Stat Methods Appl (2012) 21:411–436
DOI 10.1007/s10260-012-0195-2

Robust parameter estimation for the Ornstein–Uhlenbeck process

Sonja Rieder

Accepted: 8 May 2012 / Published online: 26 May 2012

Abstract In this paper, we derive elementary M- and optimally robust asymptotic linear (AL)-estimates for the parameters of an Ornstein–Uhlenbeck process. Simulation and estimation of the process are already well-studied, see Iacus (Simulation and inference for stochastic differential equations. Springer, New York, 2008). However, in order to protect against outliers and deviations from the ideal law the formulation of suitable neighborhood models and a corresponding robustification of the estimators are necessary. As a measure of robustness, we consider the maximum asymptotic mean square error (maxasyMSE), which is determined by the influence curve (IC) of AL estimates. The IC represents the standardized influence of an individual observation on the estimator given the past. In a first step, we extend the method of M-estimation from Huber (Robust statistics. Wiley, New York, 1981). In a second step, we apply the general theory based on local asymptotic normality, AL estimates, and shrinking neighborhoods due to Kohl et al. (Stat Methods Appl 19:333–354, 2010), Rieder (Robust asymptotic statistics. Springer, New York, 1994), Rieder (2003), and Staab (1984). This leads to optimally robust ICs whose graph exhibits surprising behavior. In the end, we discuss the estimator construction, i.e. the problem of constructing an estimator from the family of optimal ICs. Therefore we carry out in our context the One-Step construction dating back to LeCam (Asymptotic methods in statistical decision theory. Springer, New York, 1986) and compare it by means of simulations with MLE and M-estimator.

Keywords Ornstein–Uhlenbeck process · Influence curves · M-estimators · Asymptotically linear estimators

S. Rieder (✉)
Department of Mathematics, École Polytechnique Fédérale de Lausanne, Station 8,
1015 Lausanne, Switzerland
e-mail: sonja.rieder@ep.ch

1 Introduction

The Ornstein–Uhlenbeck process was introduced in Ornstein and Uhlenbeck (1930) as a model for the velocity process of Brownian particles (i.e. free particles moving in gas) being affected by a friction force. The work of Ornstein and Uhlenbeck therefore continued the work of Einstein on Brownian motion, which studied the friction-less movement. The process has nowadays many applications; e.g. in finance it is used as a model for interest rates, see Vasicek (1977). Simulation and estimation are already well-studied, see Iacus (2008). However, in order to protect against outliers and deviations from the ideal law, we formulate neighborhoods about transition distributions and robustify the parameter estimators with respect to these neighborhoods. As a measure of robustness, we consider the maximum asymptotic mean square error (maxasyMSE) determined by the influence curve (IC) of the AL estimates.

Influence curves of correlation measures have been recently studied in the literature, see Croux and Dehon (2010).

But contrary to the original definition of the IC as Gâteaux derivative of statistical functionals by Huber (1981), we define the IC via an estimator expansion. This is motivated by the observation that most proofs of asymptotic normality amount to an estimator expansion with the IC as summand. By this definition, we achieve the original interpretation of the IC (as standardized, asymptotic influence of an individual observation on the estimator given the past) and avoid suitable regularity assumptions.

The paper is organized as follows: We first summarize some basic facts of the process. Secondly, we extend the method of M-estimation from Huber (1981), leading to M-equations for the parameters of the Ornstein–Uhlenbeck process. However, a more general robustness concept seems more appropriate and necessary. Therefore, thirdly, we apply the infinitesimal approach of Kohl et al. (2010), Rieder (1994, 2003), and Staab (1984), which aims to minimize the maxasyMSE on shrinking neighborhoods about the ideal transition distributions. Afterwards we compare graphically the ICs of the two approaches.

In the last section, we discuss the estimator construction, that is, the problem of constructing an estimator from the family of optimal ICs. Therefore we carry out in our context the One-Step construction dating back to LeCam (1986). In a simulation study we compare then this One-Step estimator with Method of Moment (MoM), MLE and M-estimator for different contamination models. However, our approach is theoretical incomplete; we finalize by summarizing the remaining mathematical problems for future research.

2 Preliminaries on the Ornstein–Uhlenbeck process

In this section, we summarize some facts of the process, both in continuous and discrete time.

2.1 Continuous time

A stochastic process $X = (X_t)_{t\geq 0}$ is said to be an Ornstein–Uhlenbeck process if it satisfies the linear stochastic differential equation (SDE)

$$dX_t = \lambda(\mu - X_t)dt + \sigma dB_t \tag{1}$$

with parameters $\lambda, \sigma \in \mathbb{R}^+$, $\mu \in \mathbb{R}$ and $(B_t)_{t\geq 0}$ the standard Brownian motion. This process is well-studied in the literature, see Øksendal (2002), and the solution of (1) is given by

$$X_t = e^{-\lambda t}X_0 + \mu(1 - e^{-\lambda t}) + \sigma \int_0^t e^{\lambda(s-t)}dB_s, \tag{2}$$

where $X_0 := X(0)$ and $\int_0^t e^{\lambda(s-t)}dB_s \sim \mathcal{N}\left(0, \frac{1-e^{-2\lambda t}}{2\lambda}\right)$.
Choosing the limiting distribution for $t \to \infty$ in (2) as initial distribution, i.e. $X_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\lambda}\right)$, the process will become stationary (in the strong sense).
Furthermore, assuming X_0 independent of $\mathcal{F}(V(t), t \geq 0)$, where $V(t) := \int_0^t e^{\lambda(s-t)}dB_s$, the process is Gaussian with expectation $\mathrm{E}(X_t) = \mu$ and covariance

$$\mathrm{Cov}(X_t, X_u) = \frac{\sigma^2}{2\lambda}e^{-\lambda|t-u|}. \tag{3}$$

From Gaussianity and weak stationarity we may deduce that the Ornstein–Uhlenbeck process is strongly stationary. In addition, having independent increments the process is Markov. Furthermore, the form of the covariance function in (3) implies the ergodicity of the Ornstein–Uhlenbeck process, see Breiman (1992, p. 120).

