COMMUNICATIONS IN STATISTICS—SIMULATION AND COMPUTATION®
2018, VOL. 00, NO. 0, 1–21
https://doi.org/10.1080/03610918.2018.1468456

Taylor & Francis
Taylor & Francis Group

Semi-parametric estimation of the autoregressive parameter in non-Gaussian Ornstein–Uhlenbeck processes

S. Rao Jammalamadaka[a] and Emanuele Taufer[b]

[a]Department of Statistics and Applied Probability, University of California, Santa Barbara, Santa Barbara, CA, USA;
[b]Department of Economics and Management, University of Trento, Trento, Italy

ABSTRACT
This paper considers the problem of estimating the autoregressive parameter in discretely observed Ornstein–Uhlenbeck processes. Two consistent estimators are proposed: one obtained by maximizing a kernel-based likelihood function, and another by minimizing a Kolmogorov-type distance from independence. After establishing the consistency of these estimators, their finite-sample performance and possible normality in large samples, is investigated by means of extensive simulations. An illustrative example to credit rating is discussed.

ARTICLE HISTORY
Received 18 December 2017
Accepted 6 April 2018

KEYWORDS
Adaptive estimation; Kernel density estimation; Lévy process; self-decomposable distribution; minimum squared distance to independence

1. Introduction

A continuous stationary process $\{X(t), t \geq 0\}$ is defined to be of the Ornstein–Uhlenbeck type (OU for short) if it is the solution of the stochastic differential equation

$$dX(t) = -\lambda X(t)dt + d\grave{Z}(t) \tag{1}$$

here $\lambda > 0$, and $\grave{Z}(t)$ is a homogeneous Lévy process, commonly referred to as the background driving Lévy process (BDLP), which satisfies the condition $\mathrm{E}[\log(1 + |\grave{Z}(1)|)] < \infty$ (see, e.g., Barndorff-Nielsen and Shephard 2001). Modeling via the use of general Lévy processes, other than Brownian motion, allows one to introduce specific non-Gaussian distributions for the marginal law of $X(t)$, and has received considerable attention in recent literature in an attempt to accommodate features such as jumps, semi-heavy tails and asymmetry, which are quite evident in real phenomena and are of practical interest in fields of application such as finance and econometrics.

Most notable examples include OU processes with marginal distributions such as the normal inverse Gaussian and the inverse Gaussian (Barndorff-Nielsen 1998), the variance gamma (Seneta 2004), the Meixner (Schoutens and Teugels 1998), the t-distribution (Heyde and Leonenko 2005), the normal, the stable and the gamma distributions. OU processes with positive jumps with marginal distributions such as the inverse Gaussian are often used as building blocks in stochastic volatility models (see, e.g., Barndorff-Nielsen and Shephard 2001).

A key concept related to these processes is that of self-decomposability. Recall that a random variable X with characteristic function $\psi(\zeta)$, is said to be self-decomposable if, for

CONTACT Emanuele Taufer emanuele.taufer@unitn.it Department of Economics and Management, University of Trento, Trento, Italy. AQ1
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

all $c \in (0,1)$, there exists a characteristic function $\psi_c(\zeta)$ such that $\psi(\zeta) = \psi(c\zeta)\psi_c(\zeta)$. Self-decomposability is closely related to stationary linear autoregressive time series of order 1, i.e. an AR(1) process: essentially the only possible AR(1) processes are those for which the one-dimensional marginal law is self-decomposable and similarly for the OU process, i.e. an "AR(1)" in continuous time. For further details on self-decomposable, infinitely divisible distributions and Lévy processes see Sato (1999).

This paper is concerned with estimation of the autoregressive parameter λ. Maximum likelihood estimation of λ is generally infeasible except for a few special cases and the large availability of marginal distributions for X calls for efficient estimation in a broad range of situations. To this end we propose two estimators: an estimator using a kernel estimate of the likelihood; another based on minimum distance from independence, which addresses some of the problems that the kernel-based estimator encounters in certain cases.

Suppose we observe the process Eq. (1) at equi-spaced time points $0 < t_1 < \cdots < \cdots t_n$ with $\Delta = t_j - t_{j-1}, j = 1, \ldots n$, $t_0 = 0$. In order to slightly simplify notation, denote the observation at time $t_j, X(t_j)$, by X_j. It follows from the discussion in Wolfe (1982) that, for self-decomposable distributions, a discrete AR(1) process can be embedded into a continuous OU process. In our case, this amounts to saying that the discretely observed OU process Eq. (1) can be written as AQ2

$$X_j = e^{-\lambda\Delta} X_{j-1} + \varepsilon_j, \quad j = 1, 2, \ldots, n \tag{2}$$

where the ε_j's are *i.i.d.* random variables. Note that in practical applications, determining the timing of observations is quite arbitrary, which amounts to saying that from a practical point of view one is not able to distinguish between Δ and λ. In this paper, contrary to other approaches where Δ is assumed to be known, we will actually consider estimation of, say, $\lambda' = \lambda\Delta$ so that, from now on it will be assumed without loss of generality that $\Delta = 1$. Denote $\theta = e^{-\lambda\Delta}$ and rewrite Eq. (2) as

$$X_j = \theta X_{j-1} + \varepsilon_j, \quad \theta \in \Theta, \quad \Theta = (0,1), \quad j = 1, 2, \ldots, n \tag{3}$$

With X_0 having distribution corresponding to the characteristic function $\psi(\zeta)$, model Eq. (3) is strictly stationary with marginal distribution having characteristic function $\psi(\zeta)$ and *i.i.d.* innovations with characteristic function $\psi_\theta(\zeta) = \psi(\zeta)/\psi(\theta\zeta)$.

Estimation of these models and in particular the estimation of the parameter θ (or λ) has attracted considerable interest in recent literature. When X is normal, the sample counterpart of the auto-correlation $Cor(X_1, X_2)$ provides, after transformation, the maximum likelihood estimator of λ. This turns out to be an estimator widely used in practice; Long (2009) has shown that the auto-correlation (AC) estimator is consistent for the model Eq. (3) with stable innovations with index of stability $1 < \alpha < 2$ with $\Delta = \Delta_n = 1/n$ when $n \to \infty$ and dispersion approaching 0. Zhang and Zhang (2013) show that the AC-based estimator of λ to be consistent for symmetric α-stable innovations for $0 < \alpha < 2$ either for fixed Δ and $\Delta \to 0$. Again, Hu and Long (2009) consider a least squares estimator for the case of α-stable innovations and show its consistency for $1 < \alpha < 2$ and $\Delta \to 0$. These approaches are equivalent when $\Delta \to 0$. Notwithstanding, the AC estimator turns out to be inefficient in many non-normal cases; to correct this situation Koul (1986) introduced a class of L_2-distance estimators of θ when the errors have an unknown symmetric distribution.

Jongbloed, Van der Meulen, and Van der Waart (2005) have proposed a highly efficient estimator of θ for the case where model $\grave{Z}$ is a subordinator, i.e., a process with positive increments. In this case, for the discretely observed model Eq. (3), $\hat{\theta} = \min_{1 \leq j \leq n} X_j / X_{j-1}$ which had also been discussed by Nielsen and Shephard (2003) in a model with exponential innovations. For other estimation problems for non-negative Lévy-driven OU processes, see Brockwell, Davis, and Yang (2007).

Restricting attention to non-negative Lévy-driven OU processes, however, excludes a whole range of possible marginal distributions for the model Eq. (1). A general parametric approach is considered by Taufer and Leonenko (2009a) which uses the characteristic function to estimate θ together with the parameters of the marginal distribution of X, while Andrews, Calder, and Davis (2009) discuss estimation of α-stable auto-regressive processes; see also Taufer, Leonenko, and Bee (2011) and Meintanis and Taufer (2012) for extensions to stochastic volatility models. Other papers of interest here are those of Diop and Yode (2010) who study a minimum distance estimator of θ when dispersion of the innovations approaches 0, and Ma (2010) who shows that the results of Long (2009) hold also under weaker conditions and Zhang, Lin, and Zhang (2015) which discuss LSE estimation for Lévy-driven moving averages.

The problem discussed here is closely connected to the works on adaptive estimation; in particular of direct relevance here are the papers of Kreiss (1987), Drost, Klaassen, and Werker (1997), Koul and Schick (1997) and Hallin et al. (2000) in time series contexts; Linton and Xiao (2007), Linton, Sperlich, and Van Keilegom (2008), and Yao and Zhao (2013) in semi-parametric and regression contexts; these approaches have in common the requirement that a preliminary consistent estimator of the parameter of interest is available while the approach proposed here has a one-step structure without using any preliminary estimator: only Eq. (3) is exploited and a simple maximization of a kernel density estimator is required. In this sense, the paper closest to our approach is Yuan and De Gooijer (2007) which considers a one-step adaptive procedure in the regression context. The problem discussed here may be seen as an extension to the dependent case of Yuan and De Gooijer (2007), although we adopt different techniques and require a minimal set of conditions, such as not requiring symmetry and placing very mild moment conditions in proving consistency of the estimators which generally hold in a large variety of self-decomposable distributions for OU processes.

