interestingly, the exact maximum likelihood estimator $\hat{\theta}_{MLE}$, Kessler's (1997) estimator $\hat{\theta}_K$ and the estimator $\hat{\theta}^{(\epsilon)}$ obtained using (3.15) and (4.10) as an approximate likelihood, all satisfy Theorem 4.1 and in particular are first-order efficient. However, it turns out that, in an asymptotic sense, the distance between $\hat{\theta}^{(\epsilon)}$ and $\hat{\theta}_{MLE}$ is an order of magnitude smaller than the distance between $\hat{\theta}_K$ and $\hat{\theta}_{MLE}$. These findings are summarised in the theorem below.

Theorem 4.2. *Assume we are in the framework of model (1.1), with observations $x(t_0), x(t_1),\ldots,x(t_n)$ at equally-spaced time points $t_0, t_1,\ldots,t_n$, with $t_i - t_{i-1} = \Delta$. Suppose that $n \to \infty$, $\Delta \to 0$ and $n\Delta \to \infty$. If assumptions (A0), (A1), $(A2)_4$ and (A3) are all satisfied then the following results hold.*

(a) *If* $n\Delta^5 \to 0$ *then*

$$(n\Delta)^{1/2}||\hat{\alpha}_K - \hat{\alpha}_{MLE}|| = O_p\left\{(n\Delta)^{-1/2} + \Delta^{1/2} + (n\Delta^5)^{1/2}\right\}$$
$$= \begin{cases} O_p\left\{(n\Delta)^{-1/2}\right\} & \text{if } n\Delta^2 \to 0 \\ O_p\left\{\Delta^{1/2}\right\} & \text{if } n\Delta^2 \to \infty \quad \text{and} \quad n\Delta^4 \to 0\,; \\ O_p\left\{\left(n\Delta^5\right)^{1/2}\right\} & \text{if } n\Delta^4 \to \infty \end{cases} \tag{4.11}$$

(b) *if* $n\Delta^7 \to 0$ *then*

$$(n\Delta)^{1/2}||\hat{\alpha}^{(\epsilon)} - \hat{\alpha}_{MLE}|| = O_p\left\{(n\Delta)^{-1/2} + \Delta^{3/2} + (n\Delta^7)^{1/2}\right\}$$
$$= \begin{cases} O_p\left\{(n\Delta)^{-1/2}\right\} & \text{if } n\Delta^4 \to 0 \\ O_p\left\{\left(n\Delta^7\right)^{1/2}\right\} & \text{if } n\Delta^4 \to \infty \end{cases}; \tag{4.12}$$

(c) *if* $n\Delta^5 \to 0$ *then*

$$n^{1/2}||\hat{\beta}_K - \hat{\beta}_{MLE}|| = O_p\left\{(n\Delta)^{-1/2} + \Delta + (n\Delta^5)^{1/2}\right\} = \begin{cases} O_p\left\{(n\Delta)^{-1/2}\right\} & \text{if } n\Delta^3 \to 0 \\ O_p\left\{\left(n\Delta^5\right)^{1/2}\right\} & \text{if } n\Delta^3 \to \infty \end{cases}; \tag{4.13}$$

(d) *and if* $n\Delta^7 \to 0$ *then*

$$n^{1/2}||\hat{\beta}^{(\epsilon)} - \hat{\beta}_{MLE}|| = O_p\left\{(n\Delta)^{-1/2} + \Delta^2 + (n\Delta^7)^{1/2}\right\} = \begin{cases} O_p\left\{(n\Delta)^{-1/2}\right\} & \text{if } n\Delta^4 \to 0 \\ O_p\left\{\left(n\Delta^7\right)^{1/2}\right\} & \text{if } n\Delta^4 \to \infty \end{cases}. \tag{4.14}$$

Moreover, the rates specified in (4.11)–(4.14) are typically achieved.

Remark 4.1. In words, Theorem 4.2 tells us that $(n\Delta)^{1/2}||\hat{\alpha}^{(\epsilon)} - \hat{\alpha}_{MLE}||$ is never of larger order, in probability, than $(n\Delta)^{1/2}||\hat{\alpha}_K - \hat{\alpha}_{MLE}||$, and it will be of strictly smaller order, in probability, when $n\Delta^2 \to \infty$. Similarly, $n^{1/2}||\hat{\beta}^{(\epsilon)} - \hat{\beta}_{MLE}||$ is never of larger order, in probability, than $n^{1/2}||\hat{\beta}_K - \hat{\beta}_{MLE}||$, and it will be of strictly smaller order, in probability, when $n\Delta^3 \to 0$.

Remark 4.2. The final statement in Theorem 4.2, that the stated rates are typically achieved, follows in each case from careful scrutiny of the leading-order terms in the remainder: from inspection it is clear that these terms will typically be non-zero.

Remark 4.3. Consider the version of Ait-Sahalia's (2002, p.38 and formulae (4.10) and (4.11)) transition density approximation which includes all terms up to and including those of size Δ^3, as opposed to all terms up to and including size Δ^2 in (3.15). Under suitable assumptions, it is possible to derive an analogue of Theorem 4.2 in this case, using a similar proof to that given in the proof of Theorem 4.2. We omit this as the details are rather involved, but the main message is that in some circumstances, Ait-Sahalia's (2002) estimator is closer by an order of magnitude to the MLE than the version of $\hat{\theta}^{(\epsilon)}$ based on (3.15). However, the results in the next section indicate that these theoretical differences are difficult to detect numerically and the practical benefits are negligible in typical examples.

5 The results of numerical experiments

5.1 Preliminaries

In this section numerical results comparing transition density approximations and approximate maximum likelihood estimators are given. The following notation is used throughout this section. We refer to the ϵ-expansion developed in (3.15) as *epsilon*. This is the proposed method of the paper. We also look at three approximations from the recent literature: the direct Hermite polynomial approximation approach proposed by Ait-Sahalia (2002) as implemented in the Matlab code found at `https://www.princeton.edu/~yacine/closedformmle.htm` for the current version dated July 19, 2018, which we refer to by *AS*; the local linearisation approach of Shoji and Ozaki (1998), which we refer to by *SO*; and the method of Kessler (1997), which we refer to by *Kessler*.

We carry out the numerical studies using the well-known Cox-Ingersoll-Ross process, often used to model interest rates, given by

$$\mathrm{d}X_t = \theta_1 \left(\theta_2 - X_t\right) \mathrm{d}t + \theta_3 \sqrt{X_t} \mathrm{d}B_t. \tag{5.1}$$

This process has been widely used in similar numerical studies, such as Durham and Gallant (2002) and Preston and Wood (2012), due to the fact that this is among the few stochastic differential equations for which the transition density is the density of a well-known distribution (non-central χ^2 in this case). In part A of the Supplementary Material, we give numerical results for the Ornstein-Uhlenbeck (OU) process, which has also been used in the above reference.

We mention one point concerning linear changes of the time variable. If $a > 0$ is a constant and $t \mapsto at$ then $dt \mapsto adt$ and $dB_t \mapsto \sqrt{a} dB_t$. Consequently, when $t \mapsto at$, $\theta_1 \mapsto a\theta_1$ and $\theta_3 \mapsto \sqrt{a}\theta_3$, with θ_2 unchanged. Therefore a linear time-change is equivalent to a reparametrisation of the model. Hence there is nothing special about the time increment $\Delta = 1/12$ used in many of the examples below, except that it was chosen to match the value used in many previous papers. A similar remark applies to the OU model specified in formula (A1) in Section A1 of the Supplementary Material.

The absolute error of the log density is used as a measure to compare the closed-form transition density approximations. We plot the known transition density functions for the Cox-Ingersoll-Ross processes and the transition density approximations. Consider the following definition.

Definition 5.1. *Given an initial value x_0, and an interval between observations Δ, define $p(x_\Delta|x_0, \theta)$ as the true density of a diffusion process, and $\tilde{p}(x_\Delta|x_0, \theta)$ as the approximate density. Then the Absolute Error of the Log Density is defined as*

$$AELD(x_\Delta|x_0, \theta) = |\ln \tilde{p}(x_\Delta|x_0, \theta) - \ln p(x_\Delta|x_0, \theta)|.$$

Each figure consists of two plots. The top plot depicts the transition density over a subset of the domain of the process and beneath it $AELD(x_\Delta|x_0, \theta)$ is given for each approximation method.

In addition, we compare the accuracy of the approximate maximum likelihood estimators found from using the various closed-form transition density approximations. The metrics used for determining the accuracy of these estimators are the median absolute difference and root

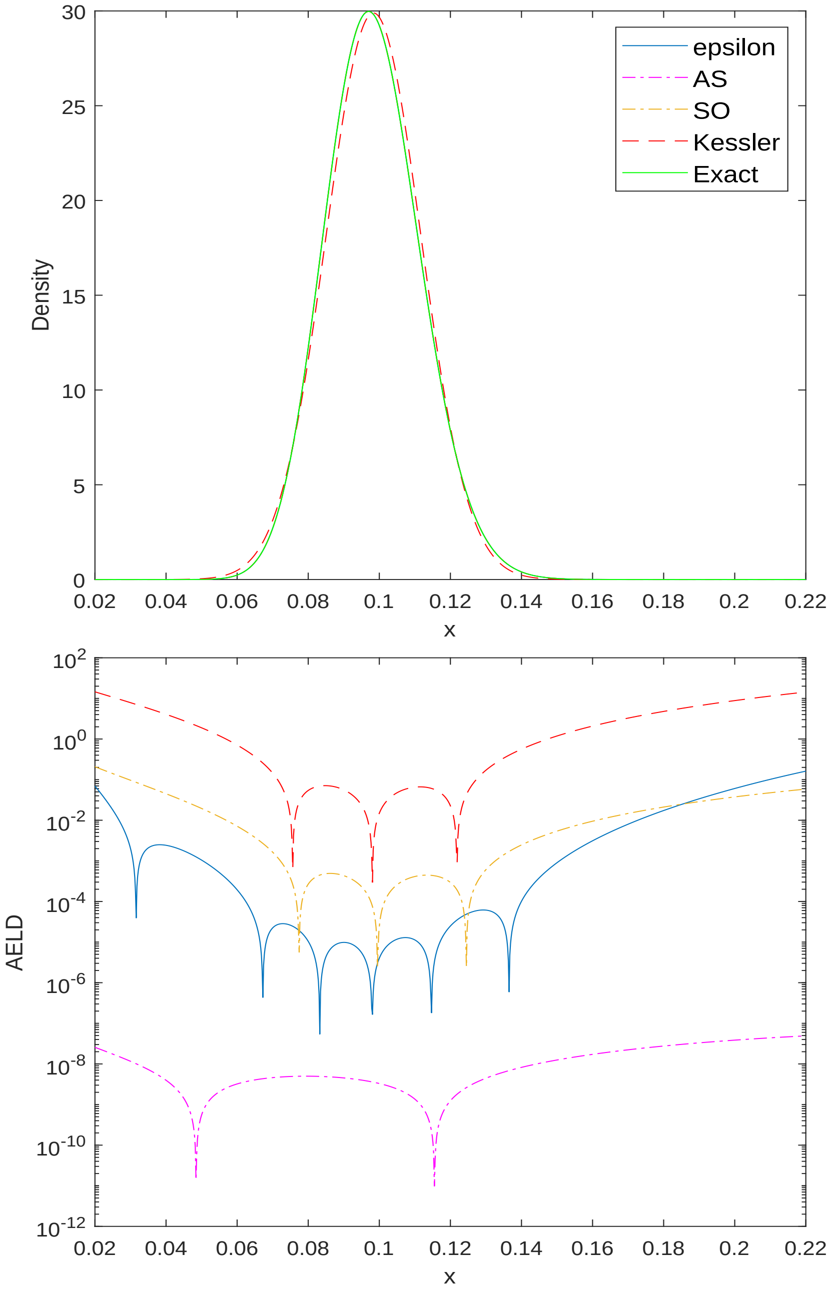

Figure 1: (Top) Approximations of the transition density, specified in §5.1, of the Cox-Ingersoll-Ross process (5.1) and (bottom) corresponding absolute error of the log densities. The model parameters are $(\theta_1, \theta_2, \theta_3) = (0.5, 0.06, 0.15)$, $x_0 = 0.1$ and $\Delta = \frac{1}{12}$.

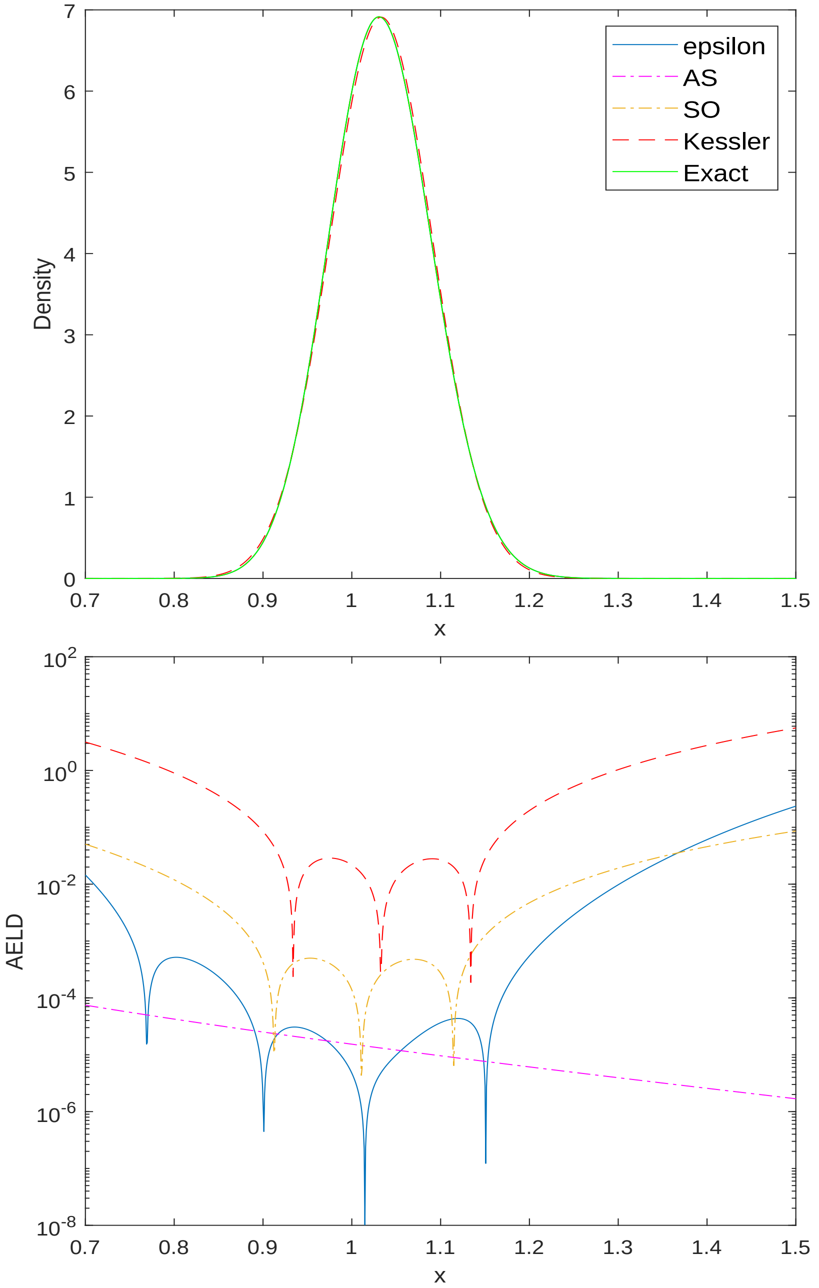

Figure 2: (Top) Approximations of the transition density, specified in §5.1, of the Cox-Ingersoll-Ross process (5.1) and (bottom) corresponding absolute error of the log densities. The model parameters are $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $x_0 = 1$ and $\Delta = \frac{1}{12}$.