2.2 Discrete time

The process observed in discrete time is more relevant in statistics. Therefore we denote by $x_{\leq n} := (x_n, x_{n-1}, \ldots, x_0)$ discrete-time observations on a regular grid. Then, by (2), the Ornstein–Uhlenbeck time series is for $i = 1, \ldots, n$ given by

$$x_i = e^{-\lambda d}x_{i-1} + \mu(1 - e^{-\lambda d}) + \sigma\sqrt{\frac{(1 - e^{-2\lambda d})}{2\lambda}} \cdot z_i, \tag{4}$$

where $Z_i \sim \mathcal{N}(0, 1)$ i.i.d and with equidistant time lag d, fixed in advance.
For $\mu = 0$, and $\sigma = 1$ we obtain as a special case from (4) the Standard Ornstein–Uhlenbeck which is $AR(1)$

$$x_i = \theta x_{i-1} + v_i, \tag{5}$$

where $V_i \sim \mathcal{N}(0, (1 - e^{-2\lambda d})/(2\lambda))$ i.i.d, and $0 < \theta = e^{-\lambda d} < 1$ is in the so-called region of stationarity of an $AR(1)$ time series.

3 M-estimates

In this section, we extend the idea of M-estimates for location and regression to the Ornstein–Uhlenbeck time series. We understand M-estimation according to Huber (1981) in an elementary way: We replace the quadratic function in the likelihood by Huber's ρ-function, and as far as scale is concerned we introduce Huber's Proposal 2. This leads to our M-estimates for $\theta = (\mu, \sigma, \lambda)$.

We should be aware, however, that this extension is rather intuitive: Huber (1981) has only treated location, the theory remains incomplete for simultaneous location and scale, and does not cover more general models or the case of dependent observations.

In literature several other authors have extended M-estimation to dynamic location-scale models: Muler and Yohai (2002, 2008), consider instead of Huber's Proposal 2 the so-called τ-scale and modify the likelihood for ARCH and GARCH models in this way.

There exist also several other approaches for simultaneous location and scale problems: Besides Kohl et al. (2010), Rieder (1994), and Staab (1984), Mancini et al. (2005) propose robust estimators for conditional location and scale parameters for a strictly stationary time series model, the Ornstein–Uhlenbeck time series being a special case. However, contrary to Huber's approach where his explicit ψ-function is used, the ψ-function in these approaches is derived by an optimization criterion.

We will discuss a more general robust approach in the next section, in this section the aim is to apply the most simple and intuitive approach.

Initial density and conditional densities of the Ornstein–Uhlenbeck process, for $i = 1, \ldots, n$, are given by

$$f_\theta(x_0) = \frac{\sqrt{\lambda}}{\sqrt{\pi}\sigma} \exp\left(-\lambda \frac{(x_0 - \mu)^2}{\sigma^2}\right),$$

$$f_\theta(x_i|x_{i-1}) = \frac{\sqrt{\lambda}}{\sqrt{\pi}\sigma\sqrt{1 - e^{-2\lambda d}}} \exp\left(-\frac{\lambda\left(x_i - \mu - e^{-\lambda d}(x_{i-1} - \mu)\right)^2}{\sigma^2(1 - e^{-2\lambda d})}\right).$$

Applying the Markovian and Gaussian properties of the process, we easily obtain the following likelihood for the sample $x_0, x_1, \ldots, x_n$

$$\mathcal{L}(\theta) = -\frac{n+1}{2}\log\left(\frac{\pi\sigma^2}{\lambda}\right) - \frac{1}{2}\left(\frac{\sqrt{2\lambda}(x_0 - \mu)}{\sigma}\right)^2 - \frac{n}{2}\log(1 - e^{-2\lambda d})$$
$$-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{\sqrt{2\lambda}\left((x_i - \mu) - e^{-\lambda d}(x_{i-1} - \mu)\right)}{\sigma\sqrt{1 - e^{-2\lambda d}}}\right)^2. \tag{6}$$

We now replace the quadratic function $\frac{1}{2}x^2$ in (6) by the Huber function

$$\rho(x) = \rho_k(x) = \begin{cases} \frac{1}{2}x^2 & |x| \le k \\ \text{sign}(x)k(x-k) + \frac{1}{2}k^2 & |x| > k \end{cases}, \tag{7}$$

for some $k;\ 1 \le k \le 2$, see Huber (1981).
This gives

$$\mathcal{L}_\rho(\theta) = -\frac{n+1}{2}\log\left(\frac{\pi\sigma^2}{\lambda}\right) - \rho\left(\frac{\sqrt{2\lambda}(x_0-\mu)}{\sigma}\right) - \frac{n}{2}\log(1-e^{-2\lambda d}) - \sum_{i=1}^{n}\rho\left(\frac{\sqrt{2\lambda}\left((x_i-\mu)-e^{-\lambda d}(x_{i-1}-\mu)\right)}{\sigma\sqrt{1-e^{-2\lambda d}}}\right).$$