AQ3

As for our second estimator based on the minimum distance to independence, previous related literature dates back to Manski (1983), Brown and Wegkamp (2002), and Linton, Sperlich, and Van Keilegom (2008); in these papers a minimum mean squared distance to independence is considered; this approach would require the existence of a finite mean, while here with the aim of requiring a minimal set of conditions, we use a Kolmogorov-type distance instead; the use of this distance has been discussed by Manski (1983) for the case where a parametric form of the distribution function is given while here a semi-parametric setting is discussed.

In the next section we will precisely define the estimators and present the main results. In Section 3 the small sample performance of the estimators will be analyzed my means of extensive simulations. An Appendix presents the proofs of the results.

In the paper we will use the notation $X_n = O_p(a_n)$ meaning that, for any $\varepsilon > 0$ there exists a finite M such that $P(|X_n/a_n| > M) < \varepsilon \ \forall n$ and $X_n = o_p(a_n)$ meaning that, for any $\varepsilon > 0$, $\lim_{n\to\infty} P(|X_n/a_n| > \varepsilon) = 0$. $X_n \to_D X$ is used to indicate convergence in distribution.

2. Semi-parametric estimators for θ and main results

If we denote by $\theta_0 \in \Theta$ the true parameter value, then the sequence of innovations $\varepsilon_j = X_j - \theta_0 X_{j-1}, j = 1, 2, \ldots, n$ is *i.i.d*. More generally, define the residuals $e_j = e_j^\theta = X_j - \theta X_{j-1}$, $j = 1, 2, \ldots, n$. Note that only the choice $\theta = \theta_0$ assures that X is strictly stationary with *i.i.d*. innovations ε; other choices of θ will lead to dependent innovations e. In fact, writing $e_j = (\theta_0 - \theta)X_{j-1} + \varepsilon_j, j = 1, 2, \ldots, n$ we note that the sequence of the e_j's is not independent due to the dependence of the X_j's.

2.1. A Kernel-based estimator

Let $f_\theta = f_\theta(e)$ denote the density of the residuals and, for $\theta = \theta_0, f_{\theta_0} = f_{\theta_0}(\varepsilon)$ denotes the density of the innovations. Also, let $f_\theta(x_0, x_1)$ be the bivariate density of X_0 and X_1.

Define the kernel estimator of f_θ, based on $e_j, j = 1, 2, \ldots, n$, as

$$\hat{f}_\theta(x) := \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - e_j^\theta}{h}\right) \tag{4}$$

where K is a scalar kernel and $h = h(n)$ is a bandwidth sequence. The following estimator of θ is proposed:

$$\hat{\theta}_1 = \arg\max_{\theta \in \Theta} \sum_{i \in S} \log \hat{f}_\theta(e_i) := \arg\max_{\theta \in \Theta} L_n(\theta) \tag{5}$$

where S is a subset of $\{1, 2, \ldots, n\}$ and it is introduced in case it is felt necessary to trim out some summands. Typically S will coincide with the full set $\{1, 2, \ldots, n\}$, i.e. all observations are used to estimate f_θ however, in some instances, one could get very small positive estimates of $\hat{f}_\theta$ which can cause numerical problems due to un-boundedness of the logarithmic function near the origin. Also, negative estimates of $\hat{f}_\theta$ could arise if higher order kernels are used.

To avoid these problems it is quite common in entropy estimation to assume that the support of f_θ is bounded, see, e.g. Hall (1986), van Es (1992), Hall and Morton (1993), and Yuan and De Gooijer (2007). This is not the approach followed here where OU processes require unbounded distributions.

From a purely practical point of view it might be a sensible precaution to exclude those e_i such that $\hat{f}_\theta(e_i) < b$ for some prescribed positive b or, alternatively, omitting those e_i such that $|e_i| > M$.

In our simulations (Section 3) in all but the stable cases with index of stability less than 1 the whole set of data was used without noticing any problem. When some trimming is necessary, this is usually quite evident as one gets unreasonable estimates of θ, i.e. 0 or 1 or a seemingly unbounded numerical likelihood. In the simulations we have set some common level of trimming for a given distribution. For an actual application, close inspection of data and estimates would suggest which data values should be excluded from the computations.

From a theoretical point of view, in order to ensure consistency one needs to allow arbitrary values of b or M. This point will be discussed more fully in the appendix. For showing the consistency of $\hat{\theta}_1$, we need the following standard conditions in kernel estimation:

A1 The sequence $\{X_i\}_{0 \le i \le n}$ follows model Eq. (3) and is strictly stationary with non-degenerate self-decomposable marginal distribution such that, for some $p > 0$ $\mathrm{E}(X_0^p) < \infty$.

A2 The density $f_\theta(x)$ is bounded away from 0 and Lipschitz continuous *wrt* θ on compact intervals of $x \in \mathbb{R}$ and $\sup_e f_\theta(e) < \infty$ for any θ.

A3 The joint density $f_\theta(x_0, x_1)$, is bounded away from 0 on compact sets of $x_0, x_1 \in \mathbb{R}$ and $\sup_{x_0,x_1} f_\theta(x_0, x_1) < \infty$ for any θ.

A4 $\int_{-\infty}^{\infty} |\log f_\theta(x)| f_\theta(x) dx < \infty$ for any θ.

A5 $|K(u)| < \infty$, $\int_{-\infty}^{\infty} |K(u)|du < \infty$, $\int_{-\infty}^{\infty} |uK(u)|du < \infty$.

A6 For some $M_1 < \infty$ and $M_2 < \infty$, either $K(u) = 0$ for $|u| > M_2$ and for all $u, u' \in \boldsymbol{R}$, $|K(u) - K(u')| \leq M_1|u - u'|$ or $K(u)$ is differentiable, $|(\partial/\partial u)K(u)| \leq M_1$, and for some $\nu > 1$, $|(\partial/\partial u)K(u)| \leq M_1|u|^{-\nu}$ for $|u| > M_2$.

A7 $h \to 0$, $nh \to \infty$ as $n \to \infty$.

Assumption A1 specializes the situation to the context of OU processes and has some relevant consequences for our results. First of all we note that any non-degenerate self-decomposable distribution is absolutely continuous (Sato 1999, Thm. 27.13). This, in turn, together with the postulated conditions on boundedness of the density and its derivatives (A2 and A10 below) implies that the density $f = f_\theta$ belongs to the class of densities that satisfies

$$\sup_x f(x) + \sup_{x,x'} \frac{|f(x) - f(x')|}{|x - x'|} \leq M, \qquad 0 < M < \infty \tag{6}$$

Second, from Masuda (2004, Theorem 4.3) it follows that $\{X_i\}_{0 \leq i \leq n}$ is ergodic and β-mixing with coefficients, for some $a > 0$, $\beta_X(t) = O(e^{-at})$. Recall that if X is a strictly stationary Markov process with initial distribution π and t^{th} step transition probability $P^t(x, .)$, then the β-mixing coefficients are defined as

$$\beta_X(t) = \int ||P^t(x, .) - \pi(.)||\pi(dx)$$

where $||\mu||$ denotes the total variation norm of a signed measure μ. The fact that $\{X_i\}_{0 \leq i \leq n}$ is α-mixing follows from the inequality $2\alpha(t) \leq \beta(t)$. Conditions A2 and A3 require that all densities involved are bounded and A4 introduces a very mild tail restriction. Conditions A5 and A7 are quite standard in kernel density estimation while A6, introduced in Hansen (2008), is satisfied by most kernels including the normal one.

Our proof of consistency has a very mild restriction on existence of moments (A1 and A4) and uses boundedness and continuity (but not differentiability) conditions on the densities involved (A2, A3). On the other hand, it will require that the density estimates be restricted on a compact interval $\{x : |x| \leq c_n\}$ with $c_n \to \infty$ as $n \to \infty$ so that, ultimately, consistency will hold on a set of probability 1. The truncating device is defined in the Appendix.

Theorem 1. *Assume conditions A1–A7; then* $|\hat{\theta}_1 - \theta_0| = o_p(1)$.

Asymptotic normality of $\hat{\theta}_1$ appears to need additional regularity assumptions, as well as existence of third order moments of X. This issue is investigate further in the simulations section.

2.2. A minimum distance to independence estimator

As we will see, a kernel based estimator suffers some problems when distributions with very heavy tails are involved. In such cases it may be sensible to resort to an alternative; here, in order to provide an estimator which could be used under a minimal set of conditions and which could be a computationally attractive competitor, we introduce an estimator based on a minimum distance from independence. Define, with $\boldsymbol{I}_A$ being the indicator function of A,

$$\hat{F}_\theta(t) = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{I}_{(e_j^\theta \le t)} \tag{7}$$

and

$$\hat{F}_\theta(t_1, t_2) = \frac{1}{n(n-1)}\sum_{i \neq j}^{n} \boldsymbol{I}_{(e_j^\theta \le t_1)}\boldsymbol{I}_{(e_i^\theta \le t_2)} \tag{8}$$

An estimator of θ can be obtained as

$$\hat{\theta}_2 = \arg\min_{\theta \in \Theta} \sup_{t_1, t_2 \in R} \left|\hat{F}_\theta(t_1, t_2) - \hat{F}_\theta(t_1)\hat{F}_\theta(t_2)\right| \tag{9}$$

The use of the *sup* norm rather than other measures of distance is dictated by the desire to construct an estimator based on a minimal set of conditions on F. We then have (see Appendix for the proof).

Theorem 2. *Assume A1 then* $|\hat{\theta}_2 - \theta_0| = o_p(1)$.