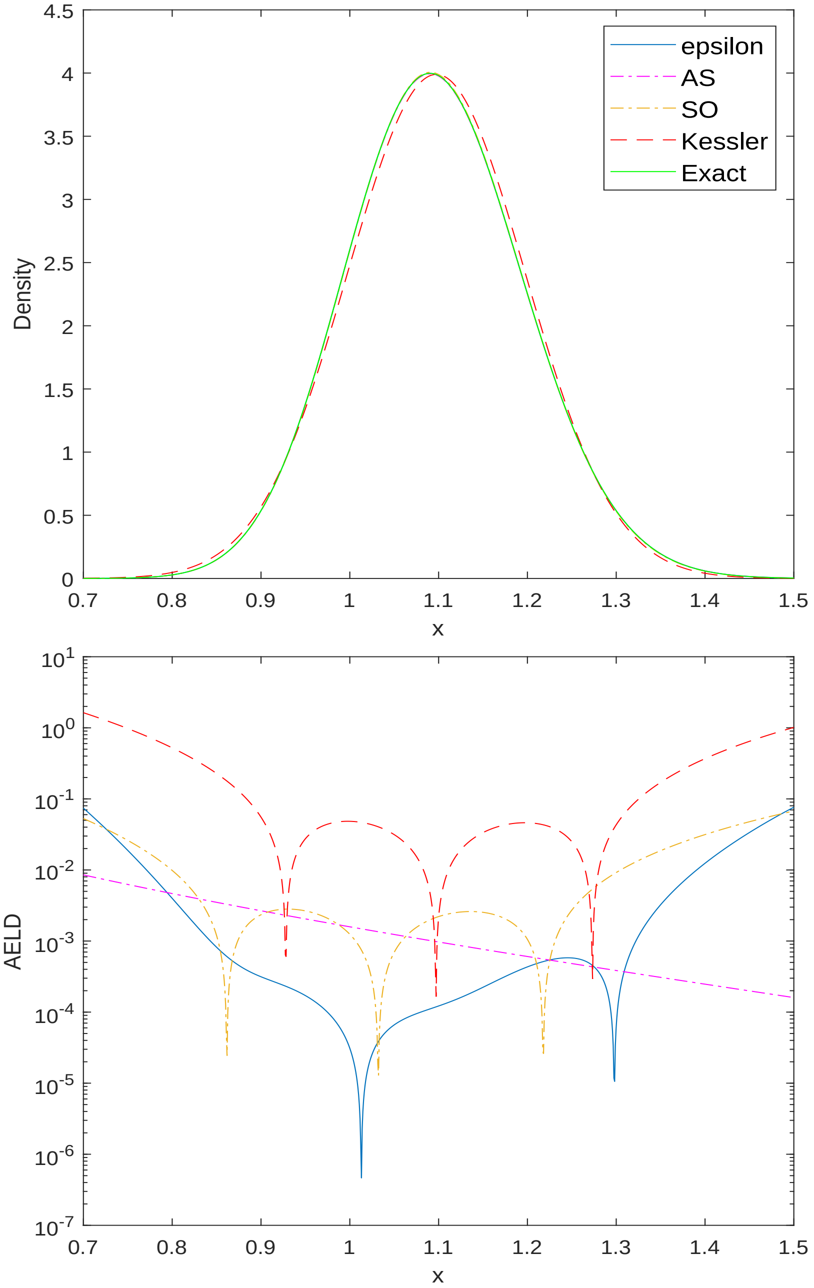

Figure 3: (Top) Approximations of the transition density, specified in §5.1, of the Cox-Ingersoll-Ross process (5.1) and (bottom) corresponding absolute error of the log densities. The model parameters are $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $x_0 = 1$ and $\Delta = \frac{1}{4}$.

mean square error (rmse). The exact method is the exact maximum likelihood estimator $\hat{\theta}_{MLE}$. For the exact method, we compare $\hat{\theta}_{MLE}$ with the true parameter vector θ_0 used to generate the data. For all the approximate methods, comparisons with both θ_0 and $\hat{\theta}_{MLE}$ are given. We report the median because we are looking at a skewed distribution of absolute differences between the $\hat{\theta}_{MLE}$ and the other estimators. These methods were run on Matlab, and we used the the Nelder-Mead method for optimisation as implemented by the function fminsearch, which is consistent with the implementation of the *AS* method. We also report the average time to compute the estimator in seconds per Monte Carlo run. The simulations were run on a desktop computer with 3GHz processor speed. All simulations used $n = 1000$ simulated data points (that is, 1001 observations including the fixed initial value x_0) from the known distribution and there were 5000 Monte Carlo runs.

5.2 Numerical results

		Exact	**epsilon**	**AS**	**SO**	**Kessler**
θ_1	median $\lvert\hat{\theta} - \theta_0\rvert$	0.07904	0.07913	0.07899	0.07917	0.08119
	rms $\lvert\hat{\theta} - \theta_0\rvert$	0.13124	0.13103	0.13127	0.13096	0.13478
	median $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00158	0.00006	0.00242	0.02125
	rms $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.01378	0.00328	0.00573	0.03313
θ_2	median $\lvert\hat{\theta} - \theta_0\rvert$	0.00531	0.00531	0.00530	0.00531	0.00531
	rms $\lvert\hat{\theta} - \theta_0\rvert$	0.00791	0.00791	0.00791	0.00791	0.00791
	median $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00001	0.00000	0.00001	0.00002
	rms $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00005	0.00001	0.00002	0.00006
θ_3	median $\lvert\hat{\theta} - \theta_0\rvert$	0.00235	0.00234	0.00235	0.00235	0.00231
	rms $\lvert\hat{\theta} - \theta_0\rvert$	0.00347	0.00347	0.00347	0.00347	0.00354
	median $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00001	0.00000	0.00001	0.00042
	rms $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00011	0.00005	0.00005	0.00064
Time (sec)		12.83800	0.19411	1.16876	0.22743	0.17332

Table 1: Performance of the various estimators, specified in §5.1, based on a trajectory of 1000 simulated data points from a Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.5, 0.06, 0.15)$, $\Delta = \frac{1}{12}$ and $x_0 = 0.1$ (the same parameter values as used by Durham and Gallant (2002) and Preston and Wood (2012)). Here, $\hat{\theta}$ is $\hat{\theta}_{MLE}$ for the exact method or the estimated value for the corresponding approximate method. Time is the average time to compute the estimator in seconds per Monte Carlo run.

In the results for the first Cox-Ingersoll-Ross process with $\Delta = \frac{1}{12}$ in Figure 1 and Table 1, *AS* is the best performing method, followed by *epsilon*. Results for the second Cox-Ingersoll-Ross process with $\Delta = \frac{1}{12}$ are shown in Figure 2 and Table 2. In this case, the initial location, x_0, is far from the mean-reversion level, and we see that *epsilon* is the best performing method, though the difference between all the approximate estimators and the MLE are small. We also consider the effect of changing the sampling interval to $\Delta = \frac{1}{4}$ for the the second Cox-Ingersoll-Ross process in Figure 3 and Table 3. The error of the approximate estimators compared to the MLE is larger than in Table 2, and the conclusion is similar, with *epsilon* or *SO* being the best performing method depending on the metric used. *Kessler* is the worst performing method in this study. The results of the numerical studies shown in the figures and tables suggest that there is no uniformly best method for approximating the transition density of a diffusion process, though the approximation *epsilon* is either best or competitive in all three examples. In Figures 1 and 3, where the exact transition density is skewed, we see that *Kessler* is unable

		Exact	epsilon	AS	SO	Kessler
θ_1	median $\lvert\hat{\theta}-\theta_0\rvert$	0.04077	0.04076	0.04076	0.04077	0.04088
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.07097	0.07097	0.07063	0.07098	0.07099
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00001	0.00003	0.00005	0.00174
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00003	0.00059	0.00008	0.00265
θ_2	median $\lvert\hat{\theta}-\theta_0\rvert$	0.13484	0.13479	0.13484	0.13484	0.13476
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.19792	0.19792	0.19792	0.19792	0.19796
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00003	0.00003	0.00004	0.00076
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00005	0.00005	0.00007	0.00159
θ_3	median $\lvert\hat{\theta}-\theta_0\rvert$	0.00301	0.00301	0.00301	0.00301	0.00300
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.00447	0.00447	0.00447	0.00447	0.00447
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00000	0.00009
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00000	0.00014
Time (sec)		14.32346	0.21599	1.29910	0.25316	0.19248

Table 2: Performance of the various estimators, specified in §5.1, for a Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $\Delta = \frac{1}{12}$ and $x_0 = 1$. Other details are the same as in Table 1.

to exhibit the required skewness, due to being a normal approximation.

An issue that may arise in practice is that the transition density approximation is not strictly positive. In most examples this only happens rarely if at all, but in some more challenging examples it can happen with higher frequency. In the results in Table 1, in 46 and 4 out of 5000 Monte Carlo runs for *epsilon* and *AS*, respectively, the approximate likelihood function is negative at the true parameter vector $(\theta_1, \theta_2, \theta_3)$. This does not occur in the results in Tables 2 and 3. We avoid this problem by instead using the absolute value of the transition density approximation. This is a "quick fix" which in our examples has done an effective job of sidestepping the negativity problem.

A significant advantage of *epsilon* relative to *AS* is that the former is substantially faster—typically around 6 times faster—than the latter; see Tables 1-3 in this section and the further tables in part A of the Supplementary Material. The exact method is the slowest as it involves evaluating a non-central χ^2 density, which demonstrates the usefulness of the approximation methods even in this example where the transition density is known. Moreover, to apply the transition density approximation *epsilon* to a new diffusion process the user simply has to calculate the relevant derivatives of the drift function given in (3.8) and substitute these into (3.15). From the numerical studies we can see that, despite the simple implementation, the approximation *epsilon* is very accurate. The Matlab programs that we used for all the calculations presented in this paper are available in the Supplementary Materials.

5.3 Distribution of the estimates

In this subsection, we consider the distribution of the estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ based on 5000 Monte Carlo runs of the Cox-Ingersoll-Ross process (5.1) for the exact method (MLE), and the approximation methods *epsilon*, *AS*, *SO*, and *Kessler*. The parameters are as given in Table 2, where we saw that *epsilon* was the best performing approximation method.

Specifically, we provide the histograms (Figure 4), normal QQ plots (Figure 5), and a table of sample moments (Table 4) for the mean, variance, standardised skewness and standardised kurtosis. Under the assumptions of Theorem 4.1, the MLE is asymptotically normal. From

		Exact	epsilon	AS	SO	Kessler
θ_1	median $\lvert\hat{\theta}-\theta_0\rvert$	0.02628	0.02621	0.02591	0.02625	0.02619
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.04185	0.04183	0.04056	0.04186	0.04200
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00007	0.00020	0.00010	0.00175
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00018	0.00223	0.00015	0.00268
θ_2	median $\lvert\hat{\theta}-\theta_0\rvert$	0.07501	0.07499	0.07497	0.07502	0.07503
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.11013	0.11013	0.11013	0.11013	0.11013
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00006	0.00006	0.00004	0.00033
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00011	0.00019	0.00006	0.00060
θ_3	median $\lvert\hat{\theta}-\theta_0\rvert$	0.00304	0.00304	0.00304	0.00305	0.00305
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.00452	0.00452	0.00452	0.00452	0.00455
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00001	0.00000	0.00000	0.00022
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00001	0.00002	0.00001	0.00032
Time (sec)		13.17059	0.19897	1.19451	0.23277	0.17751

Table 3: Performance of the various estimators, specified in §5.1, for a Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $\Delta = \frac{1}{4}$ and $x_0 = 1$. Other details are the same as in Table 1.

these results, the estimates are slightly skewed for $\hat{\theta}_1$ and $\hat{\theta}_2$, and approximately normal for $\hat{\theta}_3$. Based on Figure 4 and Table 4, the distribution of the estimates for all the approximation methods appear to be virtually identical to that of MLE.

However, if we instead consider the histogram of the absolute difference between the estimates of the approximation methods and MLE (see Figure 6), then there is a clear separation. The variance of this absolute difference is much lower for *epsilon* than *Kessler*, which is consistent with Theorem 4.2, and also lower than for *AS*.

		Exact	epsilon	AS	SO	Kessler
$\hat{\theta}_1$	mean	0.23117	0.23117	0.23102	0.23118	0.23108
	variance	0.00407	0.00407	0.00403	0.00407	0.00407
	skewness	0.77308	0.77301	0.75345	0.77339	0.76759
	kurtosis	3.83514	3.83479	3.76104	3.83637	3.81484
$\hat{\theta}_2$	mean	2.99696	2.99697	2.99697	2.99696	2.99706
	variance	0.03917	0.03917	0.03917	0.03917	0.03919
	skewness	0.18363	0.18365	0.18364	0.18360	0.18338
	kurtosis	3.06971	3.06973	3.06965	3.06953	3.06820
$\hat{\theta}_3$	mean	0.20006	0.20006	0.20006	0.20006	0.20008
	variance	0.00002	0.00002	0.00002	0.00002	0.00002
	skewness	0.04583	0.04582	0.04581	0.04580	0.04765
	kurtosis	3.01893	3.01898	3.01900	3.01895	3.01781

Table 4: Various sample moments of 5000 estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ for the Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $\Delta = \frac{1}{12}$ and $x_0 = 1$.

5.4 Discussion

We now provide some discussion of the numerical results in this section and their implications for practice. We use the acronyms defined in §5.1.

1. All four methods perform at least respectably in most of the examples. However, as pre-

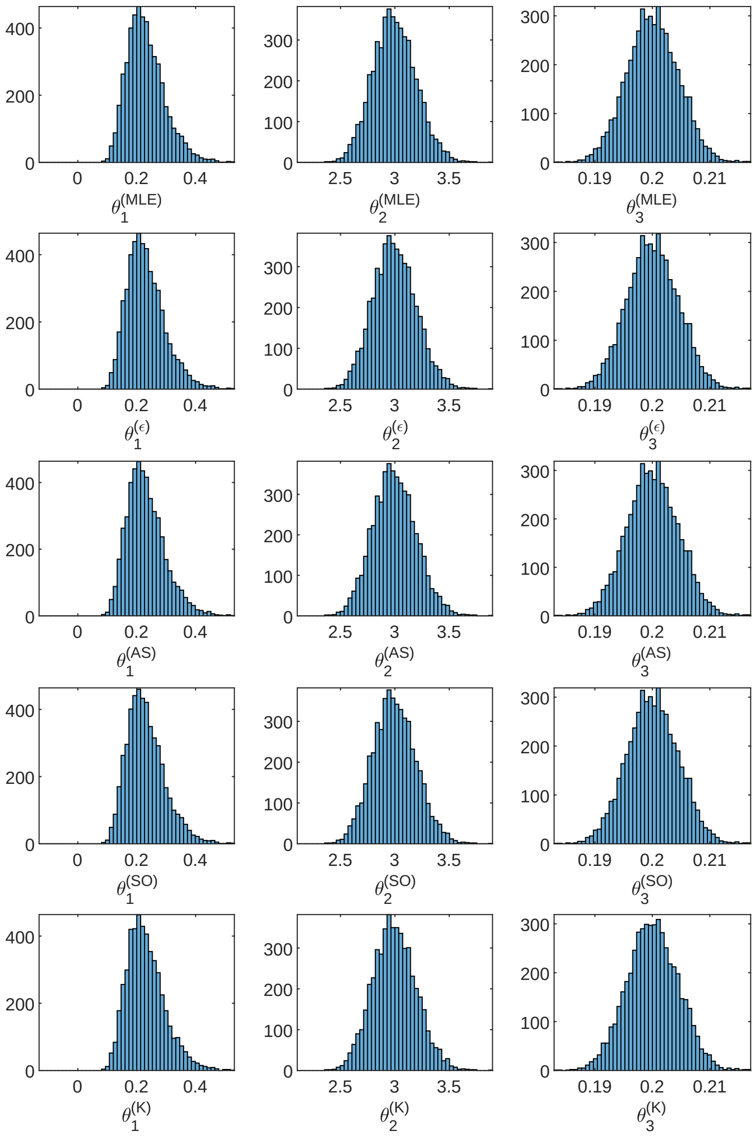

Figure 4: Histogram of estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ for 5000 realisations of the Cox-Ingersoll-Ross process (5.1) with model parameters $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $x_0 = 1$ and $\Delta = \frac{1}{12}$. The plots from top to bottom are for the exact, *epsilon*, *AS*, *SO* and *Kessler* methods. The y-axis is frequency.