Denote

$$\psi(x) = \psi_k(x) = \rho_k'(x) = \begin{cases} x & |x| \le k \\ \text{sign}(x)k & |x| > k \end{cases}, \tag{8}$$

and

$$r_i = \frac{\sqrt{2\lambda}\left((x_i-\mu)-e^{-\lambda d}(x_{i-1}-\mu)\right)}{\sigma\sqrt{1-e^{-2\lambda d}}} \qquad i = 1,\ldots,n. \tag{9}$$

Then the first derivatives of $\mathcal{L}_\rho(\theta)$ may be written as

$$\frac{\partial\mathcal{L}_\rho}{\partial\mu} = \frac{\sqrt{2\lambda}}{\sigma}\psi\left(\frac{\sqrt{2\lambda}(x_0-\mu)}{\sigma}\right) - \frac{\sqrt{2\lambda}(e^{-\lambda d}-1)}{\sigma\sqrt{1-e^{-2\lambda d}}}\sum_{i=1}^{n}\psi(r_i),$$

$$\frac{\partial\mathcal{L}_\rho}{\partial\sigma} = -\frac{n+1}{\sigma} + \frac{\sqrt{2\lambda}(x_0-\mu)}{\sigma^2}\psi\left(\frac{\sqrt{2\lambda}(x_0-\mu)}{\sigma}\right) + \frac{1}{\sigma}\sum_{i=1}^{n}\psi(r_i)\,r_i,$$

$$\frac{\partial\mathcal{L}_\rho}{\partial\lambda} = \frac{n+1}{2\lambda} - \frac{x_0-\mu}{\sqrt{2\lambda}\sigma}\psi\left(\frac{\sqrt{2\lambda}(x_0-\mu)}{\sigma}\right) - \frac{nde^{-2\lambda d}}{1-e^{-2\lambda d}} - \frac{1}{2\lambda}\sum_{i=1}^{n}\psi(r_i)\,r_i - \sum_{i=1}^{n}\psi(r_i)\frac{\sqrt{2\lambda}e^{-\lambda d}d(x_{i-1}-\mu)}{\sigma\sqrt{1-e^{-2\lambda d}}} + \frac{e^{-2\lambda d}d}{1-e^{-2\lambda d}}\sum_{i=1}^{n}\psi(r_i)\,r_i.$$

Hence, the M-estimator $\hat{\mu}$ is given as the solution of

$$\psi\left(\frac{\sqrt{2\hat{\lambda}}(x_0-\hat{\mu})}{\hat{\sigma}}\right) - \frac{e^{-\hat{\lambda}d}-1}{\sqrt{1-e^{-2\hat{\lambda}d}}}\sum_{i=1}^{n}\psi\left(\hat{r}_i\right) = 0, \tag{10}$$

where the arguments of ψ are

$$\hat{r}_i = \frac{\sqrt{2\hat{\lambda}}\left((x_i - \hat{\mu}) - e^{-\hat{\lambda}d}(x_{i-1} - \hat{\mu})\right)}{\hat{\sigma}\sqrt{1 - e^{-2\hat{\lambda}d}}} \quad i = 1, \ldots, n. \tag{11}$$

And the M-equations for σ and λ are

$$\chi\left(\frac{\sqrt{2\hat{\lambda}}(x_0 - \mu)}{\hat{\sigma}}\right) + \sum_{i=1}^{n} \chi\left(\hat{r}_i\right) = 0, \tag{12}$$

$$-\frac{\sqrt{2\hat{\lambda}}}{\hat{\sigma}} \sum_{i=1}^{n} \psi\left(\hat{r}_i\right)(x_{i-1} - \hat{\mu}) + \frac{e^{-\hat{\lambda}d}}{\sqrt{1 - e^{-2\hat{\lambda}d}}} \sum_{i=1}^{n} \chi\left(\hat{r}_i\right) = 0, \tag{13}$$

where

$$\chi(x) = \psi(x)^2 - a(k), \tag{14}$$

and $a(k) = 2\Phi(k) - 1 - 2k\varphi(k) + 2k^2(1 - \Phi(k))$, with $\varphi(k)$ and $\Phi(k)$ denoting the standard Gaussian probability density function and the standard Gaussian distribution function, respectively, see Huber (1981).

Although we adopt Huber's choice of the functions ψ and χ in (8) and (14), we should be aware however, that the Ornstein–Uhlenbeck process is quite different from a location model with i.i.d errors.

The M-equations (10), (12), and (13) can be solved numerically, e.g. by applying the Newton method with the robustified MoM estimates as initial values: $\mu_0 = \text{med}(x)$, $\sigma_0 = \sqrt{2\lambda_0}\,\text{mad}(x)$, and $\lambda_0 = -\frac{\log(\widehat{\text{ACF}}(d))}{d}$, $\widehat{\text{ACF}}(d)$ denoting the empirical autocorrelation function with time lag d, med the median, and mad the median absolute deviation.

4 Approach based on LAN and shrinking neighborhoods

However, since the theory in Huber (1981) does not cover the dependent case, the M-approach from the previous section may be reasonable practical but is not optimal in general. Therefore we apply a more general concept to the Ornstein–Uhlenbeck time series, the approach based on local asymptotic normality (LAN) and shrinking neighborhoods due to Kohl et al. (2010), Rieder (1994, 2003), and Staab (1984). It will lead in fact to IC and estimates of a different kind than those based on the M-approach. In the simulation study in Sect. 6 we then compare the estimators of the two approaches, and use also the M-estimator as initial estimator for the One-Step construction of the more general approach.