In terms of computing $\hat{\theta}_2$, one may note that

$$\begin{aligned}\hat{F}_\theta(t_1, t_2) - \hat{F}_\theta(t_1)\hat{F}_\theta(t_2) &= \frac{1}{n(n-1)}\sum_{i \neq j}^{n} \boldsymbol{I}_{(e_j \le t_1)}\boldsymbol{I}_{(e_i \le t_2)} - \frac{1}{n^2}\sum_{i,j=1}^{n} \boldsymbol{I}_{(e_j \le t_1)}\boldsymbol{I}_{(e_i \le t_2)} \\ &= \frac{1}{n^2(n-1)}\sum_{i \neq j}^{n} \boldsymbol{I}_{(e_j \le t_1)}\boldsymbol{I}_{(e_i \le t_2)} - \frac{1}{n^2}\sum_{i=1}^{n} \boldsymbol{I}_{(e_i \le t_1)}\boldsymbol{I}_{(e_i \le t_2)}\end{aligned}$$

The actual computation of the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ can be done by a simple grid search.

3. Performance in finite samples

In this section the finite-sample performance of the proposed estimators is analyzed by simulations. The base-line to which we will compare the performance of our estimators will be the AC based estimator which is equivalent to several approaches proposed in the literature (see the introductory section for discussion about this) and, for the case of processes with positive increments, with the highly efficient estimator $\hat{\theta} = \min_{1 \le j \le n} X_j / X_{j-1}$ proposed by Jongbloed, Van der Meulen, and Van der Waart (2005); it is expected that $\hat{\theta}_1$ and $\hat{\theta}_2$ will not perform better than $\hat{\theta}$ however it is of interest here to give an overall evaluation of their performance.

Distributions over the real line such as the normal, the normal inverse Gaussian, the t-Student and the stable are considered; inverse Gaussian and stable OU processes with positive increments will also be used. The notation used will be a standard one, i.e., a normal distribution with mean μ and variance σ^2 will be denoted as $N(\mu, \sigma^2)$; the normal inverse

Gaussian distributions is indicated with $NIG(\alpha, \beta, \mu, \sigma)$ where α, β, μ and σ are related, respectively, to the tail, asymmetry, location and scale, $0 \le \beta \le \alpha$, $\mu \in \boldsymbol{R}$, $\sigma > 0$; t_ν stands for a t-Student distribution with ν degrees of freedom while $S(\alpha, \beta, \mu, \sigma)$ denotes a stable distribution with index of stability α, and where β, μ and σ indicate, respectively, asymmetry, location and scale; here we have $0 < \alpha \le 2$, $0 \le \beta \le 1$, $\mu \in \boldsymbol{R}$, $\sigma > 0$; the inverse Gaussian distribution with mean μ and shape σ will be indicated by $IG(\mu, \sigma)$.

As to the choice of kernel, we will compare two possibilities: a normal kernel, which is a standard choice in many computer packages, as well as a heavy tail kernel which should work better for heavy tailed distributions, namely

$$K(u) = \frac{1}{2} e^{-|u|} \tag{10}$$

The choice of the smoothing bandwidth h exhibits a strong influence on the resulting estimate and it may not be optimal to consider automatic choices in running extensive simulations. Several alternatives have been compared: Silverman's rule of thumb, least squares cross validation, Sheather-Jones, over-smooth rule, standard deviation; we found that the choice of simply using the standard deviation as bandwidth works generally quite well for our problem and here we report estimation results based on that choice without any changes on single cases; this will allow a fair comparison on the estimators.

We found that the kernel-based estimators suffer some problems when facing distributions with heavy tails, where it is clear that in some cases the estimation procedure is failing completely, e.g. illogical results or improper kernel estimates. Hence implementation of formula Eq. (5) was carried out by eliminating those data for which $e > M$ for a given M. In the tables, simulated results with trimming and without trimming are reported; the value of M is indicated in the tables by writing $e \le M$, i.e. all values $e > M$ have been eliminated. The choice of M is the result of a trial and error procedure by which the problems noted above are eliminated. For non-stable distributions no trimming was used. In the simulations, to prevent any bias in the comparison of the estimators we have chosen a general rule for trimming outliers and report the results as they are; we suspect that considering data-driven techniques would improve substantially the performance of the kernel-based method in the case of heavy-tailed distributions.

The OU processes with given marginal distribution have been generated according to the technique suggested in Taufer and Leonenko (2009b). All simulations have been run using the Mathematica ® 8 software and the commands there automatically defined for kernel density estimation ("Smooth Kernel Distribution" with the "Standard deviation" bandwidth selection method) as well as for random number generation. The grid search for the value of θ maximizing the estimated likelihood or minimizing the Kolmogorov distance from independence has been set from 0.01 to 0.99 with 0.01 increments.

Tables 1–7 respectively report the estimation results for OU processes with marginal distributions: 1) $N(0, 3)$; 2) $NIG(2, 1.7, -1, 1)$; 3) t_4; 4) $S(1.5, -0.8, 0, 1)$; 5) $S(0.5, 0.5, 0, 1)$; 6) $S(0.8, 1, 0, 1)$; 7) $IG(2, 2)$, the last two cases being positive distributions. For all cases but the IG one, where $\lambda = 0.5$, λ has been set to one. The examples proposed cover a variety of cases with symmetric and asymmetric, heavy and semi-heavy tailed marginal distributions. The Monte Carlo estimates of the mean and mean squared error ($\widehat{MSE}$) of the estimators of $\theta = e^{-\lambda}$ are based on 1000 simulations of samples with sizes $n = 50, 100, 200, 300$, where,

Table 1. Monte Carlo simulation results: N(0,3); $\theta = 0.3679$ ($\lambda = 1$). Mean, MSE and Relative efficiency (RE) of the estimators with respect to AC. Estimates based on 1000 replications.

		n = 50	*n* = 100	*n* = 200	*n* = 300
AC	*Mean*	0.3236	0.3445	0.3560	0.3591
	MSE	0.0191	0.0953	0.0047	0.0027
SE	*Mean*	0.4476	0.3969	0.3784	0.3732
	RE	0.3086	0.6677	0.5550	0.5746
NO	*Mean*	0.3312	0.3476	0.3577	0.3603
	RE	1.0240	0.9777	0.9857	0.9958
HT	*Mean*	0.3297	0.3473	0.3576	0.3602
	RE	0.9822	0.9295	0.9487	0.9790

Table 2. Monte Carlo simulation results: NIG(2,1.7,−1,1); $\theta = 0.3679$ ($\lambda = 1$). Mean, MSE and Relative efficiency (RE) of the estimators with respect to AC. Estimates based on 1000 replications.

		n = 50	*n* = 100	*n* = 200	*n* = 300
AC	*Mean*	0.3188	0.3403	0.3544	0.3588
	MSE	0.0170	0.0087	0.0041	0.0028
SE	*Mean*	0.4120	0.3759	0.3721	0.3694
	RE	0.5532	1.0148	1.1043	1.2807
NO	*Mean*	0.3319	0.3511	0.3619	0.3642
	RE	1.1018	1.3368	1.9082	2.0078
HT	*Mean*	0.3194	0.3590	0.3638	0.3632
	RE	1.1059	2.6348	2.7717	2.5150

Table 3. Monte Carlo simulation results: t_4; $\theta = 0.3679$ ($\lambda = 1$). Mean, MSE and Relative efficiency (RE) of the estimators with respect to AC. Estimates based on 1000 replications.

		n = 50	*n* = 100	*n* = 200	*n* = 300
AC	*Mean*	0.3171	0.3375	0.3531	0.3590
	MSE	0.0180	0.0097	0.0045	0.0031
SE	*Mean*	0.4235	0.3832	0.3717	0.3670
	RE	0.3757	0.5184	0.6946	0.7699
NO	*Mean*	0.3302	0.3451	0.3568	0.3611
	RE	1.2514	1.2296	1.2833	1.2597
HT	*Mean*	0.3301	0.3462	0.3576	0.3617
	RE	1.2134	1.2472	1.3492	1.2905

for an estimator $\hat{\theta}$:

$$\widehat{MSE}(\hat{\theta}) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\theta}_i - \theta_0)^2 \tag{11}$$

with $\hat{\theta}_i$ the estimator obtained at the i-th Monte Carlo replicate, $i = 1, 2, \ldots, M$.