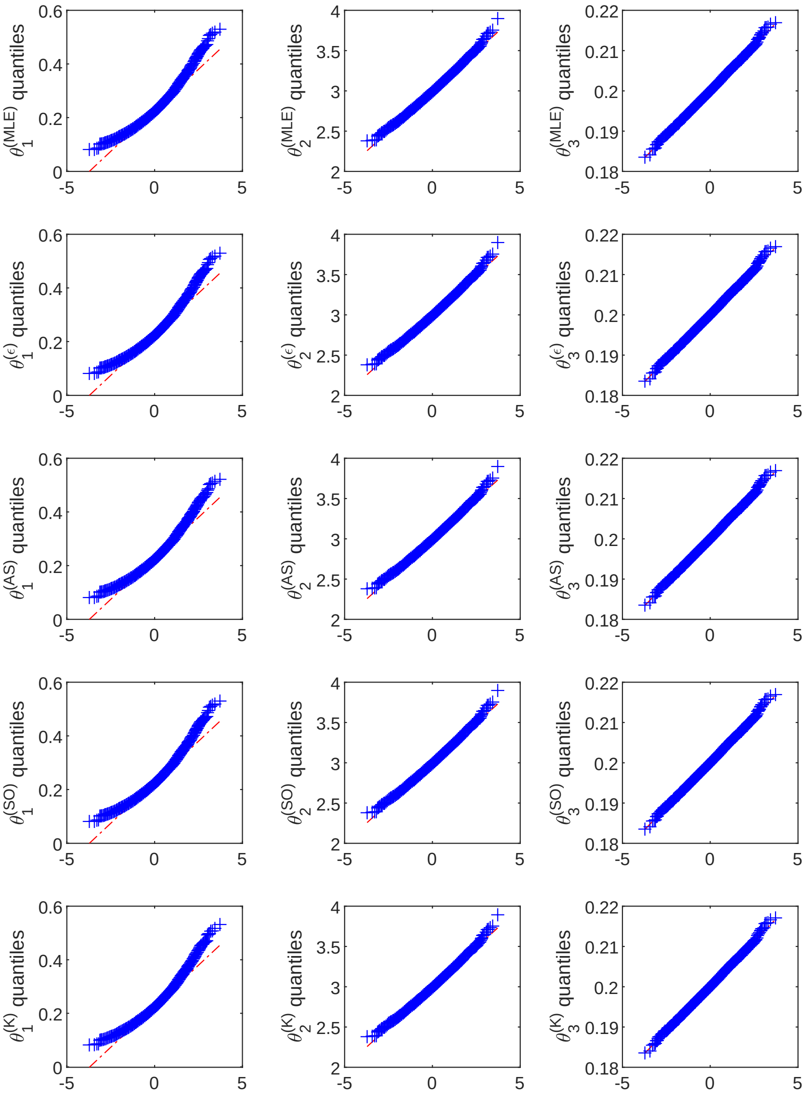

Figure 5: Normal QQ plots of 5000 estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ for the Cox-Ingersoll-Ross process (5.1) with model parameters $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $x_0 = 1$ and $\Delta = \frac{1}{12}$. The plots from top to bottom are for the exact, *epsilon*, *AS*, *SO* and *Kessler* methods. The x-axis are the standard normal quantiles.

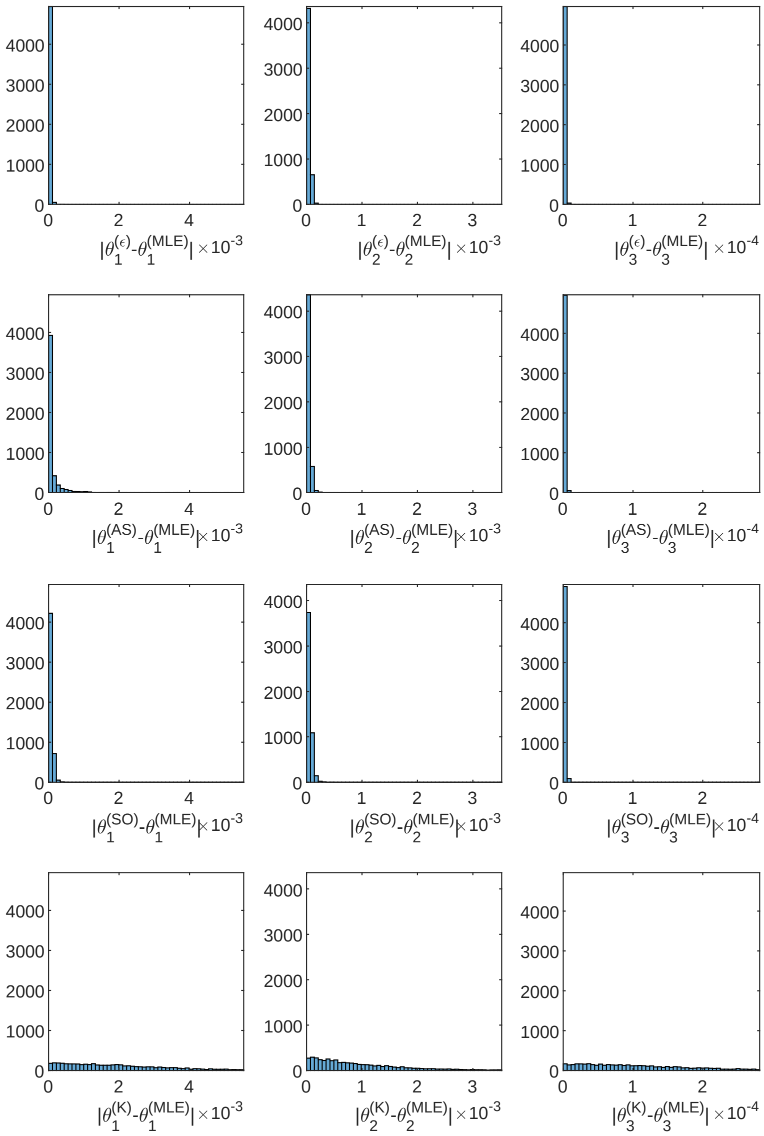

Figure 6: Histogram of 5000 absolute differences $|\hat{\theta} - \hat{\theta}_{MLE}|$ for the Cox-Ingersoll-Ross process (5.1) with model parameters $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $x_0 = 1$ and $\Delta = \frac{1}{12}$. The plots from top to bottom are for the exact, *epsilon*, *AS*, *SO* and *Kessler* methods. The y-axis is frequency.

dicted by the theory, *Kessler* tends to be less accurate than the *epsilon* and *AS* estimators; it also tends to be less accurate than *SO*.

2. In terms of accuracy, there is little to choose between *epsilon*, *AS* and *SO* in most of the examples. Usually the most accurate is one of *epsilon* and *AS*, with *AS* more frequently the best, but usually by a small margin. Occasionally *SO* performs best.

3. In terms of speed, *epsilon* is faster than *AS* by roughly a factor of six and approximately 15-20% faster than *SO*, while *Kessler* is roughly 10% faster than *epsilon*.

4. The implementation of *AS* we have used represents an approximation with error $O(\Delta^3)$ while *epsilon* represents an approximation with error $O(\Delta^2)$, both measured on the scale of the conditional density of z_Δ given by (3.15). This suggests that the rather modest degree of additional accuracy achieved by the former comes at a substantial cost in terms of computational speed.

If both speed and accuracy are important then the numerical results in the examples considered suggest that *epsilon* will be the best choice.

6 Auxiliary Results

6.1 A result on perturbed M-estimators

We now present a result which plays a key role in simplifying the proof of Theorem 4.2. Although we use notation suggestive of a likelihood framework, the result applies to a large class of M-estimators.

Suppose we have a sequence of estimators $(\hat{\theta}_0^{(n)})$ of a parameter vector θ_0 such that $\hat{\theta}_0^{(n)}$ solves the vector of estimating equations $S_0(\hat{\theta}_0^{(n)}) = 0_p$. Consider also a sequence of modified estimators $(\hat{\theta}_1^{(n)})$ which satisfy $S_0(\hat{\theta}_1^{(n)}) + S_1(\hat{\theta}_1^{(n)}) = 0_p$, where in the present context, $S_1(\theta)$ is of smaller order in probability than $S_0(\theta)$. To simplify notation we will drop the superscript (n). Define the $p \times p$ square matrices

$$J_i(\theta) = -\partial_{\theta^\top} S_i(\theta), \quad i = 0, 1, \tag{6.1}$$

and assume that J_0 and J_1 are continuously differentiable in θ. Consider the identity

$$\hat{\theta}_0 - \theta_0 = (J_0^*)^{-1} S_0(\theta_0), \tag{6.2}$$

where J_0^* is suitably chosen. Since the exact form of Taylor's theorem does not hold for vector-valued functions, there does not exist a single $\hat{\theta}_0^*$ lying on the line segment connecting $\hat{\theta}_0$ and θ_0 such that (6.2) holds with $J_0^* = J_0(\theta^*)$. However, it is possible to construct such a J_0^* if the $\hat{\theta}_0^*$ we use can be different in different rows of $J_0(\theta)$. It is assumed below that J_0^* is constructed in this way. We also consider the corresponding perturbed identity

$$\hat{\theta}_1 - \theta_0 = (J_0^{**} + J_1^{**})^{-1} \left\{S_0(\theta_0) + S_1(\theta_0)\right\}, \tag{6.3}$$

where $J_0^{**} + J_1^{**}$ is defined similarly to J_0^*, i.e. a different $\hat{\theta}_1^*$ is used for each row of $J_0(\theta) + J_1(\theta)$.

In the present context we need to allow for the situation in which different components of the vector of parameter estimators converge at different rates. Define $\theta_0 = (\theta_{01}^\top, \ldots, \theta_{0r}^\top)^\top$, where for $j = 1, \ldots, r$, θ_{0j} is a sub-vector of dimension d_j. Similarly, we write $\hat{\theta}_i = (\hat{\theta}_{i1}^\top, \ldots, \hat{\theta}_{ir}^\top)^\top$ and $S_i(\theta) = (S_{i1}^\top, \ldots S_{ir}^\top)^\top$ for $i = 0, 1$; and we define the block diagonal matrix

$$D_n = \operatorname{diag}\left(\delta_{n,1} I_{d_1}, \ldots, \delta_{n,r} I_{d_r}\right). \tag{6.4}$$

Proposition 6.1 *Assume that for all n, $\delta_{n,1} \leq \delta_{n,2} \ldots \leq \delta_{n,r}$ and that $\delta_{n,1} \to \infty$ as $n \to \infty$. Suppose that, with probability approaching 1 as $n \to \infty$, the identities (6.2) and (6.3) both hold. Moreover, suppose that for $j = 1, \ldots, r$*

$$||S_{0j}(\theta_0)|| = O_p(\delta_{n,j}), \qquad ||S_{1j}(\theta_0)|| = o_p(\delta_{n,j}); \tag{6.5}$$

and that, after suitable re-scaling, J_0^* and J_0^{**} converge in probability in the sense that

$$D_n^{-1} J_0^* D_n^{-1} \xrightarrow{p} \mathcal{I}(\theta_0), \qquad D_n^{-1} J_0^{**} D_n^{-1} \xrightarrow{p} \mathcal{I}(\theta_0), \tag{6.6}$$

where $\mathcal{I}(\theta) = diag(\mathcal{I}_1(\theta), \ldots, \mathcal{I}_r(\theta))$ is a block-diagonal matrix, with diagonal block j of dimension $d_j \times d_j$, which is strictly positive definite in some neighbourhood of θ_0; for $j = 1, \ldots, r$,

$$||J_{1j}^{**}|| = o_p(\delta_{n,j}^2); \tag{6.7}$$

and, finally,

$$||D_n \left\{ (J_0^{**})^{-1} - (J_0^*)^{-1} \right\} D_n|| = O_p(\delta_{n,1}^{-1}). \tag{6.8}$$

Then
(i) $\hat{\theta}_0$ and $\hat{\theta}_1$ are both consistent estimators of θ_0 and
(ii) for $j = 1, \ldots, r$,

$$||\delta_{n,j}(\hat{\theta}_{1,j} - \hat{\theta}_{0,j})|| = O_p\left(\delta_{n,j}^{-1} ||S_{1j}(\theta_0)||\right) + O_p\left(\delta_{n,j}^{-2} ||J_{1j}^{**}||\right) + O_p(\delta_{n,1}^{-1}) = o_p(1). \tag{6.9}$$

Remark 6.1. In a standard repeated sampling setting we take $r = 1$ and $\delta_{n,1} = n^{1/2}$. In the setting of this paper, with $\theta^\top = (\alpha^\top, \beta^\top)^\top$ as in (1.1), we have $r = 2$, where we may take $j = 1 < r$ to correspond to the α parameters, and we should choose $\delta_{n,1}$ to be $(n\Delta)^{1/2}$; while $j = 2 = r$ corresponds to the β parameters, where we should choose $\delta_{n,2} = n^{1/2}$.

Remark 6.2. In most of our applications of Proposition 6.1, the first term on the right-hand side of (6.9) will be the dominant one, and the third term will be smaller than the second term.

6.2 Mixing Results

The mixing result that we need is based on a theorem in Veretennikov (1988) combined with a result in Hall and Heyde (1980, Corollary A.2).

Let $(\Omega, \mathcal{F}, P)$ denote a probability space and let $\mathcal{G}$ and $\mathcal{H}$ denote two sub-sigma-fields of $\mathcal{F}$. Then the α-mixing, or strong-mixing, coefficient $\alpha(\mathcal{G}, \mathcal{H})$ between $\mathcal{G}$ and $\mathcal{H}$ is defined by

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}} |P(G \cap H) - P(G)P(H)|.$$

It will be sufficient for our purposes to consider a stochastic differential equation with unit diffusion coefficient of the form

$$dX_t = \mu(X_t)dt + dB_t, \tag{6.10}$$

subject to the initial condition $X_0 = x$, where $\{B_t : t \geq 0\}$ is a standard Brownian motion in $\mathbb{R}$. It is assumed that the solution of (6.10) is (a) unique in distribution, and (b) satisfies the strong Markov property. Sufficient conditions for (a) and (b) to hold are that μ is bounded and Lipschitz, i.e.

$$\sup_{x \in \mathbb{R}} |\mu(x)| < \infty \quad \text{and} \quad |\mu(x) - \mu(y)| < K|x - y|, \quad x, y \in \mathbb{R}, \tag{6.11}$$

for some constant $K \in (0, \infty)$. Note that the conditions in (6.11) follow from $(\mathrm{A2})_0$ in §3.4. We also define $\mathcal{F}^x_{[a,b]}$ to be the sub-sigma-field of $\mathcal{F}$, subject to the initial condition $X_0 = x$, defined by

$$\mathcal{F}^x_{[a,b]} = \sigma(X_t : a \leq b \leq t), \tag{6.12}$$

for any a and b satisfying $0 \leq a \leq b \leq \infty$. The one-dimensional version of Veretennikov's result is now stated.

Result A (Veretennikov, 1988). *Assume that the conditions in (6.11) hold and that $(X_t)_{t \geq 0}$ is the solution of the SDE (6.10). Suppose that, for some $r > 0$ and $\gamma > 0$, and all $|x| \geq r$,*

$$\mu(x)\mathrm{sgn}(x) \leq -\gamma,$$

where $sgn(x) = -1, 0, 1$ depending on whether x is negative, zero or positive, respectively. Then, for some $\lambda > 0$ and any $s > 0$,

$$\alpha(s) \equiv \sup_{t>0} \alpha(\mathcal{F}^x_{[0,t]}, \mathcal{F}^x_{[t+s,\infty)}) \leq C_0 \exp(-\lambda s), \tag{6.13}$$

where $\mathcal{F}^x_{[a,b]}$ is defined in (6.12) and C_0 is a constant independent of s.

We shall also make use of the following result.