4.1 LAN condition

For this approach the LAN-condition from Hájek (1972) and LeCam (1986) is required. We give this condition in terms of the following general notation.

Let for $n \geq 1$ $(X_1, \ldots, X_n)$ be a sequence of random vectors defined on a probability space $(\Omega, \mathcal{F}, P_\theta)$, with k-dimensional parameter $\theta \in \Theta$, and $\Theta \subset \mathbb{R}^k$ an open subset.

Further we denote by $\mathcal{F}_n := \sigma(X_1, \ldots, X_n)$ the σ-field induced by the random vector $(X_1, \ldots, X_n)$ and by $P_\theta^{(n)}$ the restriction of P_θ to $\mathcal{F}_n$.

Then the family of probability measures $\mathcal{P} := \left\{P_\theta^{(n)}; \theta \in \Theta\right\}$ on $(\Omega_n, \mathcal{F}_n)$ is local asymptotic normal (LAN) at θ if there exists a sequence of random variables $Z_n : (\Omega_n, \mathcal{A}_n) \to (\mathbb{R}^k, \mathcal{B}^k)$ that are asymptotically normal, i.e. $Z_n \to \mathcal{N}(0, C)$ in law under $P_\theta^{(n)}$, such that for each $h \in \mathbb{R}^k$

$$\log\left(dP^{(n)}_{\theta+\frac{h}{\sqrt{n}}} \Big/ dP^{(n)}_\theta\right) - \left(h' Z_n - \frac{1}{2} h' C h\right) \longrightarrow 0 \tag{15}$$

stochastically in $P_\theta^{(n)}$-probability.

So far in literature the LAN property has been shown for general ergodic diffusion processes, containing the Ornstein–Uhlenbeck process, in Gobet (2002). In case of the Standard Ornstein–Uhlenbeck process the property has been also verified by Hallin et al. (2000). Since the LAN property remains valid under location-scale transformations, Hallin et al. (2000) contains also the general Ornstein–Uhlenbeck process. In addition, Swensen (1985) showed the LAN property for general AR(p) time series, which includes AR(1), thus also the Ornstein–Uhlenbeck process by a location-scale transformation.

Another possibility is to verify the LAN-condition of the Ornstein–Uhlenbeck time series by means of L_2-differentiability in the sense of Jeganathan (1982), Theorem 1, in the special case of LAN instead of LAMN. For the definition of L_2-differentiability we refer to Jeganathan (1982), (2.A.1)–(2.A.5).

Then the random sequence and asymptotic variance from (15) are given by $Z_n = \frac{1}{\sqrt{n}} \sum \Lambda_\theta(x_{\leq j})$, where

$$\Lambda_\theta(x_{\leq j}) = \partial_\theta \log f(x_j | x_{<j}),$$

and $C = I_\theta$ with $I_\theta = \mathrm{E}(\Lambda'_\theta \Lambda_\theta)$.

In case of the Ornstein–Uhlenbeck process, we obtain $\Lambda_\theta = (\Lambda_\mu, \Lambda_\sigma, \Lambda_\lambda)'$ with coordinates

$$\Lambda_\mu = \frac{2\lambda z_j}{\sigma^2(1+e^{-\lambda d})} \tag{16}$$

$$\Lambda_\sigma = -\frac{1}{\sigma} + \frac{2\lambda z_j^2}{\sigma^3(1-e^{-2\lambda d})} \tag{17}$$

$$\Lambda_\lambda = \frac{1}{2\lambda} - \frac{d}{e^{2\lambda d}-1} - \frac{z_j^2 + 2\lambda z_j (x_{j-1} - \mu) d e^{-\lambda d}}{\sigma^2(1-e^{-2\lambda d})} + \frac{\lambda z_j^2 e^{-2\lambda d} 2d}{\sigma^2(1-e^{-2\lambda d})^2}, \tag{18}$$

where $z_j = x_j - \mu - e^{-\lambda d}(x_{j-1} - \mu),\ j = 1, \ldots, n$, and

$$I_\theta = \begin{pmatrix} \frac{2\lambda}{\sigma^2}\frac{1-e^{-\lambda d}}{1+e^{-\lambda d}} & 0 & 0 \\ 0 & \frac{2}{\sigma^2} & -\frac{1}{\lambda\sigma} + \frac{2d}{\sigma(e^{2\lambda d}-1)} \\ 0 & -\frac{1}{\lambda\sigma} + \frac{2d}{\sigma(e^{2\lambda d}-1)} & \frac{1}{2\lambda^2} + \frac{d^2(1+e^{2\lambda d})}{(e^{2\lambda d}-1)^2} - \frac{2d}{\lambda(e^{2\lambda d}-1)} \end{pmatrix}. \quad (19)$$

Proposition 1 *Being L_2-differentiable, the Ornstein–Uhlenbeck time series is LAN with derivative (score-function) Λ_θ and Fisher-Information I_θ, given by* (16)–(19).

From the properties of $(x_j)_{j=1}^n$, it follows that the sequence $(\Lambda_\theta(x_{\le j}))_{j=1}^n$ is a stationary, ergodic martingale difference sequence. And since the sequence $(z_j)_{j=1}^n$ from (17) and (18) depends only on x_j and x_{j-1}, we have that

$$\Lambda_\theta(x_{\le j}) = \Lambda_\theta(x_j, x_{j-1}), \quad (20)$$

which reflects the Markovian property of the Ornstein–Uhlenbeck process.

4.2 Asymptotically linear estimators

We define now asymptotically linear estimators. Since many estimators (e.g. MLE and M-estimates) have an asymptotic expansion as given in (21) below, which is often derived in proofs of asymptotic normality, we define our more general estimators by such an expression.