Each table reports: mean and $\widehat{MSE}$ for the auto-correlation estimator (AC); mean and relative efficiency (RE) with respect to the AC estimator for:

1. the minimum distance to independence estimator $\hat{\theta}_2$, indicated with SE;
2. the normal kernel-based $\hat{\theta}_1$ estimator, indicated with NO;
3. the Eq. (10) kernel-based $\hat{\theta}_1$ estimator, indicated with HT;

Table 4. Monte Carlo simulation results: Stable(1.5,–0.8,0,1); $\theta = 0.3679$ ($\lambda = 1$). Mean, MSE and Relative efficiency (RE) of the estimators with respect to AC. Estimates based on 1000 replications.

		n = 50	*n = 100*	*n = 200*	*n = 300*
AC	*Mean*	0.3156	0.3427	0.3565	0.3594
	MSE	0.0152	0.0071	0.0032	0.0023
SE	*Mean*	0.4341	0.3941	0.3783	0.3735
	RE	0.3864	0.5180	0.6030	0.7908
NO	*Mean*	0.3398	0.3561	0.3646	0.4271
	RE	1.5493	1.7451	1.3533	0.0660
HT	*Mean*	0.3436	0.3592	0.3660	0.4279
	RE	1.6244	1.8969	1.4965	0.0670
HTC	*Mean*	0.3435	0.3592	0.3642	0.3670
$e \leq 50$	*RE*	1.6141	1.8171	1.8012	1.4878

Table 5. Monte Carlo simulation results: Stable(0.5, 0.5, 0,1); $\theta = 0.3679$ ($\lambda = 1$). Mean, MSE and Relative efficiency (RE) of the estimator with respect to AC. Estimates based on 1000 replications.

		n = 50	*n = 100*	*n = 200*	*n = 300*
AC	*Mean*	0.3276	0.3462	0.3569	0.3622
	MSE	0.0085	0.0042	0.0024	0.0004
SE	*Mean*	0.3920	0.3763	0.3720	0.3706
	RE	1.4754	4.6537	14.960	16.7523
NO	*Mean*	0.3568	0.3598	0.4931	–
	RE	2.9792	2.9395	0.0383	–
HT	*Mean*	0.3704	0.3698	0.4977	–
	RE	4.9750	3.2251	0.0384	–
HTC	*Mean*	0.3788	0.3701	0.3652	0.3687
$e \leq 500$	*RE*	0.6538	0.7316	0.7937	0.5607

Table 6. Monte Carlo simulation results: S(0.8,1,0,1); $\theta = 0.3679$ ($\lambda = 1$). Mean, MSE and Relative efficiency (RE) of the estimators with respect to AC. Estimates based on 1000 replications.

		n = 50	*n = 100*	*n = 200*	*n = 300*
AC	*Mean*	0.3238	0.3476	0.3582	0.3611
	MSE	0.0096	0.0038	0.0017	0.0009
SE	*Mean*	0.3855	0.3759	0.3725	0.3704
	RE	1.5854	3.3078	4.6422	4.7578
NO	*Mean*	0.3545	0.3625	0.4432	–
	RE	2.8035	3.6082	0.0480	–
HT	*Mean*	0.3683	0.3707	0.4459	–
	RE	6.3683	15.2199	0.0485	–
HTC	*Mean*	0.3654	0.3636	0.3640	0.3684
$e \leq 100$	*RE*	3.7733	2.8597	2.0772	2.1075
RA	*Mean*	0.4025	0.3851	0.3761	0.3728
	RE	4.0470	5.9336	10.3867	15.7843

4. the ratio-estimator of Jongbloed, Van der Meulen, and Van der Waart (2005) is indicated with *RA*.

In the case of stable distributions for which, as mentioned, automatic simulations with standard settings suffered some problems, results for the kernel-based estimator *HT* computed

Table 7. Monte Carlo simulation results: IG(2,2); $\theta = 0.6065$ ($\lambda = 0.5$). Mean, MSE and relative efficiency (RE) of the estimators with respect to AC. Estimates based on 1000 replications.

		n = 50	*n* = 100	*n* = 200	*n* = 300
AC	*Mean*	0.5491	0.5763	0.5927	0.5953
	MSE	0.0129	0.0058	0.0028	0.0020
SE	*Mean*	0.6493	0.6298	0.6151	0.6077
	RE	0.6597	0.6625	0.9352	0.9704
NO	*Mean*	0.5759	0.5904	0.6001	0.6008
	RE	1.9786	1.9568	1.8757	1.8876
HT	*Mean*	0.5881	0.5968	0.6034	0.6036
	RE	3.5109	3.8107	3.8725	4.0709
RA	*Mean*	0.6327	0.6282	0.6238	0.6222
	RE	15.9005	10.9213	8.6006	7.3762

with extremes outliers censored out are reported; this is indicated as *HTC* and the level M above which residuals have been eliminated is indicated as $e \leq M$.

The choice of reporting the RE with respect to the AC estimator is in order to emphasize the comparisons with respect to a cornerstone for all estimators. If $\hat{\theta}_{AC}$ denotes the AC estimator and $\hat{\theta}_O$ denotes any other estimator used in the simulations, then

$$RE(\hat{\theta}_O) = \frac{\widehat{MSE}(\hat{\theta}_{AC})}{\widehat{MSE}(\hat{\theta}_O)} \tag{12}$$

An RE higher than one results in a better performance of the estimator under analysis with respect to the AC estimator.

In terms of investigating whether these estimators are asymptotically normal, Figures 1–4 show the distribution of the estimators for some of the cases discussed in the tables, namely we consider the OU processes with $N(0, 3)$ and $IG(2, 2)$ marginal distribution either where λ is estimated using the normal kernel or the heavy kernel Eq. (10). In each figure the histogram of the standardized data is super-imposed with the standard normal density and PP and QQ plots for normality are reported. As we note from the figures, a normal approximation works quite well in all cases for sample sizes of around $n = 100$.

To summarize, the results in the tables and the figures are quite clear and indicate that generally the *NO* and *HT* estimators perform better with respect to the *AC* estimator having some problems only in the case of extremely heavy tails where in this case the *SE* estimator performs very well. Specifically we can summarize the results as follows:

a) in the Normal case the relative efficiency of *NO* and *HT* is always quite close to unity essentially indicating (as suggested by the theoretical results) no loss in efficiency with respect to the maximum likelihood estimator AC, even for small sample sizes.

b) In all other cases the performance of *NO* and *HT* is generally better with respect to *AC* and relative efficiency can be quite high. In the stable case, some distinction needs to be made: it appears that, as sample size increases, the large number of extreme observations has a serious effect on the efficiency of the estimators, trimming can improve the situation. The tables, reporting the results of standardized simulations, may not show the effective performance of *NO* and *HT* in these cases.

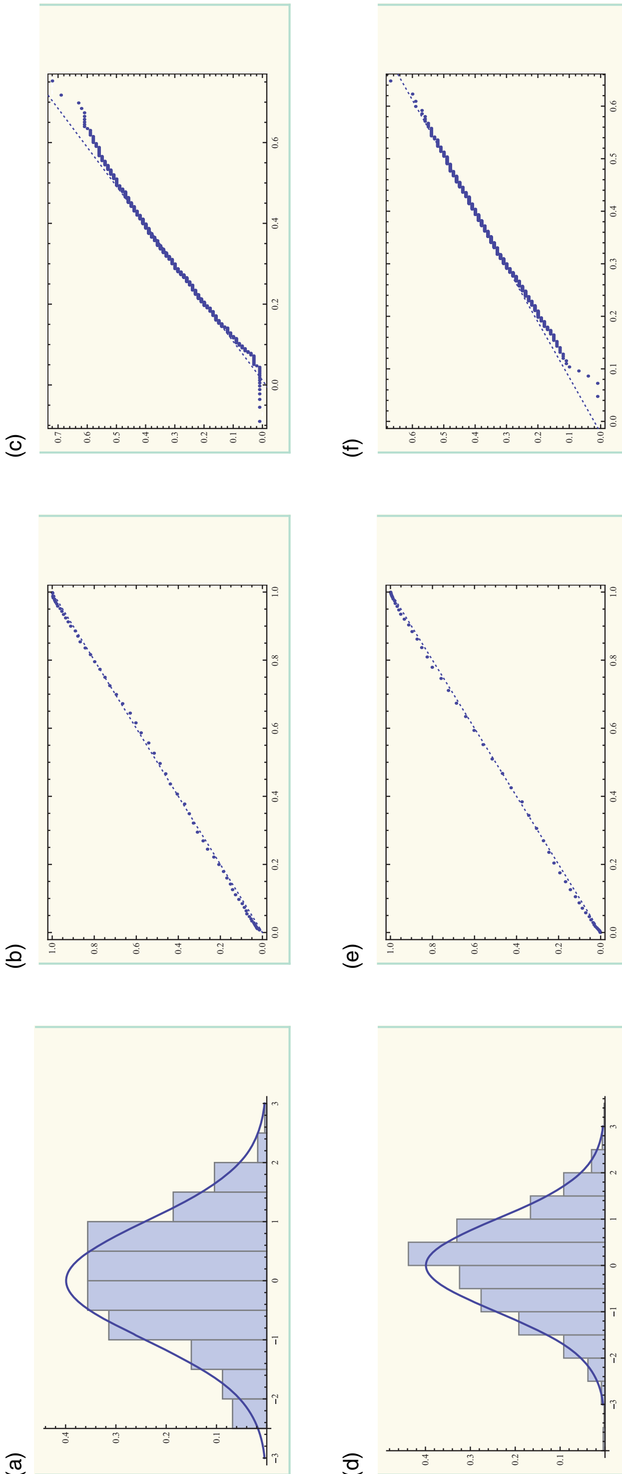

Figure 1. OU process with $N(0, 3)$ marginal distribution. Empirical summaries of 1000 estimates of $\hat{\theta}_1$ based on the normal kernel; $n = 50$ (top) and $n = 100$ (bottom) sample size.