Result B (Hall and Heyde, 1980, Corollary A.2) *Suppose that X and Y are random variables which are $\mathcal{G}$-measurable and $\mathcal{H}$-measurable, respectively, and that $E[|X|^p] < \infty$ and $E[|Y|^q] < \infty$, where $p > 1$, $q > 1$ and $p^{-1} + q^{-1} < 1$. Then*

$$|E[XY] - E[X]E[Y]| \leq 8 \left\{E[|X|^p]\right\}^{1/p} \left\{E[|Y|^q]\right\}^{1/q} \left\{\alpha(\mathcal{G}, \mathcal{H})\right\}^{1/r},$$

where $r > 1$ is chosen such that $p^{-1} + q^{-1} + r^{-1} = 1$.

These results lead to the following proposition which is used repeatedly in the proof of Theorem 4.2.

Proposition 6.2. *In the context of Theorem 4.2, suppose that U_i, a function of X_i only, has, for some fixed $\epsilon > 0$, uniformly bounded moments of order $2 + \epsilon$, i.e. $\sup_i E[|U_i|^{2+\epsilon}] < \infty$. Let $\bar{Y}_i$ be defined as in (3.14). Then for any fixed non-negative integer k, with H_k denoting the Hermite polynomial of degree k,*

$$\mathrm{Var}\left\{\sum_{i=1}^{n} U_{i-1} H_k(\bar{Y}_i)\right\} = O(n\Delta^{-1}).$$

The proof of Proposition 6.2 is given in part B4 in the Supplementary Material.

Acknowledgements

We thank three referees and an Associate Editor for careful readings of this paper which have led to a substantially improved presentation and to a more extensive numerical study.

References

[1] Ait-Sahalia, Y. (1999). Transition densities for interest rate and other nonlinear diffusions. *The Journal of Finance* **54** 1361-1395.

[2] Ait-Sahalia, Y. (2002). Maximum likelihood estimation of discretely-sampled diffusions: a closed-form approximation approach. *Econometrica* **70** 223-262.

[3] Ait-Sahalia, Y., Li, C. and Li, C.X. (2021a). Implied stochastic volatility models. *The Review of Financial Studies* **34**, 394-450.

[4] Ait-Sahalia, Y., Li, C. and Li, C.X. (2021b). Closed-form implied volatility surfaces for stochastic volatility models with jumps. *Journal of Econometrics* **222** 364-392.

[5] Barndorff-Nielsen, O.E. and Cox, D.R. (1989). *Asymptotic Techniques for Use in Statistics.* Chapman and Hall, London.

[6] Beskos, A., Papaspiliopoulos, O. and Roberts, G.O. (2009). Monte Carlo maximum likelihood estimation for discretely observed diffusion processes. *The Annals of Statistics* **37** 223-245.

[7] Beskos, A., Papaspiliopoulos, O., Roberts, G.O. and Fearnhead, P. (2006). Exact and efficient likelihood-based inference for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society Series B* **68** 333-382.

[8] Chang, J. and Chen, S.X. (2011). On the approximate likelihood estimation for diffusion processes. *The Annals of Statistics* **39**, 2820-2851.

[9] Cox, D.R. and Reid, N. (1987). Approximations to noncentral distributions. *The Canadian Journal of Statistics* **15** 105-114.

[10] Cox, J.C., Ingersoll, J.E. and Ross, S.A. (1985). A theory of the term structure of interest rates. *Econometrica* **53** 385-408.

[11] Dacunha-Castelle, D. and Florens-Zmirou, D. (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics* **19** 263-284.

[12] Davis, A.W. (1976). Statistical distributions in univariate and multivariate Edgeworth populations. *Biometrika* **63** 661-670.

[13] Durham, B.G. and Gallant, A.R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes, *Journal of Business and Economic Statistics* **20** 297-338.

[14] Gardiner, C.W. (2003). *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. 3rd Edition. Springer, New York.

[15] Gobet, E. (2002). LAN property for ergodic diffusions with discrete observations. *Ann. I. H. Poincaré -PR* **38** 711-737.

[16] Hall, P. and Heyde, C.C. (1980). *Martingale limit theory and its applications.* Academic Press, New York.

[17] Iacus, S.M. (2008). *Simulation and Inference for Stochastic Differential Equations with R Examples*. Springer, New York.

[18] van Kampen, N. G. (2004). *Stochastic Processes in Physics and Chemistry*. 2nd Edition. Elsevier, Amsterdam.

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

[20] Kloeden, P.E. and Platen, R.A. (1992). *Numerical Solutions of Stochastic Differential Equations.* Springer-Verlag, Berlin.

[21] Li, C. (2013). Maximum likelihood estimation for diffusion processes via closed-form density expansions. *The Annals of Statistics* **41**, 1350-1380.

[22] Li, C. (2014). Closed-form expansions, conditional expectations, and option valuation. *Mathematics of Operational Research*, **39**, 487-516.

[23] Li, C. and Chen, D. (2016). Estimating jump-diffusions using closed-form likelihood expansions. *Journal of Econometrics* **195**, 51-70.

[24] Li, C., An, Y., Chen, D., Lin, Q. and Si, N. (2016). Efficient computation of likelihood expansions for diffusion models. *IIE Transactions*, **48**, 1156-1171.

[25] McCullagh, P. (1987). *Tensor Methods in Statistics.* Chapman and Hall, London.

[26] Øksendal, B. K. (1998). *Stochastic Differential Equations: An Introduction with Applications.* Springer-Verlag, Berlin.

[27] Pedersen (1995). A new approach to maximum likelihood estimation for stochastic differential equations. *Scandinavian Journal of Statistics* **22** 55-71.

[28] Prakasa Rao, B.L.S. (1983). Asymptotic theory for non-linear least squares estimator for diffusion processes. *Math. Operationsforsch. Statist.Ser. Stat.* **14** 195-209.

[29] Prakasa Rao, B.L.S. (1988). Statistical inference from sampled data for stochastic processes. *Contemp. Math.* **80** 249-284.

[30] Preston, S.P. and Wood, A.T.A. (2012). Approximation of transition densities of stochastic differential equations by saddlepoint methods applied to small-time Ito-Taylor sample-path expansions. *Statistics & Computing* **22** 205-217.

[31] Shepp, L.A. (1992). On the integral of the absolute value of the pinned Wiener process. *Annals of Probability* **10** 234-239.

[32] Shoji, I. and Ozaki, T. (1998). Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications* **16** 733-752.

[33] Shreve, S.E. (2010). *Stiochastic Calculus for Finance II: Continuous-Time Models*. Springer-Verlag, New York.

[34] Stramer, O. and Yan, J. (2007). On simulated likelihood of discretely observed diffusion processes and comparison to closed-form approximation. *Journal of Computational and Graphical Statistics* **16** 672-691.

[35] Veretennikov, A.Yu. (1988). Bounds for the mixing rate in the theory of stochastic equations. *Theory of Probability and its Applications* **32** 273-281.

[36] Wan, X. and Yang, N. (2021). Hermite expansions for transition densities of irreducible diffisusions with an application to option pricing. *Journal of Economic Dynamics and Control*, forthcoming.

[37] Watanabe, S. (1987). Analysis of Wiener functionals (Malliavin calculus) and its application to heat kernels. *The Annals of Probability*, **15**, 1-39.

[38] Wilkinson, D. J. (2006). *Stochastic Modelling for Systems Biology*. Chapman and Hall/CRC, London.

[39] Yang, N., Chen, N. and Wan, X. (2019). A new delta expansion for multivariate diffusions via the Itô-Taylor expansion. *Journal of Econometrics* **209** 256-288.

[40] Yoshida, N. (1992). Estimation for diffusion processes from discrete observation. *Journal of Multivariate Analysis* **41** 220-242.

SUPPLEMENTARY MATERIAL

A: Further numerical results

A1: Numerical results for the Ornstein-Uhlenbeck process

We also carried out the numerical studies using the Ornstein-Uhlenbeck process

$$\mathrm{d}X_t = \theta_1 \left(\theta_2 - X_t\right) \mathrm{d}t + \theta_3 \mathrm{d}B_t, \tag{A1}$$

with the same parameters as were used by Ait-Sahalia (1999) and Preston and Wood (2012); see the caption of Table 4. Note that the Shoji and Ozaki (1998) local linearisation approach is exact for the Ornstein-Uhlenbeck process and hence this is omitted from simulations using this model as a benchmark.

		Exact	**epsilon**	**AS**	**Kessler**
θ_1	median $\lvert\hat{\theta} - \theta_0\rvert$	0.03238	0.03238	0.03238	0.03237
	rms $\lvert\hat{\theta} - \theta_0\rvert$	0.06549	0.06549	0.06549	0.06550
	median $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00003
	rms $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00002	0.00000	0.00005
θ_2	median $\lvert\hat{\theta} - \theta_0\rvert$	0.02439	0.02440	0.02439	0.02440
	rms $\lvert\hat{\theta} - \theta_0\rvert$	0.03830	0.03830	0.03830	0.03831
	median $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00002
	rms $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00001	0.00000	0.00003
θ_3	median $\lvert\hat{\theta} - \theta_0\rvert$	0.00034	0.00034	0.00034	0.00034
	rms $\lvert\hat{\theta} - \theta_0\rvert$	0.00050	0.00050	0.00050	0.00050
	median $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00000
	rms $\lvert\hat{\theta} - \hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00000
Time (sec)		0.09483	0.15956	0.13531	0.14268

Table 5: Performance of the various estimators, specified in §5.1, based on a trajectory of 1000 observations from an Ornstein-Uhlenbeck process (A1) with $(\theta_1, \theta_2, \theta_3) = (0.0717, 0.261, 0.02237)$, $\Delta = \frac{1}{12}$ and $x_0 = 0.1$. Other details are the same as in Table 1.

In Table 5 and Figure 7, *AS* is the best performing method, followed by *epsilon*, though all the approximate methods give almost exactly the same results as the exact method in this example.

A2: Numerical results for changing the sampling interval

For the second Cox-Ingersoll-Ross process in Figure 2 and Table 2, we considered increasing the sampling interval to $\Delta = \frac{1}{4}$ in Figure 3 and Table 3. Here we consider the effect of decreasing the sampling interval to $\Delta = \frac{1}{24}$.

In Table 6, the error of the approximate estimators compared to the MLE are smaller than in Table 2, as expected, and *epsilon* remains the best performing method. When looking at Tables 2, 3 and 6 together, changing the sampling interval does not appear to affect the relative performance of the four approximate methods in this example.

A3: Numerical results with differing numbers of observations

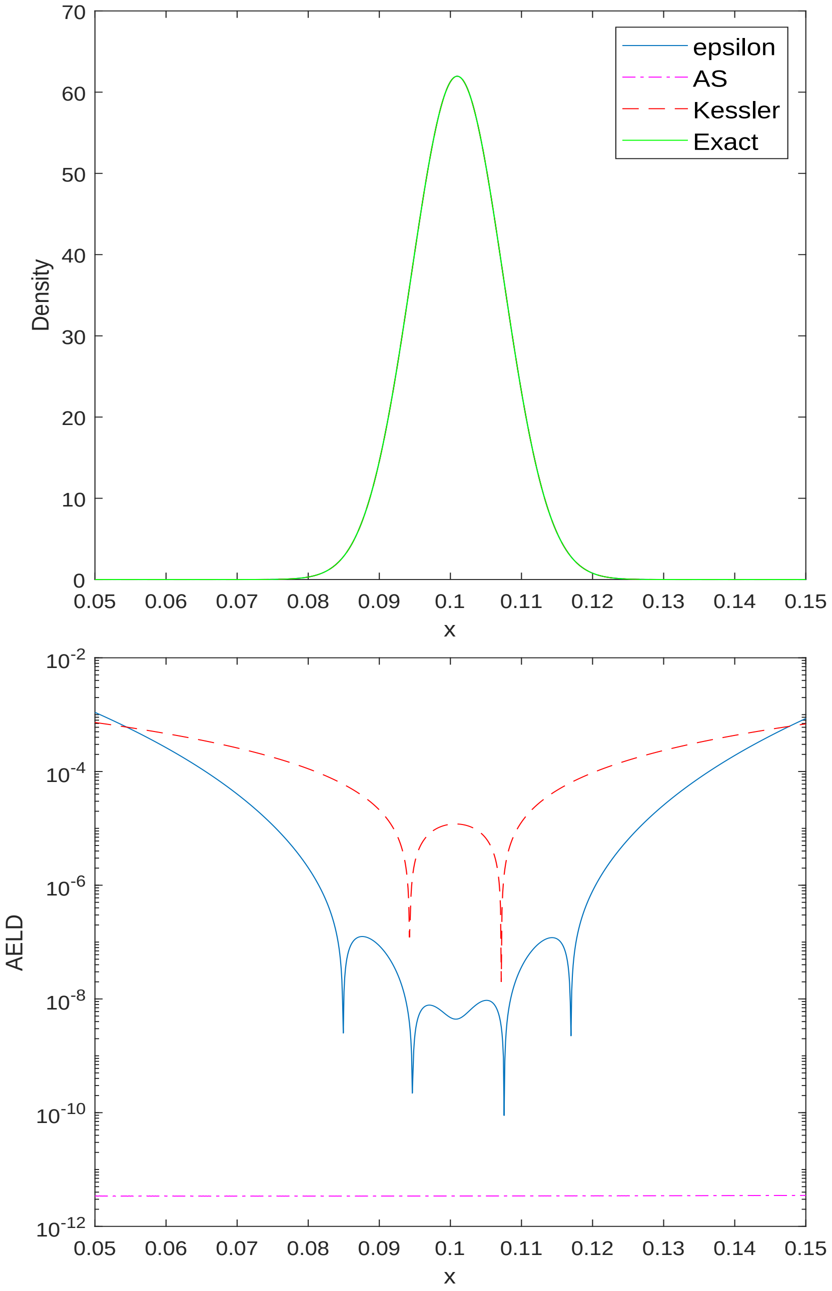

Figure 7: (Top) Approximations of the transition density, specified in §5.1, of the Ornstein-Uhlenbeck process (A1) and (bottom) corresponding absolute error of the log densities. The model parameters are $(\theta_1, \theta_2, \theta_3) = (0.0717, 0.261, 0.02237)$, $x_0 = 0.1$ and $\Delta = \frac{1}{12}$.

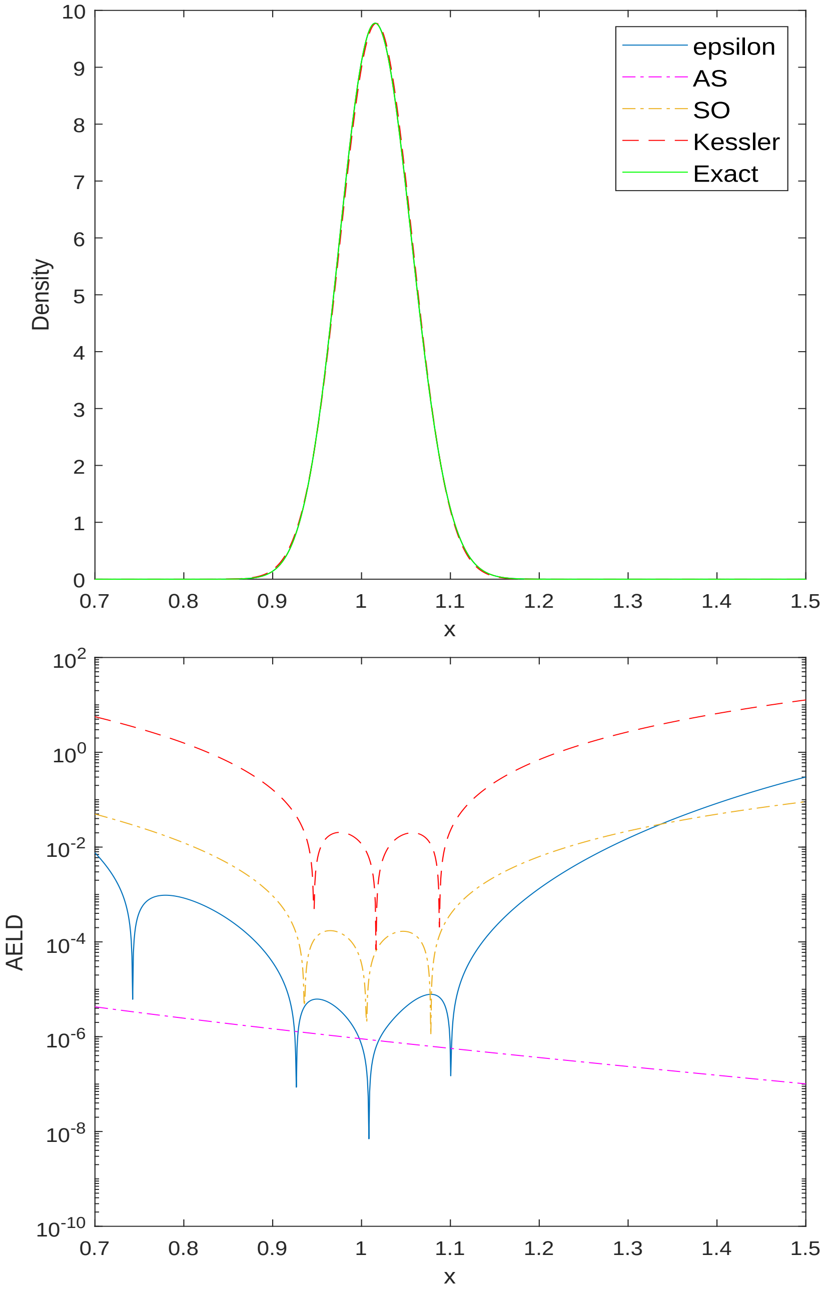

Figure 8: (Top) Approximations of the transition density, specified in §5.1, of the Cox-Ingersoll-Ross process (5.1) and (bottom) corresponding absolute error of the log densities. The model parameters are $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $x_0 = 1$ and $\Delta = \frac{1}{24}$.