More precisely, an estimator $S = (S_n)$, $S_n \colon (\Omega_n, \mathcal{F}_n) \to (\mathbb{R}^k, \mathcal{B}^k)$, is asymptotically linear (AL) at θ if there exists an IC ψ_θ, necessarily unique (P_θ-a.s.), such that

$$\sqrt{n}(S_n - \theta) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi_\theta(x_{\le j}) + o_{P_\theta^{(n)}}(n^0) \quad (21)$$

where $o_{P_\theta^{(n)}}(n^0) \to 0$ in product $P_\theta^{(n)}$-probability as $n \to \infty$, and $n^0 = 1$ indicates the index tending to infinity.

Remark 1 For mathematical reasons—ergodicity, stationarity, and the martingale difference property of $(\psi_\theta(x_{\le j}))_{j=1}^n$—we consider in (21) the whole past of $x_{\le j}$, that is, also observations at negative time points. Later, for the implementation, we have to make sure that the influence of $x_{\le 0}$ on S_n is asymptotically negligible.

Moreover, we define the set $\Psi(\theta)$ of all (conditionally centered, square integrable, $\mathbb{R}^k$-valued) ICs by

$$\psi_\theta \in L_2^k(P_\theta), \quad \mathrm{E}_\theta(\psi_\theta(x_{\le j})|x_{<j}) = 0\ P_\theta(dx_{<j})\text{-a.s.}, \quad \mathrm{E}_\theta\, \psi_\theta \Lambda_\theta' = 1_k, \quad (22)$$

where E_θ denotes expectation, respectively conditional expectation, under P_θ, and 1_k is the $k \times k$ identity matrix.

Remark 2 (a) The originally intended interpretation of an IC by Bickel (1984), Hampel et al. (1986), and Huber (1981) is extended to the dependent setup: The summand $\psi_\theta(x_{\le j})$ in (21) represents the asymptotic influence of observation x_j given the past $x_{<j}$ on the standardized estimator $\sqrt{n}(S_n - \theta)$.

(b) The class of AL estimators covers the MLE $\psi_\theta = \widehat{\psi}_\theta := I_\theta^{-1} \Lambda_\theta$, and other M-estimates, see for example Bustos (1982).

(c) Influence curves in the statistical literature have traditionally been defined as Gâteaux derivatives of statistical functionals, see Mancini et al. (2005). But even if a Gâteaux derivative exist it is too weak to cover the empirical measure (since for sample size $n > 1$ the empirical measure is not a Dirac measure), and thus too weak to derive properties from the Gâteaux derivative corresponding to the functional.

The asymptotic derivations for time series can be simplified by the method of Fréchet differentiability, see Bednarski (2010).

Because the Ornstein–Uhlenbeck process is stationary and ergodic, and since $\mathrm{E}_\theta(\psi_\theta(x_{\le j})|x_{<j}) = 0$, the sequence $(\psi_\theta(x_{\le j}))_{j=1}^n$ is a stationary and ergodic martingale difference sequence. Then, with $\psi_\theta \in L_2^k(P_\theta)$, the CLT in McLeish (1974), Theorem 2.3, applies. This gives us the asymptotic normality at θ of the AL estimators under P_θ.

Assuming only $\psi_\theta \in L_2^k(P_\theta)$ and $\mathrm{E}_\theta(\psi_\theta(x_{\le j})|x_{<j}) = 0$, the condition $\mathrm{E}_\theta\, \psi_\theta \Lambda_\theta' = 1_k$ (Fisher-consistency), turns out to be equivalent to a locally uniform extension of this asymptotic normality under P_θ. That is

$$\mathcal{L}_{P_{\theta+h/\sqrt{n}}^{(n)}} \sqrt{n}(S_n - (\theta + h/\sqrt{n})) \to \mathcal{N}(0, \mathrm{Cov}_\theta(\psi_\theta)) \qquad \forall h \in \mathbb{R}^3. \tag{23}$$

This follows by one of LeCam's contiguity lemmas.

Furthermore, we get the Cramér-Rao bound for the asymptotic variance under P_θ

$$\mathrm{Cov}_\theta(\psi_\theta) \ge I_\theta^{-1} = \mathrm{Cov}_\theta(\widehat{\psi}_\theta)$$

with equality iff $\psi_\theta = \widehat{\psi}_\theta = I_\theta^{-1} \Lambda_\theta$.

This easily follows from the positive semidefinity of the expectation and the Fisher-consistency, $\mathrm{E}_\theta\, \psi_\theta \Lambda_\theta' = 1_k$. More precisely

$$\begin{aligned} 0 &\le \mathrm{E}_\theta \left(\psi_\theta - I_\theta^{-1} \Lambda_\theta\right) (\psi_\theta - I_\theta^{-1} \Lambda_\theta)' \\ &= \mathrm{E}_\theta \left(\psi_\theta \psi_\theta'\right) - \mathrm{E}_\theta \left(\psi_\theta \Lambda_\theta' \left(I_\theta^{-1}\right)'\right) - \mathrm{E}_\theta \left(I_\theta^{-1} \Lambda_\theta \psi_\theta'\right) + \mathrm{E}_\theta \left(I_\theta^{-1} \Lambda_\theta \Lambda_\theta' \left(I_\theta^{-1}\right)'\right) \\ &= \mathrm{E}_\theta \left(\psi_\theta \psi_\theta'\right) - I_\theta^{-1}, \end{aligned}$$

since $\mathrm{E}_\theta\left(I_\theta^{-1}\Lambda_\theta\Lambda_\theta'\left(I_\theta^{-1}\right)'\right) = I_\theta^{-1}\,\mathrm{E}_\theta\left(\Lambda_\theta\Lambda_\theta'\right)I_\theta^{-1} = I_\theta^{-1}$.
And we have equality $0 = \mathrm{E}_\theta\left(\psi_\theta - I_\theta^{-1}\Lambda_\theta\right)\left(\psi_\theta - I_\theta^{-1}\Lambda_\theta\right)'$ iff $\psi_\theta = I_\theta^{-1}\Lambda_\theta$, as asserted.