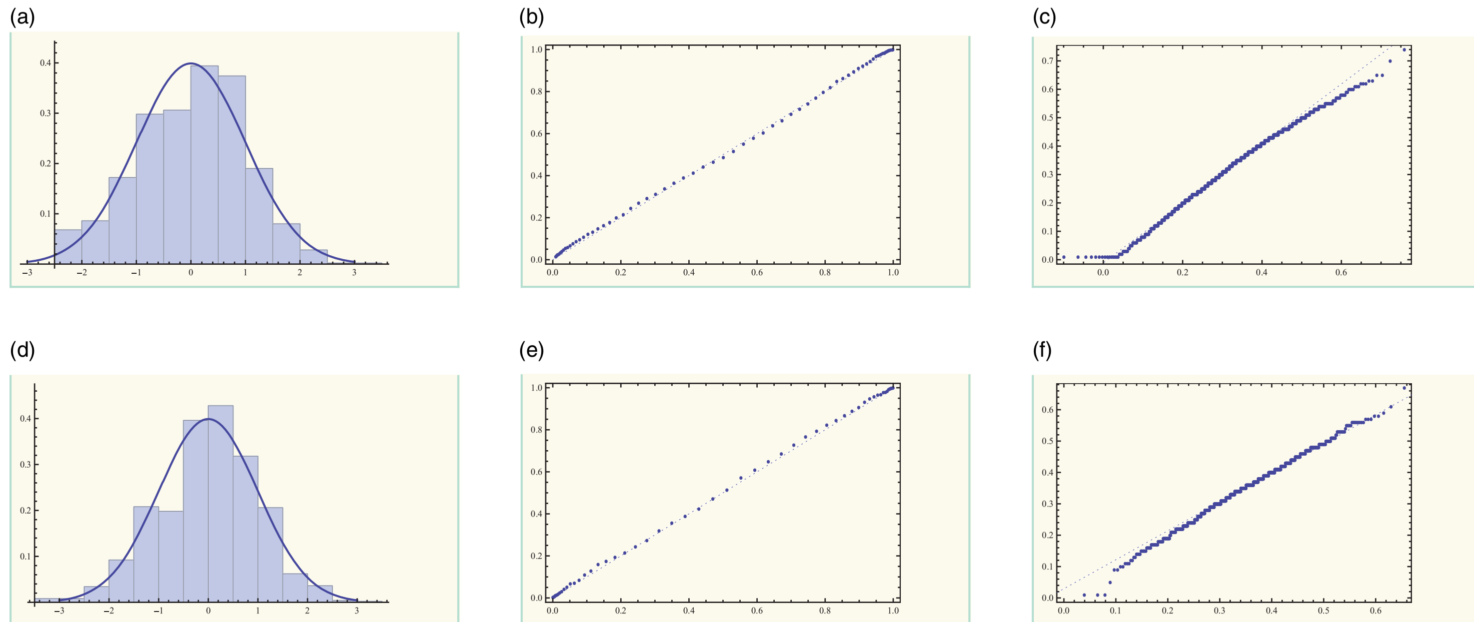

Figure 2. OU process with $N(0,3)$ marginal distribution. Empirical summaries of 1000 estimates of $\hat{\theta}_1$ based on the heavy kernel Eq. (10); $n = 50$ (top) and $n = 100$ (bottom) sample size.

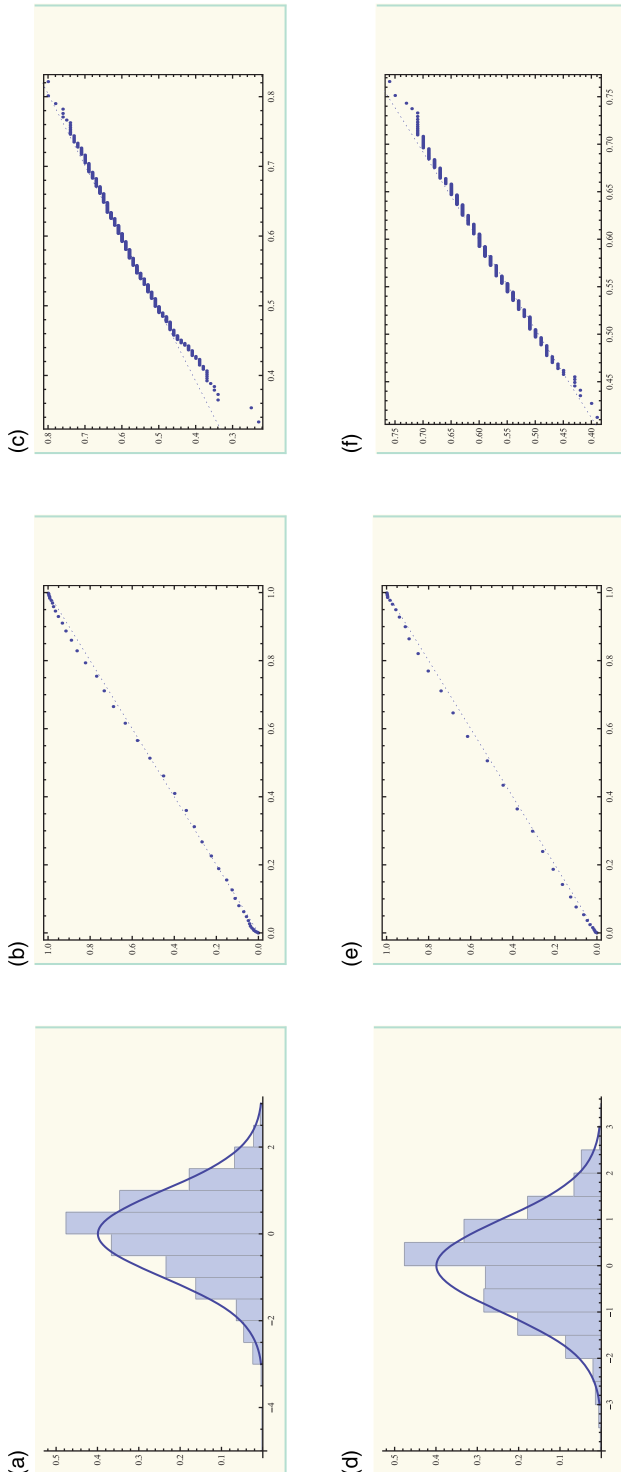

Figure 3. OU process with $IG(2, 2)$ marginal distribution. Empirical summaries of 1000 estimates of $\hat{\theta}_1$ based on the normal kernel; $n = 50$ (top) and $n = 100$ (bottom) sample size.

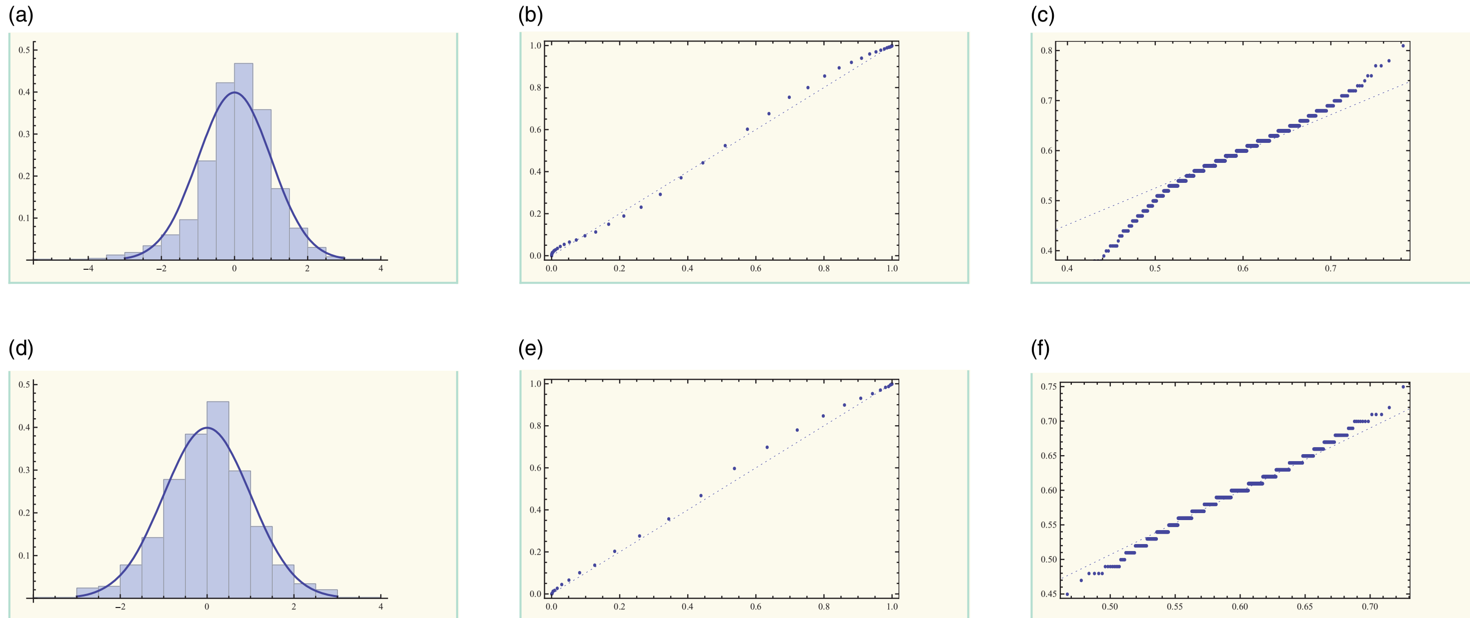

Figure 4. OU process with $IG(2, 2)$ marginal distribution. Empirical summaries of 1000 estimates of $\hat{\theta}_1$ based on the heavy kernel Eq. (10); $n = 50$ (top) and $n = 100$ (bottom) sample size.

c) The performance of *HT* is generally better than *NO* and its relative efficiency can be much higher in semi-heavy or heavy tail cases.
d) In the case of OU processes with positive increments, the performance of *RA* is generally better than all the other estimators and its efficiency can be substantially larger than one. Note however that the *HT* estimator can perform extremely well for small sample sizes in the stable case and overcome the performance of *RA*.
e) The performance of *SE* is generally poorer with respect to the other estimators but in the case of distributions with very heavy tails, e.g. stable with $\alpha < 1$, for which *SE* does not suffer from the presence of extremely large observations.
f) A general rule, which seems to be efficient in a large variety of cases is the following: use *SE* if very heavy tails are present otherwise use *HT*. The *RA* estimator should be used for OU processes with positive increments.
g) Asymptotic normality seems to hold very well in all cases discussed.