		Exact	epsilon	AS	SO	Kessler
θ_1	median $\|\hat{\theta}-\theta_0\|$	0.05416	0.05416	0.05417	0.05417	0.05423
	rms $\|\hat{\theta}-\theta_0\|$	0.09754	0.09754	0.09737	0.09754	0.09746
	median $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00000	0.00001	0.00003	0.00166
	rms $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00002	0.00042	0.00005	0.00252
θ_2	median $\|\hat{\theta}-\theta_0\|$	0.19355	0.19355	0.19355	0.19353	0.19276
	rms $\|\hat{\theta}-\theta_0\|$	0.29793	0.29793	0.29793	0.29792	0.29789
	median $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00001	0.00002	0.00004	0.00126
	rms $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00004	0.00005	0.00008	0.00362
θ_3	median $\|\hat{\theta}-\theta_0\|$	0.00308	0.00309	0.00308	0.00308	0.00309
	rms $\|\hat{\theta}-\theta_0\|$	0.00454	0.00454	0.00454	0.00454	0.00454
	median $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00000	0.00000	0.00000	0.00006
	rms $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00000	0.00000	0.00000	0.00010
Time (sec)		15.06433	0.22798	1.36738	0.26711	0.20362

Table 6: Performance of the various estimators, specified in §5.1, for a Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $\Delta = \frac{1}{24}$ and $x_0 = 1$. Other details are the same as in Table 1.

Next, we consider the effect of changing the number of observations on the numerical results. Using the second Cox-Ingersoll-Ross model in Table 2, we consider the effect of having $n = 500$ and $n = 2000$ simulated data points instead of $n = 1000$.

		Exact	epsilon	AS	SO	Kessler
θ_1	median $\|\hat{\theta}-\theta_0\|$	0.05372	0.05371	0.05355	0.05378	0.05360
	rms $\|\hat{\theta}-\theta_0\|$	0.09710	0.09709	0.09586	0.09711	0.09719
	median $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00001	0.00005	0.00006	0.00224
	rms $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00005	0.00241	0.00011	0.00352
θ_2	median $\|\hat{\theta}-\theta_0\|$	0.19228	0.19230	0.19227	0.19229	0.19234
	rms $\|\hat{\theta}-\theta_0\|$	0.29238	0.29238	0.29240	0.29238	0.29271
	median $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00003	0.00004	0.00006	0.00180
	rms $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00006	0.00021	0.00013	0.00521
θ_3	median $\|\hat{\theta}-\theta_0\|$	0.00417	0.00417	0.00418	0.00418	0.00419
	rms $\|\hat{\theta}-\theta_0\|$	0.00625	0.00625	0.00625	0.00625	0.00625
	median $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00000	0.00000	0.00000	0.00013
	rms $\|\hat{\theta}-\hat{\theta}_{MLE}\|$		0.00000	0.00000	0.00000	0.00020
Time (sec)		7.43730	0.11348	0.67716	0.13286	0.10160

Table 7: Performance of the various estimators, specified in §5.1, based on a trajectory of 500 simulated data points from a Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $\Delta = \frac{1}{12}$ and $x_0 = 1$. Other details are the same as in Table 1.

The results in Tables 7 and 8 show that *epsilon* performs best in this case. When comparing these results with Table 2, changing the sample size does not appear to have an effect on the relative performance of the four approximate methods in the example for the range of sample sizes considered.

A4: Breakdown of the approximations with large sampling intervals

The approximations to the transition densities are based on a sampling scheme that re-

		Exact	epsilon	AS	SO	Kessler
θ_1	median $\lvert\hat{\theta}-\theta_0\rvert$	0.03014	0.03014	0.03015	0.03018	0.03008
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.05000	0.05000	0.04991	0.05000	0.04997
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00001	0.00002	0.00004	0.00125
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00002	0.00018	0.00006	0.00187
θ_2	median $\lvert\hat{\theta}-\theta_0\rvert$	0.08826	0.08826	0.08826	0.08832	0.08804
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.13546	0.13546	0.13546	0.13546	0.13545
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00002	0.00002	0.00003	0.00031
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00004	0.00004	0.00005	0.00058
θ_3	median $\lvert\hat{\theta}-\theta_0\rvert$	0.00207	0.00207	0.00207	0.00207	0.00206
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.00316	0.00316	0.00316	0.00316	0.00316
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00000	0.00007
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.00000	0.00000	0.00000	0.00010
Time (sec)		27.55245	0.41493	2.49948	0.48533	0.36967

Table 8: Performance of the various estimators, specified in §5.1, based on a trajectory of 2000 simulated data points from a Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.2, 3, 0.2)$, $\Delta = \frac{1}{12}$ and $x_0 = 1$. Other details are the same as in Table 1.

quires $\Delta \to 0$. We now consider a Cox-Ingersoll-Ross process with a larger value of Δ, $\Delta = 1$, where the approximation methods begin to break down.

		Exact	epsilon	AS	SO
θ_1	median $\lvert\hat{\theta}-\theta_0\rvert$	0.05399	0.06155	0.09381	0.05251
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.08093	0.13050	0.19402	0.07718
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.07567	0.08094	0.01715
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.13189	0.17987	0.01893
θ_2	median $\lvert\hat{\theta}-\theta_0\rvert$	0.00663	0.01133	0.00726	0.00720
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.00969	0.01583	0.01078	0.01043
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.01051	0.00257	0.00382
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.01272	0.00509	0.00389
θ_3	median $\lvert\hat{\theta}-\theta_0\rvert$	0.00716	0.02136	0.00847	0.00712
	rms $\lvert\hat{\theta}-\theta_0\rvert$	0.01063	0.02424	0.01606	0.01031
	median $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.02123	0.00435	0.00174
	rms $\lvert\hat{\theta}-\hat{\theta}_{MLE}\rvert$		0.02313	0.01298	0.00186
Time (sec)		13.03974	0.18826	2.28452	0.23206

Table 9: Performance of the various estimators, specified in §5.1, based on a trajectory of 1000 simulated data points from a Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (1, 1, 0.3)$, $\Delta = 1$ and $x_0 = 2$. Other details are the same as in Table 1.

No results are given for *Kessler* in Table 9 because the variance of the normal approximation is negative. In Figure 9, the *AS* approximation to the transition density is negative for a large part of the range; the epsilon approximation also goes negative, though for a smaller part of the range. Numerical results for the estimators are given in Table 9. Recall that we take the absolute value of the approximate transition densities when they go negative. In the results in Table 9, *SO* is the best performing method, followed by *epsilon*. These approximations still produce reasonably accurate results without any optimisation problems. Out of the 5000 Monte Carlo runs for the *AS* method, there are 829 simulations where the optimisation method fails to converge due to running out of function evaluations, and 345 simulations where the esti-

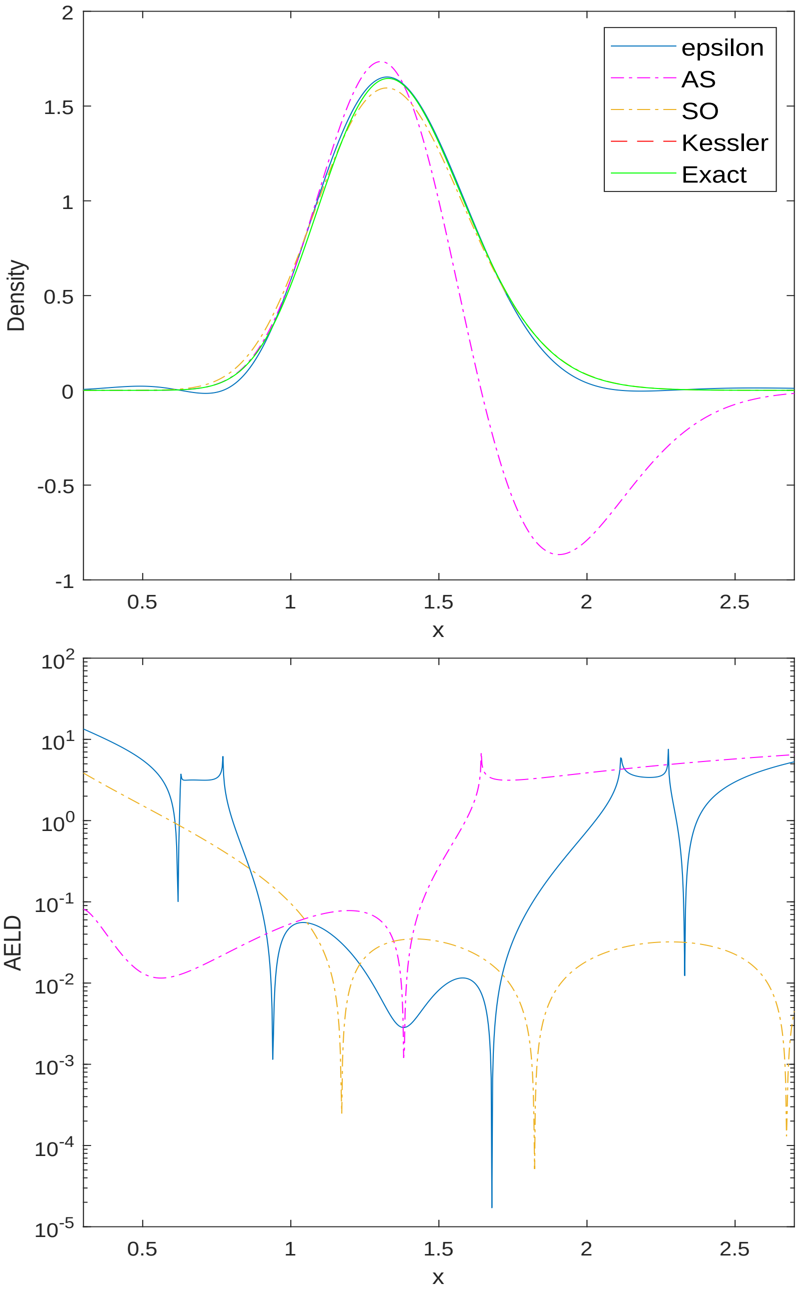

Figure 9: (Top) Approximations of the transition density, specified in §5.1, of the Cox-Ingersoll-Ross process (5.1) and (bottom) corresponding absolute error of the log densities. The model parameters are $(\theta_1, \theta_2, \theta_3) = (1, 1, 0.3)$, $x_0 = 2$ and $\Delta = 1$.

mator appears to diverge, defined as having at least one of the three parameter estimates with $|\hat{\theta}_i| > 100$. These unsuccessful simulation runs have been excluded from the *AS* results.

A5: Additional example for the distribution of estimates

Similar to §5.3, we now consider the distribution of the estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ based on 5000 Monte Carlo runs for the Cox-Ingersoll-Ross process (5.1) given in Table 1, where we saw that *AS* was the best performing approximation method. As before, the histograms (Figure 10), the normal QQ plots (Figure 11), and the table of sample moments (Table 10) show that estimates are slightly skewed for $\hat{\theta}_1$ and $\hat{\theta}_2$, approximately normal for $\hat{\theta}_3$, and the distribution of the estimates for all the approximation methods appear to be virtually identical to that of MLE, though the variance of $\hat{\theta}_1$ is slightly higher for *Kessler*. The histogram of the absolute difference between the estimates of the approximation methods and MLE (Figure 12) again shows much lower variance for *epsilon* than *Kessler*, but *AS* gives lower variance than *epsilon*.

		Exact	**epsilon**	**AS**	**SO**	**Kessler**
$\hat{\theta}_1$	mean	0.55055	0.55082	0.55034	0.54904	0.54879
	variance	0.01467	0.01459	0.01470	0.01475	0.01579
	skewness	0.61674	0.61285	0.61614	0.61097	0.60626
	kurtosis	3.65202	3.66184	3.64881	3.63846	3.59933
$\hat{\theta}_2$	mean	0.05995	0.05995	0.05995	0.05994	0.05994
	variance	0.00006	0.00006	0.00006	0.00006	0.00006
	skewness	0.39731	0.39747	0.39728	0.39705	0.39844
	kurtosis	3.21092	3.21164	3.21076	3.21029	3.21660
$\hat{\theta}_3$	mean	0.15015	0.15015	0.15015	0.15014	0.15019
	variance	0.00001	0.00001	0.00001	0.00001	0.00001
	skewness	0.00397	0.00074	0.00408	0.00391	0.01778
	kurtosis	3.08384	3.08152	3.08407	3.08455	3.14358

Table 10: Various sample moments of 5000 estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ for the Cox-Ingersoll-Ross process (5.1) with $(\theta_1, \theta_2, \theta_3) = (0.5, 0.06, 0.15)$, $\Delta = \frac{1}{12}$ and $x_0 = 0.1$.

B: Proofs

B1: Proof of Theorem 3.1

The proof of Theorem 3.1 is based on the following three lemmas. In what follows, $\mu_0 = \mu_Y(y_0, \theta_0)$, and $\mu_0^{(j)} = \partial_y^j \mu_Y(y_0, \theta_0)$, $j = 1, 2, 3$, is the jth partial derivative of $\mu_Y(y, \theta_0)$ evaluated at $y = y_0$. Note also that, from (3.14) and the fact that $y_\Delta = \gamma(x_\Delta, \beta)$ and $y_0 = \gamma(x_0, \beta)$, we have the relationship

$$\bar{y}_\Delta = \Delta^{-1/2} \left(y_\Delta - y_0 - \Delta C_0 - \Delta^2 C_{00} - \Delta^3 C_{000} \right), \tag{B1}$$

where C_0, C_{00} and C_{000} are defined in (3.8).