4.3 Infinitesimal perturbations

We are now contaminating the transition distribution of the ideal law P_θ. That means that the observations $x_{\le n}$ may now follow any law Q where the transition distribution is given in terms of the j-th transition distribution of $Q^{(n)}$, denoted by $Q^{(n;j|<j)}(dx_j|x_{<j})$, and the ideal initial distribution $Q^{(n,0)}(dx_{\le 0}) = P_\theta(dx_{\le 0})$.

More precisely, the joint law $Q^{(n)}$ is for all $f\colon (\mathbb{R}^{\le 0}\times\cdots\times\mathbb{R}, \mathbb{B}^{\le 0}\otimes\cdots\otimes\mathbb{B}) \to (\mathbb{R}^+, \mathbb{B})$ given by

$$\int\limits_{\mathbb{R}^{\le n}} f(x_{\le n})\,Q^{(n)}(dx_{\le n}) = \int\limits_{\mathbb{R}^{\le 0}}\cdots\int\limits_{\mathbb{R}} f(x_{\le n})\prod_{j=1}^{n} Q^{(n;j|<j)}(dx_j|x_{<j})\;Q^{(n,0)}(dx_{\le 0}).$$

Furthermore $Q^{(n;j|<j)}(dx_j|x_{<j})$ belongs to a neighborhood about the j-th ideal transition distribution, $P_\theta^{(n;j|<j)}(dx_j|x_{<j})$, with a radius, depending on the past.
In this article, the neighborhoods will be restricted to convex combination type. We consider contamination balls

$$B_{c,\varepsilon}\left(P_\theta^{(n;j|<j)}(dx_j|x_{<j}), r/\sqrt{n}\right)$$

consisting of all convex combinations

$$\left(1 - r/\sqrt{n}\;\varepsilon(x_{<j})\right)P_\theta^{(n;j|<j)}(dx_j|x_{<j}) + r/\sqrt{n}\;\varepsilon(x_{<j})H(dx_j|x_{<j}), \qquad (24)$$

where $H(dx_j|x_{<j})$ is any Markov kernel, $\varepsilon(x_{<j}) \ge 0$ a bounded and measurable function, a so-called contamination curve, see Bickel (1984) and Rieder (1994).

The factor $1/\sqrt{n}$ implies that the contamination neighborhoods are shrinking as $n\to\infty$. For a derivation of this factor in the i.i.d. case see Ruckdeschel (2006).
Taking the union over all convex contamination balls with respect to $\|\varepsilon\|_1 \le 1$, we obtain the average convex contamination ball, denoted by

$$B_{c,1}\left(P_\theta^{(n;j|<j)}(dx_j|x_{<j}), r/\sqrt{n}\right) = \bigcup_{\|\varepsilon\|_1\le 1} B_{c,\varepsilon}\left(P_\theta^{(n;j|<j)}(dx_j|x_{<j}), r/\sqrt{n}\right).$$

We now introduce the bounded, conditionally centered tangents

$$Z_\infty(\theta) = \left\{q\in L_\infty(P_\theta)\mid \mathrm{E}_\theta(q(x_{\le 1})|x_{<1}) = 0\;\; P_\theta(dx_{<1})\text{-a.s.}\right\}.$$

Along any $q \in Z_\infty(\theta)$, simple perturbations from the ideal transition distribution are for each $x_{<j}$ defined by

$$Q_{n,q}^{(n;j|<j)}(dx_j|x_{<j}) = \left(1 + r/\sqrt{n}\, q(x_{\le j})\right) P_\theta^{(n;j|<j)}(dx_j|x_{<j}),$$

provided that $\sqrt{n} \ge -r \inf_{P_\theta} q$, where $\inf_{P_\theta}$ denotes the P_θ-essential infimum. It holds that, for each $x_{<j}$

$$Q_{n,q}^{(n;j|<j)}(dx_j|x_{<j}) \in B_{c,\varepsilon}\left(P_\theta^{(n;j|<j)}(dx_j|x_{<j}), r/\sqrt{n}\right) \Longleftrightarrow$$
$$q(x_j, x_{<j}) \ge -\varepsilon(x_{<j}) \quad P_\theta(dx_j|x_{<j})^{(n;j|<j)}\text{-a.s.}$$

As a consequence of LAN and the AL expansion in (21) AL estimators, under such simple perturbations, are still asymptotically normal

$$\mathcal{L}_{Q_{n,q}^{(n)}} \sqrt{n}(S_n - \theta) \to \mathcal{N}_k(r\, \mathrm{E}_\theta\, \psi_\theta q, \mathrm{Cov}_\theta\, \psi_\theta), \tag{25}$$

but we obtain the bias $r\, \mathrm{E}_\theta\, \psi_\theta q$. This is shown similarly to (23).

Remark 3 The asymptotic behavior of M-estimates for general ergodic diffusion processes is studied in Sakamoto and Yoshida (1998), and under deviations from the ideal law in Yoshida (1990). Contrary to our approach, Yoshida (1990) obtains no bias in the asymptotic law under model deviations and does not treat neighborhoods of transition distributions.