4. An example

Moody's trailing 12-month default rates are widely monitored indicators of corporate credit quality and are a good source either for theoretical and empirical studies. For example, Amerio, Muliere, and Secchi (2004) have studied the historical distributions of one-year default rates for Ba-rated, B-rated and Caa-rated defaulters during the period 1970–1999; Keenan, Sobehart, and Hamilton (1999) and Taufer (2007) have used either the entire Moody's rated universe (all-corporate, AC) and a sub-grouping, i.e., the speculative-grade (SG) monthly data respectively from 1970 to 1999 and from 1920 to 2004 in order to provide forecasting models.

In this example we are going to consider the SG yearly data for the period 1920–2011 for a total of 92 observations ranging from a minimum value of 0 to a maximum of 15.641. The data are taken from Moody's website and are freely available.

To begin with, we have a look at the linear plots of the series in Figure 5(a). The path does not appear to be non-stationary, however the high spikes suggests non normality of the data, which is confirmed by analytical tests and normality plot (not shown here). The auto-correlation and partial auto-correlation function in Figures 1(b) and 1(c) suggests that a (discretely observed) OU model could be appropriate for this data.

If normality is excluded, using the AC estimator maybe inappropriate; instead, one could consider some alternative approaches. Following the results of the simulations, for positive distributions, the highly efficient ratio estimator (RA) of Jongbloed, Van der Meulen, and Van der Waart (2005) should be used. Note however that the presence of several null values in the data makes it impossible its calculation. Also, the minimum distance estimator (SE) seems inappropriate here as there is no evidence of heavy tails and it is generally less efficient with respect to the kernel ones. Then, following the recommendations given in Section 3 we proceed to compute the HT estimator with no trimming. In this case the AC and the HT estimator are in good agreement, giving an estimate of 0.64 and 0.63 respectively. A further estimation on subsets of the data with 12 series of length 80 give values of the AC and HT estimators within a range of 0.63 and 0.71. Even though in this example the two estimators are very close, using alternative approaches is important in order to substantiate our empirical analysis.

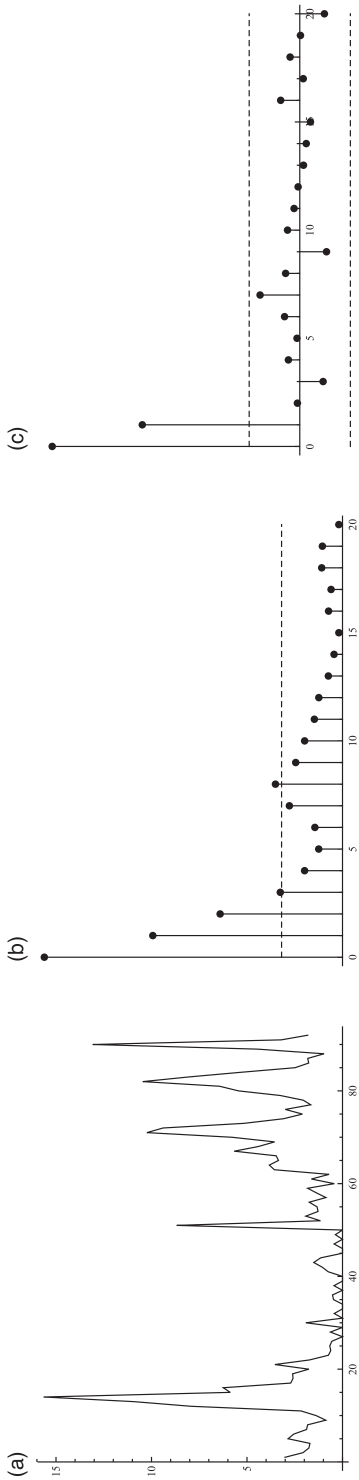

Figure 5. Autocorrelation and partial auto-correlation functions of the data.

References

Amerio, E., P. Muliere, and P. Secchi. 2004. Reinforced urn processes for modeling credit default distributions. *International Journal of Theoretical and Applied Finance* 7 (4):407–23.

Andrews, B., M. Calder, and R. A. Davis. 2009. Maximum likelihood estimation for α-stable autoregressive processes. *The Annals of Statistics* 37:1946–82.

Barndorff-Nielsen, O. E. 1998. Processes of normal inverse Gaussian type. *Finance and Stochastics* 2:41–68.

Barndorff-Nielsen, O. E., and N. Shephard. 2001. Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society, Series B. Statistical Methodology* 63:167–241.

Brockwell, P. J., R. A. Davis, and Y. Yang. 2007. Estimation for nonnegative Lévy-driven Ornstein-Uhlenbeck processes. *Journal of Applied Probability* 44:977–89.

Brown, D. J., and M. H. Wegkamp. 2002. Weighted minimum mean Â–square distance from independence estimation.*Econometrica* 70:2035–51.

Chen, X., O. Linton, and I. Van Keilegom. 2003. Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* 71:1591–1608.

Diop, A., and A. F. Yode. 2010. Minimum distance parameter estimation for Ornstein–Uhlenbeck processes driven by Lévy processes. *Statistics & Probability Letters* 80:122–7.

Drost, F. C., C. A. J. Klaassen, and B. J. M. Werker. 1997. Adaptive estimation in time-series models. *Annals of Statistics* 25:786–817.

Durrett, R. A. 1991. *Probability: theory and examples*. Belmont, CA: Wadsworth. AQ4

Hall, P. 1986. On powerful distributional tests based on sample spacings. *Journal of Multivariate Analysis* 19:201–24.

Hall, P., and S. C. Morton. 1993. On the estimation of entropy. *Annals of the Institute of Statistical Mathematics* 45:69–88.

Hansen, B. E. 2008. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory* 24:726–48.

Heyde, C. C. and N. N. Leonenko. 2005. Student processes. *Advances in Applied Probability* 37 (2):342–65.

Hu, Y., and H. Long. 2009. Least squares estimator for Ornstein-Uhlenbeck processes driven by α-stable motions. *Stochastic Processes and their Applications* 119:2465–480.

Jongbloed, G., F. H., Van der Meulen, and A. Van der Waart. 2005. Nonparametric inference for Lévy-driven Ornstein Uhlenbeck processes. *Bernoulli* 11:759–91.

Kapur, J. N., and H. K. Kesavan. 1992. *Entropy optimization principles with applications*. Boston, MA: Academic Press.

Keenan, S. C., J. Sobehart, and D. T. Hamilton. 1999. Predicting default rates: a forecasting model for Moody's issuer-based default rates. Moody's Investor Service: Special Comment.

Koul, H. L. 1986. Minimum distance estimation and goodness-of-fit tests in first-order auto-regression. *Annals of Statistics* 14:1194–213.

Koul, H. L., and A. Schick. 1997. Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli* 3:247–77.

Kreiss, J. P. 1987. On adaptive estimation in stationary ARMA processes. *Annals of Statistics* 15 (1):112–33.

Linton, O., and Z. Xiao. 2007. A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory* 23 (3):371–413.

Linton, O., S. Sperlich, and I. Van Keilegom. 2008. Estimation of a semiparametric transformation model. *Annals of Statistics* 36 (2):686–718.

Long, H. 2009. Least squares estimator for discretely observed Ornstein–Uhlenbeck processes with small Lévy noises. *Statistics & Probability Letters* 79:2076–85.

Ma, C. 2010. A note on "Least squares estimator for discretely observed Ornstein-Uhlenbeck processes with small Lévy noises". *Statistics & Probability Letters* 80:1528–31.

Manski, C. F. 1983. Closest empirical distribution estimation. *Econometrica* 51:305–19.