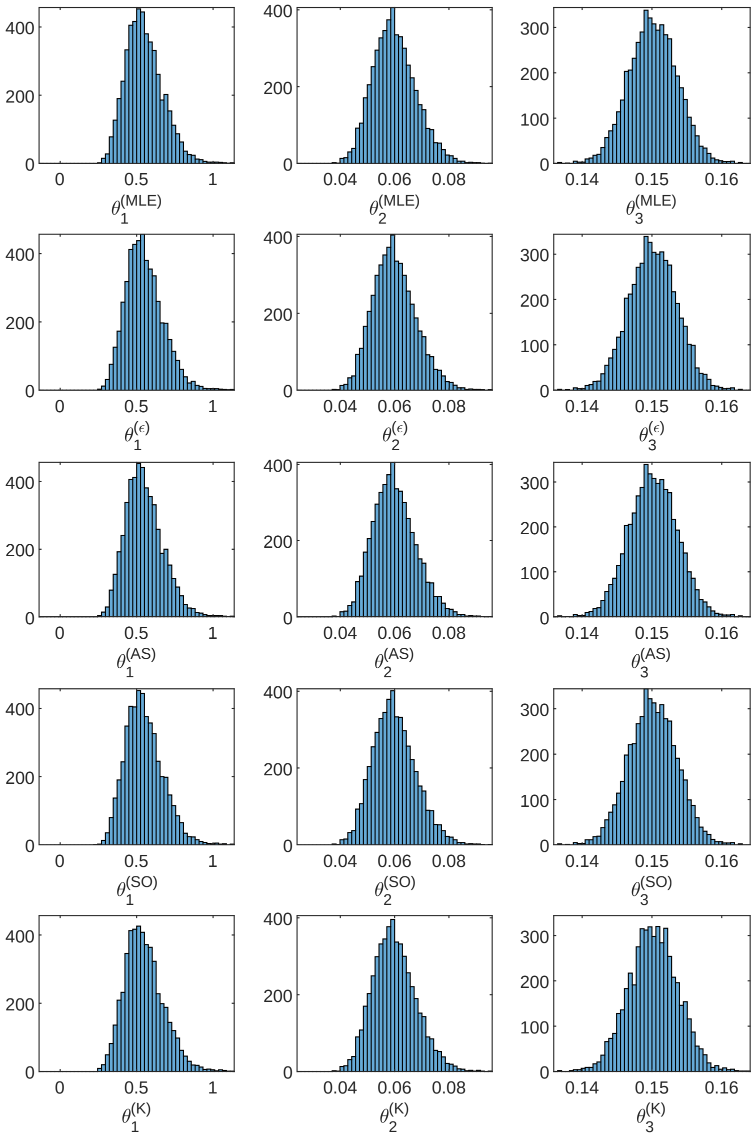

Figure 10: Histogram of 5000 estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ for the Cox-Ingersoll-Ross process (5.1) with model parameters $(\theta_1, \theta_2, \theta_3) = (0.5, 0.06, 0.15)$, $x_0 = 0.1$ and $\Delta = \frac{1}{12}$. The plots from top to bottom are for the exact, *epsilon*, *AS*, *SO* and *Kessler* methods. The *y*-axis is frequency.

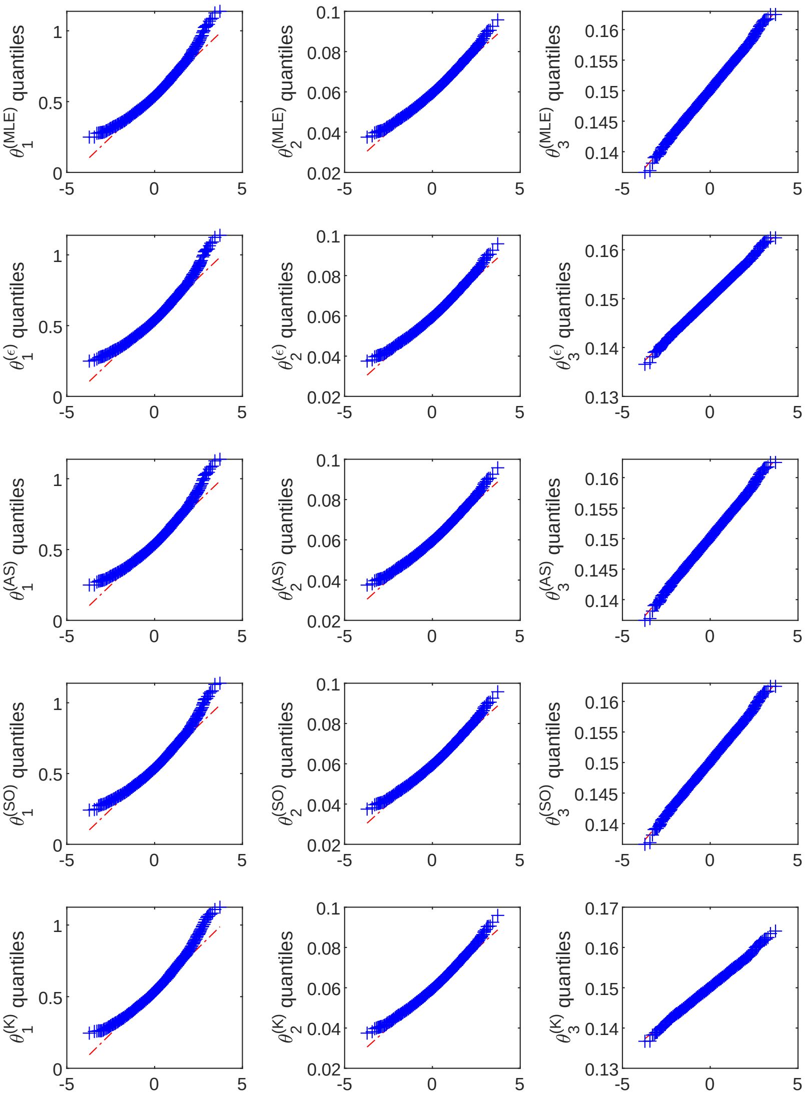

Figure 11: Normal QQ plots of 5000 estimates $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3)$ for the Cox-Ingersoll-Ross process (5.1) with model parameters $(\theta_1, \theta_2, \theta_3) = (0.5, 0.06, 0.15)$, $x_0 = 0.1$ and $\Delta = \frac{1}{12}$. The plots from top to bottom are for the exact, *epsilon*, *AS*, *SO* and *Kessler* methods. The x-axis are the standard normal quantiles.

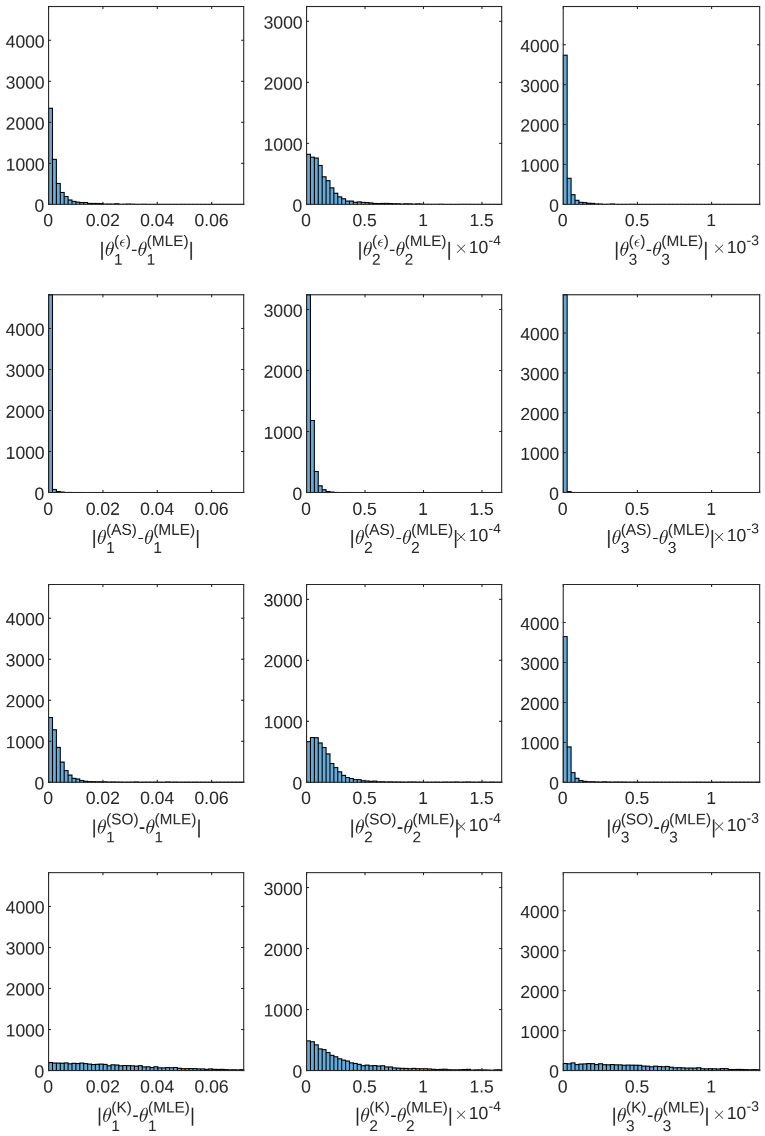

Figure 12: Histogram of 5000 absolute differences $|\hat{\theta} - \hat{\theta}_{MLE}|$ for the Cox-Ingersoll-Ross process (5.1) with model parameters $(\theta_1, \theta_2, \theta_3) = (0.5, 0.06, 0.15)$, $x_0 = 0.1$ and $\Delta = \frac{1}{12}$. The plots from top to bottom are for the exact, *epsilon*, *AS*, *SO* and *Kessler* methods. The y-axis is frequency.

Lemma B.1. *Recall the definition of $g(.)$ in (2.6). Under the assumptions of Theorem 3.1, as $\Delta \to 0$,*

$$
\begin{aligned}
E & \left[\exp \left\{ \Delta \int_0^1 g\left(y_0 + u(y_\Delta - y_0) + \Delta^{1/2} B_u \right) du \right\} \right] \\
& = 1 - \frac{\Delta}{2} \left(\mu_0^2 + \mu_0' \right) + \frac{\Delta^2}{24} \left\{ 3\mu_0^4 + \left(\mu_0' \right)^2 - 6\mu_0^2 \mu_0' - 8\mu_0 \mu_0^{(2)} - \mu_0^{(3)} \right\} \\
& \quad - \frac{\Delta^{3/2}}{4} \left(2\mu_0 \mu_0' + \mu_0^{(2)} \right) \bar{y}_\Delta - \frac{\Delta^2}{12} \left\{ 2\left(\mu_0' \right)^2 + 2\mu_0 \mu_0^{(2)} + \mu_0^{(3)} \right\} \bar{y}_\Delta^2 + \Delta^{5/2} R_1,
\end{aligned}
$$

where $R_1 \equiv R_1(\theta, y_0, \bar{y}_\Delta)$ satisfies

$$
\sup_{\theta \in \Theta} \sup_{y_0 \in \mathbb{R}} |R_1(\theta, y_0, \bar{y}_\Delta)| \leq P_1(\bar{y}_\Delta),
$$

for some fixed polynomial $P_1(\bar{y}_T)$.

Proof. From condition $(\mathrm{A2})_0$, it follows that g has three derivatives which are continuous and bounded, so we may use a 4-term Taylor expansion, including an exact remainder term; we omit the details of the calculations, which are long but straightforward. Moreover, the largest power of $\bar{Y}_\Delta$ that appears in the remainder term $R_1(\theta, \bar{y}_\Delta, y_0)$ is $\bar{y}_\Delta^3$. Consequently, Lemma B.1 holds because $|R_1|$ is bounded above by a polynomial of the form $P_1(\bar{y}_\Delta) = K_1 \left(1 + \bar{y}_\Delta^4\right)$, for some suitable constant K_1 independent of θ and y_0. □

In the next lemma, $\sigma(x_\Delta, \beta)$ is the diffusion coefficient in (1.1). However, as we are focusing on the density of $z_\Delta = (x_\Delta - x_0)/\Delta^{1/2}$ rather than x_Δ, we should think of $x_\Delta = x_0 + \Delta^{1/2} z_\Delta$ as being a function of z_Δ with x_0 fixed.

Lemma B.2. *Under the assumptions of Theorem 3.1, as $\Delta \to 0$,*

$$
\frac{1}{\{2\pi\sigma(x_\Delta, \beta)\}^{1/2}} \exp \left\{ -\frac{(y_\Delta - y_0)^2}{2\Delta} + \int_{y_0}^{y_\Delta} \mu(u) du \right\} \tag{B2}
$$

$$
\begin{aligned}
= \frac{1}{\{2\pi\sigma(x_\Delta, \beta)\}^{1/2}} & \exp\left(-\frac{1}{2} \bar{y}_\Delta^2 \right) \left[1 + \frac{\Delta}{2} \mu_0^2 + \frac{\Delta^2}{8} \left(\mu_0^4 + 4\mu_0^2 \mu_0' \right) + \frac{\Delta^{3/2}}{2} \left(\mu_0 \mu_0' - \frac{1}{2} \mu_0^{(2)} \right) \bar{y}_\Delta \right. \\
& \frac{\Delta}{2} \left\{ \mu_0' + \Delta \left(\mu_0 \mu_0^{(2)} + \frac{1}{2} \mu_0^2 \mu_0' \right) \right\} \bar{y}_\Delta^2 + \frac{\Delta^{3/2}}{6} \mu_0^{(2)} \bar{y}_\Delta^3 \\
& \left. + \frac{\Delta^2}{24} \left\{ 3\left(\mu_0' \right)^2 + \mu_0^{(3)} \right\} \bar{y}_\Delta^4 + \Delta^{5/2} R_2 \right],
\end{aligned} \tag{B3}
$$

where $R_2 \equiv R_2(\theta, y_0, \bar{y}_\Delta)$ satisfies

$$
\sup_{\theta \in \Theta} \sup_{y_0 \in \mathbb{R}} |R_2(\theta, y_0, \bar{y}_\Delta)| \leq P_2(\bar{y}_\Delta),
$$

for some fixed polynomial $P_2(\bar{y}_\Delta)$.

Proof. The proof is similar to that of Lemma B.1; in this case it follows from (B1) that we should expand

$$
\exp \left\{ \frac{1}{2} \left(\bar{y}_\Delta + \Delta^{1/2} C_0 + \Delta^{3/2} C_{00} + \Delta^{5/2} C_{000} \right)^2 + \int_{y_0}^{y_\Delta} \mu(u) du \right\}.
$$

After some straightforward but lengthy calculations using Taylor's expansion again, this leads to the expression on the right side of (B3). Moreover, in this case the highest power of $\bar{y}_\Delta$

that appears in the remainder term $R_2(\theta, \bar{y}_\Delta, y_0)$ is $\bar{y}_\Delta^5$. Consequently, the absolute value of the remainder term $|R_2|$ is bounded above by a polynomial of the form $P_2(\bar{y}_\Delta) = K_2\left(1 + \bar{y}_\Delta^6\right)$, where K_2 is a suitable constant independent of θ and y_0. □

We now combine Lemma B.1 and Lemma B.2 to obtain the following. Note that the right side in Lemma B.3 is the factor in the square brackets on the right side of (3.15).