We denote the standardized maximal asymptotic bias under shrinking neighborhoods $B_{c,1}$ by

$$\omega_{c,1}(\psi) = \sup_q |\,\mathrm{E}_\theta\, \psi_\theta q|, \tag{26}$$

provided that $q \in Z_\infty(\theta)$, and $q(x_j, x_{<j}) \ge -\varepsilon(x_{<j})\ P_\theta(dx_j|x_{<j})^{(n;j|<j)}$-a.s.

4.4 Maximum asymptotic MSE

In the case of convex contamination the standardized maximal asymptotic bias from (26) can be explicitly calculated as

$$\omega_{c,1}(\psi_\theta) = \operatorname{ess\,sup} |\psi_\theta|, \tag{27}$$

see Bickel (1984), Rieder (1994), and Staab (1984).

Our aim is to minimize the maximum asymptotic mean square error (maxasyMSE) on shrinking neighborhoods about the ideal model.

Applying the weak convergence from (25) we may write this maxasyMSE, $\text{maxasyMSE}(S_n) = \sup_q \lim_{n\to\infty} \mathrm{E}_{Q^n}\left(|\sqrt{n}(S_n - \theta)|^2\right)$, as

$$\text{maxasyMSE}(S_n) = r^2\omega_{c,1}^2(\psi_\theta) + \operatorname{tr}\operatorname{Cov}_\theta \psi_\theta \tag{28}$$

with $\omega_{c,1}(\psi_\theta)$ from (27).

4.5 Optimal influence curves

The optimally robust IC ψ_θ^* is the unique minimizer of $\text{maxasyMSE}(S_n)$ from (28) among all $\psi_\theta \in \Psi(\theta)$.

In the case of the Ornstein–Uhlenbeck process, where $\Lambda_\theta = \Lambda_\theta(x_1, x_0)$, see (20), we obtain the following sufficient condition for ψ_θ^*.

Theorem 1 *Assume there exist some* $a\colon \mathbb{R} \to \mathbb{R}^3$, $A \in \mathbb{R}^{3\times 3}$, *and* $b \in [0,\infty)$ *such that, denoting*

$$w(x_1, x_0) = \min\{1, b|A\Lambda_\theta(x_1, x_0) - a(x_0)|^{-1}\},$$

$$\mathrm{E}((A\Lambda_\theta(x_1, x_0) - a(x_0))w(x_1, x_0)|x_0) = 0, \tag{29}$$

$$A^{-1} = \mathrm{E}\left((\Lambda_\theta(x_1, x_0) - a(x_0))(\Lambda_\theta(x_1, x_0) - a(x_0))'w(x_1, x_0)\right), \tag{30}$$

$$r^2 b = \mathrm{E}(|A\Lambda_\theta(x_1, x_0) - a(x_0)| - b)_+. \tag{31}$$

Then

$$\psi_\theta^*(x_1, x_0) = (A\Lambda_\theta(x_1, x_0) - a(x_0))w(x_1, x_0) \tag{32}$$

is an IC and the optimal one.

Proof We first verify that ψ_θ^* is an IC. The conditional centeredness of ψ_θ^* is (29). For the Fisher-consistency of ψ_θ^*, we apply that, by the iterated expectation property and since $\mathrm{E}(\Lambda_\theta|x_0) = 0$, we have

$$\mathrm{E}\left((\Lambda_\theta - a)w\Lambda_\theta'\right) = \mathrm{E}\left(\Lambda_\theta w\Lambda_\theta'\right).$$

Therefore, and by (30), we have $\mathrm{E}(\psi_\theta^*\Lambda_\theta') = A\,\mathrm{E}\left((\Lambda_\theta - a)(\Lambda_\theta - a)'w\right) = 1_3$, the desired Fisher-consistency.

The uniqueness of ψ_θ^* follows from $\Psi(\theta)$ in (22) being a nonempty, closed, convex subset of the Hilbert space L_2, and the fact that every nonempty, closed, convex subset of a Hilbert space contains a unique element of smallest norm.

With (28) it remains now to show that ψ_θ^* minimizes

$$\mathrm{E}\,|\psi_\theta|^2 + r^2 \operatorname{ess\,sup}|\psi_\theta|, \tag{33}$$

among all $\psi_\theta \in \Psi(\theta)$.

In a first step, we rewrite (33). For $\psi_\theta \in \Psi(\theta)$ it is

$$\mathrm{E}(|\psi_\theta - A\Lambda_\theta - a|^2) = \mathrm{E}(|\psi_\theta|^2) + \mathrm{E}(|A\Lambda_\theta - a|^2) - 2\,\mathrm{E}(\psi'_\theta(A\Lambda_\theta - a)). \quad (34)$$

The last term in (34) may be further written as

$$\begin{aligned}\mathrm{E}\left(\psi'_\theta(A\Lambda_\theta - a)\right) &= \mathrm{E}\left(\psi'_\theta A\Lambda_\theta\right) - \mathrm{E}\left(a'\psi_\theta\right) = \mathrm{E}\left(\mathrm{tr}\left(\psi'_\theta A\Lambda_\theta\right)\right) - \mathrm{E}\left(\mathrm{E}(a'\psi_\theta|x_0)\right)\\ &= \mathrm{E}(\mathrm{tr}(A\Lambda_\theta\psi'_\theta)) = \mathrm{tr}A\,\mathrm{E}(\psi_\theta\Lambda_\theta) = \mathrm{tr}A,\end{aligned}$$

since $\mathrm{E}(\psi_\theta|x_0) = 0$ and $\mathrm{E}(\psi_\theta\Lambda_\theta) = 1_3$, see (22).
Therefore (34) is equal to

$$\mathrm{E}(|\psi_\theta - A\Lambda_\theta - a|^2) = \mathrm{E}(|\psi_\theta|^2) + c,$$

where c is a constant not depending on ψ_θ.
That means that minimizing (33) is equal to minimize, among all $\psi_\theta \in \Psi(\theta)$,

$$\mathrm{E}(|\psi_\theta - A\Lambda_\theta - a|^2) + r^2\mathrm{ess\,sup}\,|\psi_\theta|. \quad (35)$$