Masuda, H. 2004. On multidimensional Ornstein–Uhlenbeck processes driven by a general Lévy process. *Bernoulli* 10 (1):97–120.
Meintanis, S. G., and E. Taufer. 2012. Inference procedures for stable-Paretian stochastic volatility models. *Mathematical and Computer Modelling* 55 (3–4):1199–212.
Newey, W. K. 1991. Uniform convergence in probability and stochastic equicontinuity. *Econometrica* 59:1161–67.
Nielsen, B., and N. Shephard. 2003. Likelihood analysis of a first-order autoregressive model with exponential innovations. *Journal of Time Series Analysis* 24 (3):337–44. AQ5
Sato, K. 1999. *Lévy processes and infinitely divisible distributions*. Cambridge: Cambridge University Press.
Schoutens, W., and J. L. Teugels. 1998. Lévy processes, polynomials and martingales. Special issue in honor of Marcel F. Neuts. *Communications in Statistics: Stochastic Models* 14:335–49.
Seneta, E. 2004. Fitting of variance-Gamma model to financial data. *Journal of Applied Probability* 41:177–87.
Taufer, E. 2007. Modeling stylized features in default rates. *Applied Stochastic Models in Business and Industry* 23:73–82.
Taufer, E., and N. Leonenko. 2009a. Characteristic function estimation of non-Gaussian Ornstein-Uhlenbeck processes. *Journal of Statistical Planning and Inference* 139:3050–63.
Taufer, E., and N. Leonenko. 2009b. Simulation of Lévy-driven Ornstein-Uhlenbeck processes with given marginal distribution. *Computational Statistics & Data Analysis* 53:2427–37.
Taufer, E., N. Leonenko, and M. Bee. 2011. Characteristic function estimation of Ornstein-Uhlenbeck-based stochastic volatility models. *Computational Statistics & Data Analysis* 55:2525–39.
Van der Vaart, A. W. 1998. *Asymptotic statistics*. New York: Cambridge University Press.
van Es, B. 1992. Estimating functionals related to a density by a class of statistics based on spacings. *Scandinavian Journal of Statistics* 19:61–72.
Yao, W., and Z. Zhao. 2013. Kernel density based linear regression estimate. *Communications in Statistics—Theory and Methods* 42:4499–512.
Yuan, A., and J. De Gooijer. 2007. Semiparametric regression with kernel error model. *Scandinavian Journal of Statistics* 34:841–69.
Zhang, S. and X. Zhang. 2013. A least squares estimator for discretely observed Ornstein–Uhlenbeck processes driven by symmetric α-stable motions. *Annals of the Institute of Statistical Mathematics* 65:89–103.
Zhang, S., Z. Lin, and X. Zhang. 2015. A least squares estimator for Lévy-driven moving averages based on discrete time observations. *Communications in Statistics—Theory and Methods* 44:1111–29.
Zhao, Z., and W. B. Wu. 2009. Nonparametric inference of discretely sampled stable Lévy processes. *Journal of Econometrics* 153:83–92. AQ6

Appendix

For the proof of the results, a preliminary lemma due to Hansen (2008), specialized to our set-up, is needed.

Lemma A1. *Assume conditions A1–A7; then, for $c_n = O((\ln n)n^{1/2})$,*

$$\sup_{|x|\le c_n} \left|\hat{f}_\theta(x) - f_\theta(x)\right| = O_p\left(\left(\frac{\ln n}{nh}\right)^{1/2} + h^q\right), \quad \forall\theta \in \Theta \tag{A1}$$

where q denotes the order of the kernel.

Lemma A1 follows directly from Theorem 2 and Theorem 6 in Hansen (2008) by noting that, from Masuda (2004), the sequence $X_0, \dots X_n$ following model Eq. (2) is β-mixing with geometric rate; we therefore can take in the theorems of Hansen (2008), $\theta = 1$. The result of Lemma 1 can be strengthened to almost sure convergence and convergence over the whole real line by strengthening the assumptions; but the present version will suffice for our purposes.

In the proof of the results, a smooth trimming function G_b will be used, where

$$G_b(x) = \begin{cases} 0, & x < b \\ \int_b^x g_b(z)dz, & b \le x \le 2b \\ 1, & x > 2b \end{cases} \tag{A2}$$

Here $g_b(x) = \frac{1}{b}g(x/b-1)$ with $b > 0$ a trimming parameter and g any density function with support in $[0,1]$, $g(0) = g(1) = 0$. This approach has been followed, for example, by Linton and Xiao (2007) and Yao and Zhao (2013) and a proper choice of g allows to use standard Taylor series arguments; for example, if $g(z) = cz^{\alpha}(1-z)^{\alpha}$, $z \in [0,1]$ $\alpha > 0$ and c an appropriate normalizing constant, then G_b is $\alpha+1$ times continuously differentiable on $[0,1]$. Note also that $\sup_x G_b(x)/x^k \le 1/b^k$.

Lemma A2. *Assume conditions A1–A7, then,*
a)

$$\max_{1\le i\le n}\left|\frac{\hat{f}_\theta(e_i) - f_\theta(e_i)}{f_\theta(e_i)}\right| = o_p(1), \quad \forall\theta\in\Theta \tag{A3}$$

b)

$$\sup_{|\theta_1-\theta_2|\le\varepsilon}\ \max_{1\le i\le n}\left|\frac{\hat{f}_{\theta_1}(e_i) - \hat{f}_{\theta_2}(e_i)}{\hat{f}_{\theta_2}(e_i)}\right| = o_p(1) + O(\varepsilon) \tag{A4}$$

Proof of Lemma A2. For the proof of part *a*), consider first a trimmed version

$$\max_{1\le i\le n}\left|\frac{\hat{f}_\theta(e_i) - f_\theta(e_i)}{f_\theta(e_i)}\right| G_b(f_\theta(e_i)) \le \max_{1\le i\le n}\frac{|\hat{f}_\theta(e_i) - f_\theta(e_i)|}{b} \tag{A5}$$

using the fact that $\sup_x G_b(x)/x \le 1/b$. Next, for $\boldsymbol{I}_{(x)}$ the indicator function, note that

$$\max_{1\le i\le n}\left|\hat{f}_\theta(e_i) - f_\theta(e_i)\right| \boldsymbol{I}_{(|e_i|\le c_n)} \le \sup_{|x|\le c_n}\left|\hat{f}_\theta(x) - f_\theta(x)\right| \tag{A6}$$

hence Lemma A1, as $n\to\infty$, implies that Eq. (A5) is $o_p(1)O(1/b)$ and the result follows as the choice of b is arbitrary.

As far as part *b*) is concerned, using part *a*) we have,

$$\max_{1\le i\le n}\left|\frac{\hat{f}_{\theta_2}(e_i)}{f_{\theta_2}(e_i)} - 1\right| = o_p(1) \quad \text{and} \quad \max_{1\le i\le n}\left|\frac{\hat{f}_{\theta_1}(e_i) - f_{\theta_1}(e_i)}{f_{\theta_2}(e_i)}\right| = o_p(1) \tag{A7}$$

Next note that

$$\frac{\hat{f}_{\theta_1}(e)}{\hat{f}_{\theta_2}(e)} = \frac{\hat{f}_{\theta_1}(e)/f_{\theta_2}(e)}{\hat{f}_{\theta_2}(e)/f_{\theta_2}(e)} = \frac{\frac{f_{\theta_1}(e)}{f_{\theta_2}(e)} + \frac{\hat{f}_{\theta_1}(e)-f_{\theta_1}(e)}{f_{\theta_2}(e)}}{1 + \frac{\hat{f}_{\theta_2}(e)-f_{\theta_2}(e)}{f_{\theta_2}(e)}} \tag{A8}$$

results in Eq. (A7) imply that

$$\max_{1\le i\le n}\left|\frac{\hat{f}_{\theta_1}(e)}{\hat{f}_{\theta_2}(e)} - \frac{f_{\theta_1}(e)}{f_{\theta_2}(e)}\right| = o_p(1) \tag{A9}$$

Based on the above results we obtain

$$\begin{aligned} &\sup_{|\theta_1-\theta_2|\le\varepsilon}\ \max_{1\le i\le n}\left|\frac{\hat{f}_{\theta_1}(e_i) - \hat{f}_{\theta_2}(e_i)}{\hat{f}_{\theta_2}(e_i)}\right| \\ &\le \sup_{|\theta_1-\theta_2|\le\varepsilon}\ \max_{1\le i\le n}\left|\frac{\hat{f}_{\theta_1}(e_i)}{\hat{f}_{\theta_2}(e_i)} - \frac{f_{\theta_1}(e_i)}{f_{\theta_2}(e_i)}\right| + \sup_{|\theta_1-\theta_2|\le\varepsilon}\ \max_{1\le i\le n}\left|\frac{f_{\theta_1}(e_i)}{f_{\theta_2}(e_i)} - 1\right| \\ &\le o_p(1) + \frac{C\varepsilon}{b}; \quad \forall b \end{aligned} \tag{A10}$$

where the first term on the *r.h.s.* of the above expression is from Eq. (A8) whereas the second is again obtained by truncation and from condition A2. Again we can make the above term as small as desired as the choice on b and ε are arbitrary. □

Proof of Theorem 1. In order to prove consistency of $\hat{\theta}_1$ we need to show that:
a) there is a function, say $L(\theta)$, such that $\sup_{\theta\in\Theta}|L_n(\theta)-L(\theta)| = o_p(1)$;
b) $L(\theta)$ is uniquely maximized by θ_0.

In order to prove part *a*) we need to verify that: (i) the parameter space is compact; (ii) $L_n(\theta) \to_P L(\theta)$ point wise; (iii) equicontinuity in probability, i.e. there exists $\delta > 0$ such that $\sup_{|\theta_1-\theta_2|\le\delta}|L_n(\theta_1)-L_n(\theta_2)| = o_p(1)$.