Lemma B.3. *Using the definition of the C_α given in (3.8) we have*

$$\begin{aligned}
&1 + \frac{\Delta}{2} C_{10} H_2(\bar{y}_\Delta) + \frac{\Delta^{3/2}}{6} C_{110} H_3(\bar{y}_\Delta) \\
&\quad + \frac{\Delta^2}{24}\left\{\left(3C_{10}^2 + C_{1110}\right) H_4(\bar{y}_\Delta) + 4\left(C_{100} + C_{010} + C_{10}^2\right) H_2(\bar{y}_\Delta)\right\} \\
&= 1 - \frac{\Delta}{2}\mu_0' + \frac{\Delta^2}{24}\left\{\left(\mu_0'\right)^2 - 8\mu_0\mu_0^{(2)} - \mu_0^{(3)}\right\} - \frac{\Delta^{3/2}}{2}\mu_0^{(2)}\bar{y}_\Delta \\
&\quad + \left\{\frac{\Delta}{2}\mu_0' - \frac{\Delta^2}{12}\left(5\left(\mu_0'\right)^2 + \mu_0^{(3)} - 4\mu_0\mu_0^{(2)}\right)\right\}\bar{y}_\Delta^2 \\
&\quad \frac{\Delta^{3/2}}{6}\mu_0^{(2)}\bar{y}_\Delta^3 + \frac{\Delta^2}{24}\left\{3\left(\mu_0'\right)^2 + \mu_0^{(3)}\right\}\bar{y}_\Delta^4.
\end{aligned}$$

Proof. To prove Lemma B.3 we simply write the Hermite polynomials H_2, H_3 and H_4 explicitly as polynomials and then write the right side as a polynomial in $\bar{y}_\Delta$. □

To prove the first part of Theorem 3.1, we note that the product of the left side of the expression in Lemma B.1 and the left side of the expression in Lemma B.2 gives (2.5). Moreover, it is straightforward to show that multiplying the right side of the expression in Lemma B.1 by the right side of the expression in Lemma B.2 gives

$$\frac{1}{\{2\pi\sigma(x_\Delta, \theta)\}^{1/2}} \exp\left(-\frac{1}{2}\bar{y}_\Delta^2\right)\left\{A(\bar{y}_\Delta) + \Delta^{5/2} R_3\right\},$$

where $A(\bar{y}_\Delta)$ is equal to the right side, and therefore the left side, of the expression in Lemma B.3; and, in addition,

$$\sup_{\theta\in\Theta}\sup_{y_0\in\mathbb{R}} |R_3(\theta, y_0, \bar{y}_\Delta)| \le P_3(\bar{y}_\Delta),$$

where $P_3(\bar{y}_\Delta)$ can be expressed in terms of $P_1(\bar{y}_\Delta)$ and $P_2(\bar{y}_\Delta)$. Consequently, (3.15) contains all terms of order up to and including Δ^2, and part (i) of Theorem 3.1 is proved. Part (ii) of Theorem 3.1 follows because the left side of (3.16) is bounded above by

$$\sup_{\bar{y}_\Delta\in\mathbb{R}} \phi(\bar{y}_\Delta)\Delta^{5/2} P_3(\bar{y}_\Delta) = O(\Delta^{5/2}),$$

so now the proof of Theorem 3.1 is complete. □

B2: Proof of Theorem 4.2

We begin by recalling an elementary result: for any sequence of real-valued random vectors $(W_n)_{n\ge1}$ with a finite covariance matrix Σ_n and therefore well-defined mean μ_n for each n,

$$||W_n|| = O_p(||\mu_n|| + ||\Sigma_n||^{1/2}). \tag{B4}$$

We recall the definitions of the exact log-likelihood $\ell_0(\theta)$, given in (4.5); and the approximate log-likelihood, or contrast, function, $\ell_K(\theta)$, given in (4.8) and due to Kessler (1997). Define $\ell_1(\theta) = \ell_K(\theta) - \ell_0(\theta)$, the difference between the Kessler (1997) contrast function and the exact log-likelihood. Below, for ν a finite sequence of zeros and ones, we write $C_{\nu,i-1}$ for C_ν evaluated at time t_{i-1}, where the C_ν's that we need are defined in (3.8). Similarly, we write

$$\bar{y}_i = \Delta^{-1/2}\left(\gamma(x_{i-1} + \Delta^{1/2} z_i, \beta) - \gamma(x_{i-1}, \beta) - \Delta C_{0,i-1} - \frac{1}{2}\Delta^2 C_{00,i-1} - \frac{1}{6}\Delta^3 C_{000,i-1}\right),$$

where $z_i = (x_i - x_{i-1})/\Delta^{1/2}$; cf (3.14). Then it turns out that

$$\begin{aligned}\ell_1(\theta) = -\sum_{i=1}^{n}\Bigg[&\frac{1}{6}\Delta^{3/2} C_{110,i-1} H_3(\bar{y}_i) + \frac{1}{24}\Delta^2 C_{1110,i-1} H_4(\bar{y}_i)\\ &- \frac{1}{24}\Delta^{5/2} C_{10,i-1} C_{110,i-1} H_2(\bar{y}_i) H_3(\bar{y}_i) + \Delta^3 R_{K,i}\Bigg],\end{aligned} \tag{B5}$$

where $R_{K,i} = R_{K,i}(\theta) = O_p(1)$ uniformly in i. In (B5) we have picked out those leading terms which appear in $\ell_0(\theta)$ but not in $\ell_K(\theta)$, and put the remainders into the R_i. Write $S_{11}(\theta_0) = S_{1\alpha}(\theta_0)$ and $S_{12}(\theta_0) = S_{1\beta}(\theta_0)$. From (B5),

$$\begin{aligned}S_1(\theta_0) &= \frac{\partial \ell_1}{\partial\theta}\bigg|_{\theta=\theta_0}\\ &= -\sum_{i=1}^{n}\Bigg[\frac{1}{6}\Delta^{3/2}\left(\frac{\partial C_{110,i-1}}{\partial\theta} H_3(\bar{y}_i) + 3C_{110,i-1} H_2(\bar{y}_i)\frac{\partial \bar{y}_i}{\partial\theta}\right)\\ &+ \frac{1}{24}\Delta^2\left(\frac{\partial C_{1110,i-1}}{\partial\theta} H_4(\bar{y}_i) + 4C_{1110,i-1} H_3(\bar{y}_i)\frac{\partial \bar{y}_i}{\partial\theta}\right)\\ &- \frac{1}{24}\Delta^{5/2}\bigg\{\left(\frac{\partial C_{10,i-1}}{\partial\theta} C_{110,i-1} + C_{10,i-1}\frac{\partial C_{110,i-1}}{\partial\theta}\right) H_2(\bar{y}_i) H_3(\bar{y}_i)\\ &\quad + C_{10,i-1} C_{110,i-1}\left(2H_1(\bar{y}_i) H_3(\bar{y}_i) + 3H_2(\bar{y}_i)^2\right)\frac{\partial \bar{y}_i}{\partial\theta}\bigg\} + \Delta^3 \frac{\partial R_{K,i}}{\partial\theta}\Bigg],\end{aligned} \tag{B6}$$

where the C's, the $\bar{y}_i$'s and the $\partial\bar{y}_i/\partial\theta$'s are all evaluated at $\theta = \theta_0$. Writing

$$S_1 = (S_{11}^\top, S_{12}^\top) = (S_{1\alpha}^\top, S_{1\beta}^\top)^\top = \left(\frac{\partial \ell_1}{\partial\alpha^\top}, \frac{\partial \ell_1}{\partial\beta^\top}\right)^\top,$$

we now consider $S_{1\alpha}(\theta_0)$ and focus on $S_{1\beta}(\theta_0)$ later, where $S_{1\alpha}$ and $S_{1\beta}$ are the α and β components of S_1. In the case of $S_{1\alpha}$, it is easily checked that $||\partial\bar{y}_i/\partial\alpha|| = O(\Delta^{1/2})$. Moreover, under the assumed model at $\theta = \theta_0$, if follows from Theorem 3.1 and, in particular (3.15), that $|E[H_2(\bar{Y}_i)]| = O(\Delta)$, $|E[H_3(\bar{Y}_i)]| = O(\Delta^{3/2})$, and consequently

$$\left|\left|\Delta^{3/2} E\left[\left(\frac{\partial C_{110,i-1}}{\partial\alpha} H_3(\bar{y}_i) + 3C_{110,i-1} H_2(\bar{y}_i)\frac{\partial \bar{y}_i}{\partial\alpha}\right)\right]\right|\right| = O(\Delta^{3/2}\Delta^{3/2}) = O(\Delta^3),$$

uniformly over i. Applying similar calculations to the other terms on the right-hand side of (B6), we find that $||E[S_{1\alpha}(\theta_0)]|| = O(n\Delta^3)$. Also, applying Proposition 6.2 to the elements of $S_{1\alpha}$ we find that $||\Delta^{3/2}\mathrm{Cov}(S_{1\alpha})||^{1/2} = O(\Delta^{3/2} n^{1/2}\Delta^{-1/2}) = O(n^{1/2}\Delta)$. Therefore from (B4), $||S_{1\alpha}(\theta_0)|| = O_p(n\Delta^3 + n^{1/2}\Delta)$ and consequently, recalling that in the present setting

$\delta_{n,1} = (n\Delta)^{1/2}$,

$$\begin{aligned}
\delta_{n,1}^{-1}||S_{1\alpha}(\theta_0)|| + O_p(\delta_{n,1}^{-1}) &= O_p\{(n\Delta^5)^{1/2} + \Delta^{1/2} + (n\Delta)^{-1/2}\} \qquad \text{(B7)}\\
&= O_p\{(n\Delta)^{-1/2}(1 + n^{1/2}\Delta + n\Delta^3)\}\\
&= O_p\{(n\Delta)^{-1/2} + \Delta^{1/2} + (n\Delta^5)^{1/2}\}.
\end{aligned}$$

Note that the second term on the left side of (B7) corresponds to the third term on the right side of (B7) and also the third term on the right side of (6.9). So (B7) agrees with the first part of Theorem 4.2. To complete the proof of the first part of Theorem 4.2, we need to show that, in part (ii) of Proposition 6.1, the middle term on the right-hand side of (6.9) is of the same order as (B7). This is indeed the case because, due to the uniformity established in Theorem 3.1, it is seen by inspection of (B6) that

$$||J_{1\alpha}^{**}|| = \left|\left|\frac{\partial^2 \ell_1}{\partial\alpha\partial\alpha^\top}(\theta^{**})\right|\right| = O_p(n\Delta^{3/2}),$$

where, in view of the discussion at the beginning of §4 concerning the exact form of Taylor's theorem for vector valued functions, θ_1^{**} is different for each row of $\partial^2\ell_1/\partial\alpha\partial\alpha^\top$. Consequently, bearing in mind that $\delta_{n,1} = (n\Delta)^{1/2}$,

$$||\delta_{n,1}^{-2}J_{1\alpha}^{**}|| = \delta_{n,1}^{-2}O_p(n\Delta^{3/2}) = O_p(\Delta^{1/2}),$$

which is of order (B7), and so (4.11) is proved.

The proof of (4.13) is similar to that of (4.11) but with the following differences: in place of $\delta_{n,1} = (n\Delta)^{1/2}$ we use $\delta_{n,2} = n$; and $\partial\bar{y}_i/\partial\beta = O_p(1)$ rather than $O_p(\Delta^{1/2})$. In this case,

$$\left|\left|\Delta^{3/2}E\left[\left(\frac{\partial C_{110,i-1}}{\partial\beta}H_3(\bar{y}_i) + 3C_{110,i-1}H_2(\bar{y}_i)\frac{\partial\bar{y}_i}{\partial\beta}\right)\right]\right|\right| = O(\Delta^{3/2}\Delta) = O(\Delta^{5/2}).$$

Moreover, with further, similar calculations whose details we omit, it can be shown that $E[S_{1\beta}] = O(n\Delta^{5/2})$. So, using Proposition 6.2 several times, we obtain

$$||\delta_{n,2}^{-1}S_{1\beta}|| = O_p\{n^{-1/2}(n\Delta^{5/2} + \Delta^{3/2}(n/\Delta)^{1/2})\} = O_p\{(n\Delta^5)^{1/2} + \Delta\}. \qquad \text{(B8)}$$

Also, recalling that $\delta_{n,2} = n^{1/2}$, the second and third terms on the right side of (6.9) are, respectively,

$$O_p(n^{-1}n\Delta^{3/2}) = O(\Delta^{3/2}) \quad \text{and} \quad O_p\{(n\Delta)^{-1/2}\}. \qquad \text{(B9)}$$

Finally, using (6.9), (B8) and (B9), it is seen that

$$n^{1/2}||\hat{\beta}_K - \hat{\beta}_{MLE}|| = O_p\{(n\Delta)^{-1/2} + \Delta + (n\Delta^5)^{1/2}\},$$

as required.

We now sketch the proof of parts (b) and (d). First, we consider approximation of the exact density $p_Z(z_\Delta|x_0;\theta)$ in by a higher-order ϵ-expansion as follows:

$$\begin{aligned}
\hat{p}_Z(z_\Delta|x_0;\theta) = \frac{1}{\sigma(x_0 + \Delta^{1/2}z_\Delta, \theta)}\phi(\bar{y}_\Delta)\Bigg[1 &+ \frac{1}{2}\Delta C_{10}H_2(\bar{y}_\Delta) + \frac{1}{6}\Delta^{3/2}C_{110}H_3(\bar{y}_\Delta)\\
&+ \Delta^2\left\{\frac{1}{8}C_{10}^2H_4(\bar{y}_\Delta) + \frac{1}{6}(C_{100} + C_{010})H_2(\bar{y}_\Delta) + \frac{1}{24}C_{1110}H_4(\bar{y}_\Delta)\right\}\\
&+ \Delta^{5/2}\sum_{j=1}^{2}A_{5,2j+1}H_{2j+1}(\bar{y}_\Delta) + \Delta^3\sum_{j=1}^{3}A_{6,2j}H_{2j}(\bar{y}_\Delta)\\
&+ \Delta^{7/2}\sum_{j=1}^{4}A_{7,2j-1}H_{2j-1}(\bar{y}_\Delta) + \Delta^4R_\epsilon. \qquad \text{(B10)}
\end{aligned}$$

A key point to note is that the term $A_{6,0}$, the coefficient of $H_0(\bar{y}_\Delta) = 1$, is zero, so that a constant term of order Δ^3 does not appear in the above. This is a consequence of the structure of the typical term in the ϵ-expansion given by (3.5). In particular, if we put $g_0(x) = \phi(x)$ and integrate a typical term on the left side of (3.5) with respect to the standard Gaussian density $\phi(x)$, with any fixed $r \geq 1$, the result is zero provided $\phi(x)\kappa_{a_1 \ldots a_r}(x) \to 0$ as $|x| \to \infty$, as is the case here.

Following the proofs of (a) and (c), we re-define $\ell_1(\theta)$ by $\ell_1^*(\theta) = \ell^{(\epsilon)}(\theta) - \ell_0(\theta)$, where $\ell^{(\epsilon)}(\theta)$ and $\ell_0(\theta)$ are defined in (4.10) and (4.5) respectively. In this case, the analogue of (B5) is

$$\ell_1^*(\theta) = -\sum_{i=1}^{n} \Bigg[\Delta^{5/2} \sum_{j=1}^{2} A_{5,2j+1} H_{2j+1}(\bar{y}_\Delta) + \Delta^3 \sum_{j=1}^{3} A_{6,2j} H_{2j}(\bar{y}_\Delta) + \Delta^{7/2} \sum_{j=1}^{4} A_{7,2j-1} H_{2j-1}(\bar{y}_\Delta) \\ - \frac{1}{4}\Delta^{7/2} C_{10} \sum_{j=1}^{2} A_{5,2j+1} H_{2j+1}(\bar{y}_\Delta) + \Delta^4 R_{\epsilon,i} \Bigg], \tag{B11}$$

where all terms of order Δ^4 or smaller have been absorbed in the remainder term $R_{\epsilon,i}$.

The remaining steps in the proofs for cases (b) and (d) closely parallel those for parts (a) and (c) respectively and we omit the details.

□

B3: Proof of Proposition 6.1

We first prove part (i) concerning consistency of $\hat{\theta}_0$ and $\hat{\theta}_1$ under the stated assumptions. Using (6.2), and writing $(J_0^*)^{-1}$ in block form $(J_0^*)^{-1} = \left(J_0^{*jk}\right)_{j,k=1}^{r}$, where block J_0^{*jk} has dimensions $d_j \times d_k$ in the notation of §6.1, we have

$$\hat{\theta}_{0j} - \theta_{0j} = \delta_{n,j}^{-1} \sum_{k=1}^{r} \delta_{n,j}\delta_{n,k} J_0^{*jk} \delta_{n,k}^{-1} S_{0k}(\theta_0).$$

Since by assumption $\mathcal{I}(\theta)$ is strictly positive definite in a neighbourhood of θ_0, if follows from (6.6) and the continuous mapping theorem that $\delta_{n,j}\delta_{n,k}J_0^{*jk}$ converges in probability to $\mathcal{I}^{jk}(\theta_0)$, which is block (j,k) of the inverse of $\mathcal{I}(\theta_0)$. Therefore, from the first statement in (6.5) combined with Slutsky's Lemma, we conclude that

$$\begin{aligned} ||\hat{\theta}_{0j} - \theta_{0j}|| &= \delta_{n,j}^{-1} \left|\left| \sum_{k=1}^{r} \delta_{n,j}\delta_{n,k} J_0^{*jk} \delta_{n,k}^{-1} S_{0k}(\theta_0) \right|\right| \\ &\leq \delta_{n,j}^{-1} \sum_{k=1}^{r} \left(||\mathcal{I}^{jk}(\theta_0)||_\infty + ||\delta_{n,j}\delta_{n,k}J_0^{*jk} - \mathcal{I}^{jk}(\theta_0)||_\infty \right) ||\delta_{n,k}^{-1} S_{0k}(\theta_0)|| \\ &= \delta_{n,j}^{-1} \sum_{k=1}^{r} O_p(1) \\ &= O_p(\delta_{n,j}^{-1}), \end{aligned}$$

as required. In the above, $||A||_\infty$ is equal to the absolute value of the element of the matrix A with largest absolute value.