As far as point (*i*) is concerned note that although $\Theta = (0,1)$ is not compact we can consider a compact set K such that $\theta_0 \in K \subset (0,1)$. In order to verify (*ii*), define $M_n = M_n(\theta) = \frac{1}{n}\sum_{j=1}^n \log f_\theta(e_j)$ and $L(\theta) = \mathrm{E}(\ln f_\theta(e))$. Then, since $X_1,\dots,X_n$ is ergodic, under Assumption A4 it follows that $M_n \to_P L(\theta)$, $\forall\theta\in\Theta$. Since $|\ln(1+x)| \le 2|x|$ in an neighborhood of $x=0$, a sufficient condition for $L_n - M_n \to_P 0$, is

$$\max_{1\le i\le n}\left|\frac{\hat{f}_\theta(e_i)}{f_\theta(e_i)} - 1\right| = o_p(1) \quad \forall\theta\in\Theta \tag{A11}$$

which follows from Lemma A2a. It follows that $L_n(\theta) \to_P L(\theta)$ point-wise. Similarly, to show *iii*) note that,

$$\begin{aligned}\sup_{|\theta_1-\theta_2|\le\varepsilon_n}\max_{1\le i\le n}|L_n(\theta_1)-L_n(\theta_2)| &= \sup_{|\theta_1-\theta_2|\le\varepsilon_n}\max_{1\le i\le n}\frac{1}{n}\left|\sum_{i=1}^n \log\left(1+\frac{\hat{f}_{\theta_1}(e_i^{\theta_1})-\hat{f}_{\theta_2}(e_i^{\theta_2})}{\hat{f}_{\theta_2}(e_i^{\theta_2})}\right)\right| \\ &\le 2\sup_{|\theta_1-\theta_2|\le\varepsilon_n}\max_{1\le i\le n}\left|\frac{\hat{f}_{\theta_1}(e_i^{\theta_1})-\hat{f}_{\theta_2}(e_i^{\theta_2})}{\hat{f}_{\theta_2}(e_i^{\theta_2})}\right|\end{aligned}$$

which is $o_p(1)$ by Lemma A2b for suitably chosen ε_n. In order to prove part *b*), define $L(\theta) = -H(\varepsilon^\theta)$ where H is the Shannon's entropy (see, e.g., Kapur and Kesavan 1992). Then,

$$\begin{aligned}H(\varepsilon^\theta) &= H(\varepsilon^{\theta_0} + (\theta_0-\theta)X_0) \\ &\ge H(\varepsilon^{\theta_0} + (\theta_0-\theta)X_0|X_0) \\ &= H(\varepsilon^{\theta_0}|X_0) \\ &= H(\varepsilon^{\theta_0})\end{aligned} \tag{A12}$$

where we have used, in order, the facts that; conditioning reduces entropy; a constant does not change entropy; ε^{θ_0} and X_0 are independent. It follows that $L(\theta)$ is uniquely maximized by $L(\theta_0)$. □

Proof of Theorem 2. Denote for simplicity $\sup_{t_1,t_2\in R}|\hat{F}_\theta(t_1,t_2)-\hat{F}_\theta(t_1)\hat{F}_\theta(t_2)| = \rho(\hat{F},\theta)$. The proof of the theorem follows from Theorem 2 in Manski (1983) if we verify the following conditions:
B1 The parameter space Θ is compact.
B2 $\rho(F,\theta) = 0$ if and only if $\theta = \theta_0$.
B3 (Assumption 4 in Manski (1983) - continuity and uniform convergence). $\rho(F,\theta)$ is continuous as a function on Θ. Also, $\rho(\hat{F},\theta)$ converges in probability to $\rho(F,\theta)$ uniformly over Θ.

As far as B1 is concerned, as already discussed, one can consider a compact set K such that $\theta_0 \in K \subset (0,1)$. B2 follows form the discussion in Section 2, as the sequence $\{e_i^\theta\}_{1\le j\le n}$ is i.i.d only if $\theta = \theta_0$.

The firs part of B3 can be verified by first noting that F, being self-decomposable, is absolutely continuous (Sato 1999, Them 27.13) and exploiting the first part of the corollary to Theorem 2 in Manski (1983) by noting that $g(X_1,X_0,\theta) = X_1 - \theta X_0$ is continuous on $S\times\Theta$ where $S \in \boldsymbol{R}^2$ is some compact and convex set.

The second part follows if we prove that

$$\sup_{\theta\in\Theta}\sup_{t_1,t_2\in R}\left|\hat{F}_\theta(t_1,t_2)-\hat{F}_\theta(t_1)\hat{F}_\theta(t_2)-F_\theta(t_1,t_2)+F_\theta(t_1)F_\theta(t_2)\right| = o_p(1) \tag{A13}$$

In order to do this, note that

$$\begin{aligned}&\left|\hat{F}_\theta(t_1,t_2)-\hat{F}_\theta(t_1)\hat{F}_\theta(t_2)-F_\theta(t_1,t_2)+F_\theta(t_1)F_\theta(t_2)\right|\\&\quad\le\left|\hat{F}_\theta(t_1,t_2)-F_\theta(t_1,t_2)\right|+\hat{F}_\theta(t_1)\left|\hat{F}_\theta(t_2)-F_\theta(t_2)\right|+F_\theta(t_2)\left|\hat{F}_\theta(t_1)-F_\theta(t_1)\right|\end{aligned}\tag{A14}$$

Note that since the sequence $\{X_i\}_{0\le i\le n}$ is ergodic and the class of functions $\mathcal{F}=\{f_t=\boldsymbol{I}_{(-\infty,t]},t\in\boldsymbol{R}^2\}$ are Glivenko–Cantelli (see, e.g. Van der Vaart 1998, p. 270), we have that

$$\sup_{t_1,t_2\in R^2}\left|\hat{F}_\theta(t_1,t_2)-F_\theta(t_1,t_2)\right|=o_p(1),\quad\text{and}\quad\sup_{t\in\boldsymbol{R}}\left|\hat{F}_\theta(t)-F_\theta(t)\right|=o_p(1)\quad\forall\theta\in\Theta\tag{A15}$$

We claim that

$$\sup_{\theta\in\Theta}\sup_{t\in\boldsymbol{R}}\left|\hat{F}_\theta(t)-F_\theta(t)\right|=o_p(1)\tag{A16}$$

$$\sup_{\theta\in\Theta}\sup_{t_1,t_2\in\boldsymbol{R}}\left|\hat{F}_\theta(t_1,t_2)-F_\theta(t_1,t_2)\right|=o_p(1)\tag{A17}$$

The proof of Eqs. (A16) and (A17) together with compactness of Θ and Eq. (A15) will prove Eq. (A13).

In order to prove Eqs. (A16) and (A17) we'll exploit Theorem 3 in Chen, Linton, and Van Keilegom (2003) which provides primitive conditions for equicontinuity: we'll have to show that their condition (3.2) is satisfied, which require in our case to show that

$$\left[\mathrm{E}\left(\sup_{|\theta_1-\theta_2|\le\delta}\left|\boldsymbol{I}_{\{X_1-\theta_1X_0\le t\}}-\boldsymbol{I}_{\{X_1-\theta_2X_0\le t\}}\right|^r\right)\right]^{1/r}\le K\delta^s\tag{A18}$$

$$\left[\mathrm{E}\left(\sup_{|\theta_1-\theta_2|\le\delta}\left|\boldsymbol{I}_{\{X_1-\theta_1X_0\le t\}}\boldsymbol{I}_{\{X_2-\theta_1X_1\le t\}}-\boldsymbol{I}_{\{X_1-\theta_2X_0\le t\}}\boldsymbol{I}_{\{X_2-\theta_2X_2\le t\}}\right|^r\right)\right]^{1/r}\le K\delta^s\tag{A19}$$

for all $\theta\in\Theta$, all small positive values $\delta=o(1)$, $r\ge 2$ and $s\in(0,1]$ and that the bounds hold for μ-almost all (t_1,t_2). Consider Eq. (A18) and note that the expectation of the absolute value in the expression is the probability of the union of the events $\{t+\theta_1X_0<X_1<t+\theta_2X_0\}$ and $\{t+\theta_2X_0<X_1<t+\theta_1X_0\}$ which consider all possibilities arising from the cases $\theta_1\le\theta_2$ or $\theta_1>\theta_2$, $X_0\le 0$ or $X_0>0$. Since X_0 is bounded in probability there is a compact set with probability greater that $1-\varepsilon$, $\varepsilon>0$, for which there is some upper bound c such that $\sup_{|\theta_1-\theta_2|\le\delta}|\theta_1X_0-\theta_2X_0|\le\delta c$. For some $\delta>0$ we have then

$$\begin{aligned}\mathrm{E}\left(\sup_{|\theta_1-\theta_2|\le\delta}\left|\boldsymbol{I}_{\{X_1-\theta_1X_0\le t\}}-\boldsymbol{I}_{\{X_1-\theta_2X_0\le t\}}\right|\right)&\le 2P\left(t-\delta c+\theta X_0<X_1<t+\delta c+\theta X_0\right)\\&=2F_\theta(t-\delta c)-F_\theta(t+\delta c)\\&\le K\delta\end{aligned}\tag{A20}$$

for some constant $K<\infty$, from continuity of F. Therefore condition (3.2) of Theorem 3 in Chen, Linton, and Van Keilegom (2003) is satisfied with $r=2$ and $s=1/2$. The proof of Eq. (A19) resorts to an analogous device. □