The proof that $\hat{\theta}_1$ is a consistent estimator of θ_0 is similar to the above but has as starting point (6.3), and makes use of the second statement in (6.5) and (6.7) to take care of the terms $S_1(\theta_0)$ and J_1^{**} respectively.

For part (ii), we recall the result that for a $q \times q$ matrix A with $||A||$ small,

$$(I_q + A)^{-1} = I_q - A + B,$$

where $||B|| = O(||A||^2)$. Using this result, combined with (6.4), (6.6) and (6.7), we obtain

$$D_n(J_0^{**} + J_1^{**})^{-1}D_n = D_n(J_0^{**})^{-1}D_n - D_n(J_0^{**})^{-1}D_n(D_n^{-1}J_1^{**}D_n^{-1})D_n(J_0^{**})^{-1}D_n + R, \qquad \text{(B12)}$$

where

$$||R|| = o_p\left(||D_n^{-1}J_1^{**}D_n^{-1}||^2\right).$$

Therefore, using (6.2), (6.3) and (B12),

$$\begin{aligned}
D_n(\hat{\theta}_1 - \hat{\theta}_0) &= D_n(\hat{\theta}_1 - \theta_0) - D_n(\hat{\theta}_0 - \theta_0) \\
&= D_n\left(J_0^{**} + J_1^{**}\right)^{-1} D_n D_n^{-1}\left\{S_0(\theta_0) + S_1(\theta_1)\right\} - D_n\left(J_0^{*}\right)^{-1} D_n D_n^{-1} S_0(\theta_0) \\
&= D_n\left\{(J_0^{**})^{-1} - (J_0^{*})^{-1}\right\} D_n D_n^{-1} S_0(\theta_0) + \left(D_n^{-1}J_0^{**}D_n^{-1}\right)^{-1} D_n^{-1}S_1(\theta_0) \\
&\qquad - \left(D_n^{-1}J_0^{**}D_n^{-1}\right)^{-1} D_n^{-1}J_1^{**}D_n^{-1}\left(D_n^{-1}J_0^{**}D_n^{-1}\right)^{-1} D_n^{-1}S_0(\theta_0) + R_0 \\
&= a + \mathcal{I}(\theta_0)^{-1}D_n^{-1}S_1(\theta_0) - \mathcal{I}(\theta_0)^{-1}D_n^{-1}J_1^{**}D_n^{-1}\mathcal{I}(\theta_0)^{-1}D_n^{-1}S_0(\theta_0) + R_1, \qquad \text{(B13)}
\end{aligned}$$

where

$$a = (a_1^\top, \ldots, a_r^\top)^\top = \Delta_n\left\{(J_0^{**})^{-1} - (J_0^{*})^{-1}\right\}\Delta_n\Delta_n^{-1}S_0(\theta_0)$$

is such that $||a|| = O_p(\delta_{n,1}^{-1})$, and R_1 is a remainder term such that

$$||R_1|| = o_p\left(||D_n^{-1}J_1^{**}D_n^{-1}|| + ||D_n^{-1}S_1(\theta_0)||\right)$$

Since $\mathcal{I}(\theta_0)$ is block diagonal by assumption, if we focus on sub-vector j of $\hat{\theta}_1 - \hat{\theta}_0$, for $j = 1, \ldots, r$, and use (6.5)-(6.8), we obtain (6.9). □

B4: Proof of Proposition 6.2

Under the model, any given (positive) moment of $H_k(\bar{Y}_i)$ is uniformly bounded over i . Now

$$\begin{aligned}
\text{Var}\left\{\sum_{i=1}^{n} U_{i-1}H_k(\bar{Y}_i)\right\} &= \left|\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}\{U_{i-1}H_k(\bar{Y}_i), U_{j-1}H_k(\bar{Y}_j)\}\right| \\
&\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\text{Cov}\{U_{i-1}H_k(\bar{Y}_i), U_{j-1}H_k(\bar{Y}_j)\}\right|.
\end{aligned}$$

Assuming $\sup_i E[|U_i|^{2+\epsilon}] = C_0 < \infty$, choose $p = q = 2 + \epsilon/2$ and $r = \epsilon/(4+\epsilon)$ in Result B in §6.2 to obtain

$$\begin{aligned}
|\text{Cov}\{U_{i-1}H_k(\bar{Y}_i), U_{j-1}H_k(\bar{Y}_j)\}| &\leq 8\left\{E[|U_{i-1}H_k(\bar{Y}_i)|^{2+\epsilon/2}\right\}^{1/(2+\epsilon/2)} \\
&\qquad \times \left\{E[|U_{j-1}H_k(\bar{Y}_j)|^{2+\epsilon/2}\right\}^{1/(2+\epsilon/2)} \alpha(\mathcal{G}, \mathcal{H})^{\epsilon/(4+\epsilon)}.
\end{aligned}$$

Applying Hölder's inequality to each of the expectations on the right side of the above expression with

$$p' = (2+\epsilon)/(2+\epsilon/2) \quad \text{and} \quad q' = (4+2\epsilon)/\epsilon,$$

we obtain

$$\begin{aligned}|\text{Cov}\{U_{i-1}H_k(\bar{Y}_i), U_{j-1}H_k(\bar{Y}_j)\}| &\le 8C_0^{2/(2+\epsilon)}\left\{E[|H_k(\bar{Y}_i)|^{(4+\epsilon)(2+\epsilon)/\epsilon}]\right\}^{\epsilon/\{(4+\epsilon)(2+\epsilon)\}}\\ &\quad\times\left\{E[|H_k(\bar{Y}_j)|^{(4+\epsilon)(2+\epsilon)/\epsilon}]\right\}^{\epsilon/\{(4+\epsilon)(2+\epsilon)\}}\\ &\qquad\times\alpha(\mathcal{G},\mathcal{H})^{\epsilon/(4+\epsilon)}\\ &\le C_1\alpha(\mathcal{G},\mathcal{H})^{\epsilon/(4+\epsilon)}, \qquad\text{(B14)}\end{aligned}$$

for some finite constant C_1, since any given moment of $H_k(\bar{Y}_i)$ is uniformly bounded over i. Now we use Result A in §6.2. Define $\lambda_0 = \lambda\epsilon/(4+\epsilon)$. Then, using (6.13) and (B14), and bearing in mind that $\Delta \to 0$ and $n \to \infty$,

$$\begin{aligned}\text{Var}\left\{\sum_{i=1}^n U_{i-1}H_k(\bar{Y}_i)\right\} &\le \sum_{i=1}^n\sum_{j=1}^n |\text{Cov}\{U_{i-1}H_k(\bar{Y}_i), U_{j-1}H_k(\bar{Y}_j)\}|\\ &\le C_1\sum_{i=1}^n\sum_{j=1}^n \alpha(|i\Delta - j\Delta|)^{\epsilon/(4+\epsilon)}\\ &\le C_1\sum_{i=1}^n\sum_{j=1}^n\left\{e^{-\lambda\Delta|i-j|}\right\}^{\epsilon/(4+\epsilon)}\\ &= C_1\sum_{i=1}^n\sum_{j=1}^n e^{-\lambda_0\Delta|i-j|}\\ &= C_1\sum_{r=0}^{n-1}(n-r)e^{-\lambda_0 r\Delta}\\ &\le nC_1\frac{1}{1-e^{-\lambda_0\Delta}}\\ &\le \frac{nC_2}{\lambda_0\Delta}\\ &= O(n\Delta^{-1}),\end{aligned}$$

as required, where C_2 is another constant. □.

B5: Derivation of formulae in (3.13).

To derive these expressions we make use of the Schepp (1982) representation for Brownian motion defined on $[0,1]$:

$$B_t = \eta_0 t + \sum_{n=1}^\infty \eta_n\frac{\sqrt{2}}{n\pi}\sin(n\pi t), \qquad t\in[0,1],$$

where $(\eta_n)_{n\ge 0}$ is an IID sequence of $N(0,1)$ variables. Note that $B_0 = 0$ and $B_1 = \eta_0$. Using formulae on p.208 and p.210 of Preston and Wood (2012), and putting $T = \Delta = 1$, we obtain

$$
\begin{aligned}
X_0 &\equiv \check{I}_{(1)} = \eta_0;\\
X_1 &\equiv \check{I}_{(1,0)} = \frac{1}{2}\eta_0 + 2\sqrt{2}\pi^{-2}\sum_{n=1}^{\infty}\eta_{2n-1}(2n-1)^{-2};\\
X_2 &\equiv \check{I}_{(1,1,0)} = \frac{1}{6}\eta_0^2 - \frac{1}{4} + \frac{1}{2\pi^2}\sum_{n=1}^{\infty}\eta_n^2 n^{-2} + \sqrt{2}\pi^{-2}\eta_0\sum_{n=1}^{\infty}\eta_n(-1)^{n+1}n^{-2};\\
X_3 &\equiv \check{I}_{(1,0,0)} = \frac{1}{2}\eta_0 + 2\sqrt{2}\pi^{-2}\sum_{n=1}^{\infty}\eta_{2n-1}(2n-1)^{-2} - \frac{1}{3}\eta_0 - \sqrt{2}\pi^{-2}\sum_{n=1}^{\infty}\eta_n(-1)^{n+1}n^{-2}\\
&= \frac{1}{6}\eta_0 + 2\sqrt{2}\pi^{-2}\sum_{n=1}^{\infty}\eta_{2n-1}(2n-1)^{-2} - \sqrt{2}\pi^{-2}\sum_{n=1}^{\infty}\eta_n(-1)^{n+1}n^{-2};\\
X_4 &\equiv \check{I}_{(0,1,0)} = \frac{1}{6}\eta_0 - 2\sqrt{2}\pi^{-2}\sum_{n=1}^{\infty}\eta_{2n}(2n)^{-2}.
\end{aligned}
$$

Bearing in mind that the η_n are IID $N(0,1)$, it follows immediately that

$$
\begin{aligned}
\kappa_1(x) &\equiv E[X_1|X_0 = x] = \frac{1}{2}x = \frac{1}{2}H_1(x),\\
\kappa_2(x) &\equiv E[X_2|X_0 = x] = \frac{1}{6} - \frac{1}{4} + \frac{1}{2\pi^2}\frac{\pi^2}{6} = \frac{1}{6}(x^2-1) = \frac{1}{6}H_2(x),\\
\kappa_3(x) &\equiv E[X_3|X_0 = x] = \frac{1}{6}x = \frac{1}{6}H_1(x),\\
\kappa_4(x) &\equiv E[X_4|X_0 = x] = \frac{1}{6}x = \frac{1}{6}H_1(x),\\
\kappa_{11}(x) &\equiv E[X_1^2|X_0 = x] = \frac{1}{4}x^2 + \frac{8}{\pi^4}\sum_{n=1}^{\infty}(2n-1)^{-4} = \frac{1}{4}x^2 + \frac{8}{\pi^4}\frac{\pi^4}{96} = \frac{1}{4}x^2 + \frac{1}{12},
\end{aligned}
$$

where for $\kappa_2(x)$ and $\kappa_{11}(x)$ we have used the identities

$$
\sum_{i=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6} \quad \text{and} \quad \sum_{n=1}^{\infty}\frac{1}{(2n-1)^4} = \frac{\pi^4}{96},
$$

both of which can be evaluated using contour integration combined with the "coth" construction.

Finally, we need to calculate $\kappa_5(x) \equiv E[X_5|X_0 = x]$ where $X_5 = \check{I}_{(1,1,1,0)}$. We know that that $\int_0^{s_2} dB_{s_3} = B_{s_2}$, and from Itô's formula applied with $f(y) = y^2$,

$$
\int_0^{s_1}\int_0^{s_2} dB_{s_3}dB_{s_2} = \frac{1}{2}\left(B_{s_1}^2 - s_1\right).
$$

Therefore, using Itô's formula again, this time applied with $f(x) = x^3/3$,

$$
\int_0^{s}\int_0^{s_1}\int_0^{s_2} dB_{s_3}dB_{s_2}dB_{s_1} = \int_0^{s}\frac{1}{2}\left(B_{s_1}^2 - s_1\right)dB_{s_1} = \frac{1}{6}B_s^3 - \frac{1}{2}\int_0^{s}B_{s_1}ds_1 - \frac{1}{2}\int_0^{s}s_1 dB_{s_1}. \tag{B15}
$$

The conditional expectation of the terms on the RHS of (B15) given $\eta_0 = x$ are found to be

$$
E\left[\frac{1}{6}B_s^3\middle|\eta_0 = x\right] = \frac{1}{6}s^3x^3 + \frac{1}{2}sx\sum_{n=1}^{\infty}\frac{2}{n^2\pi^2}\left\{\sin(n\pi s)\right\}^2,
$$

$$\frac{1}{2}\int_0^s B_{s_1} ds_1 = \frac{1}{4}s^2 x = \frac{1}{2}\int_0^s s_1 dB_{s_1}.$$

Substituting into the RHS of (B15) and integrating each term over $s \in [0, 1]$ and we obtain

$$\frac{1}{24}x^3 + x\left(\sum_{n=1}^{\infty} \frac{1}{n^2\pi^2}\int_0^1 s\left\{\sin(n\pi s)\right\}^2 ds\right) - \frac{1}{6}x = \frac{1}{24}H_3(x),$$

as required. □

B6: Some comments on asymptotic expansions.

Suppose that for $\Delta \to 0$ and for $\alpha > 0$ and smooth functions $f, f_r, \tilde{f}_r(\cdot; \Delta), r = 1, \ldots, k,$

$$f(x; \Delta) \sim \Delta^{\alpha} \sum_{r=0}^{k} f_r(x)\Delta^r + O(\Delta^{\alpha+k+1}) \tag{6.14}$$

and

$$f(x; \Delta) \sim \Delta^{\alpha} \sum_{r=0}^{k} \tilde{f}_r(x; \Delta)\Delta^r + O(\Delta^{\alpha+k+1}). \tag{6.15}$$

Due to the coefficient functions $f_r(x)$ in (6.14) not depending on Δ, the expansion (6.14) is unique. However, there are many (in fact an infinite number) of expansions of the form (6.15) which, when we Taylor-expand each $\tilde{f}_r(x, \Delta)$ about $\Delta = 0$ and then collect terms in powers of Δ^r, we end up with the expansion (6.14). Morever, expansions of the form (6.15) are often more accurate than expansions of the form (1); e.g. compare saddlepoint approximations and Edgeworth expansions. In the case of the *AS* and *epsilon* expansions, both are of the general form (6.15), but with different $\tilde{f}_r$ functions, so each needs to be expanded into the standard form (6.14). When this is done for the two expansions, they agree up to the relevant order. Consequently, *AS* may be seen as a representative of those expansions correct up to and including all Δ^3 terms (see Ait-Sahalia, 2002, Section 4); *epsilon* may be seen as a representative of those expansions correct up to and including all order Δ^2 terms; and the Kessler approximation may be seen as a representative of those expansions only correct up to and including all order Δ terms.

In summary, to a large extent, three of the approximations considered in this paper, namely *AS*, *epsilon* and Kessler, may be seen as representatives of an expansion of a specific order. For this reason we have not compared all combinations of type (*AS*, *epsilon*) and order (correct up to Δ^2 or correct up to Δ^3); specifically, we have not included in our comparisons the *AS* approximation correct up to and including all terms of order Δ^2 with the *epsilon* approximation correct up to and including all terms of order Δ^3, due to the redundancy indicated above.