In a second step, we minimize (35) for fixed $b := \mathrm{ess\,sup}\,|\psi_\theta|$.
We obtain that the minimum of $\mathrm{E}(|\psi_\theta - A\Lambda_\theta - a|^2)$ is

$$\psi_b = \begin{cases} A\Lambda_\theta - a & \text{if } A\Lambda_\theta - a \le b \\ b\,\frac{A\Lambda_\theta - a}{|A\Lambda_\theta - a|} & \text{else} \end{cases}. \quad (36)$$

In a last step, we minimize with respect to b. Since ψ_b is the minimum and because of (36), we have

$$\begin{aligned}\mathrm{E}(|\psi_\theta - A\Lambda_\theta - a|^2) + r^2b^2 &\ge \mathrm{E}(|\psi_b - A\Lambda_\theta - a|^2) + r^2b^2\\ &= \mathrm{E}(|A\Lambda_\theta - a| - b)_+^2 + r^2b^2. \quad (37)\end{aligned}$$

The integrand in $\mathrm{E}(|A\Lambda_\theta - a)| - b)_+^2$ being a convex function in b, we may interchange integration and differentiation. Therefore differentiating (37) with respect to b gives

$$-2\,\mathrm{E}(|A\Lambda_\theta - a| - b)_+ + 2r^2b = 0,$$

that is condition (31). □

5 Calculation of influence curves

We are now comparing graphically the ICs of the two approaches.

5.1 Influence curves of M-estimates

In order to find the ICs of M-estimates, we first have to derive an asymptotic expression as in (21).

We start with the M-equations given in (10), (12), and (13). Since the initial terms in (10) and (12) are negligibly small compared to the sum, we may simplify the system of equations to

$$\sum_{i=1}^{n} \psi\left(\hat{r}_i\right) = 0 \tag{38}$$

$$\sum_{i=1}^{n} \chi\left(\hat{r}_i\right) = 0 \tag{39}$$

$$-\frac{\sqrt{2\hat{\lambda}}}{\hat{\sigma}} \sum_{i=1}^{n} \psi\left(\hat{r}_i\right)(x_{i-1} - \hat{\mu}) + \frac{e^{-\hat{\lambda}d}}{\sqrt{1-e^{-2\hat{\lambda}d}}} \sum_{i=1}^{n} \chi\left(\hat{r}_i\right) = 0 \tag{40}$$

with the arguments of ψ and χ given by

$$\hat{r}_i = \frac{\sqrt{2\hat{\lambda}}\left((x_i - \hat{\mu}) - e^{-\hat{\lambda}d}(x_{i-1} - \hat{\mu})\right)}{\hat{\sigma}\sqrt{1 - e^{-2\hat{\lambda}d}}} \quad i = 1, \ldots, n.$$

Introducing the function $\zeta : \mathbb{R}^{2\times 3} \to \mathbb{R}^3$, we may write (38), (39), and (40) as

$$\sum_{i=1}^{n} \zeta_{\hat{\theta}}(x_i, x_{i-1}) = \underline{0}, \tag{41}$$

where $\hat{\theta} = (\hat{\mu}, \hat{\sigma}, \hat{\lambda})'$, $\underline{0} = (0, 0, 0)'$, and

$$\zeta_{\hat{\theta}} = \begin{pmatrix} \zeta_{\hat{\theta},1} \\ \zeta_{\hat{\theta},2} \\ \zeta_{\hat{\theta},3} \end{pmatrix} = \begin{pmatrix} \psi\left(\hat{r}_i\right) \\ \chi\left(\hat{r}_i\right) \\ -\frac{\sqrt{2\hat{\lambda}}}{\hat{\sigma}} \psi\left(\hat{r}_i\right)(x_{i-1} - \hat{\mu}) + \frac{e^{-\hat{\lambda}d}}{\sqrt{1-e^{-2\hat{\lambda}d}}} \chi\left(\hat{r}_i\right) \end{pmatrix}. \tag{42}$$

Applying now a first order Taylor expansion about $\theta = (\mu, \sigma, \lambda)'$ to (41), we get

$$\sum_{i=1}^{n} \zeta_{\theta}(x_i, x_{i-1}) + \sum_{i=1}^{n} D\zeta_{\theta}(x_i, x_{i-1})(\hat{\theta} - \theta) \approx \underline{0}, \tag{43}$$

where the components of ζ_θ are given by (42), $\hat{\theta}$ replaced by θ, and with the Jacobian

$$D\zeta_\theta(x_i, x_{i-1}) = \begin{pmatrix} \partial_\mu \zeta_{\theta,1} & \partial_\sigma \zeta_{\theta,1} & \partial_\lambda \zeta_{\theta,1} \\ \partial_\mu \zeta_{\theta,2} & \partial_\sigma \zeta_{\theta,2} & \partial_\lambda \zeta_{\theta,2} \\ \partial_\mu \zeta_{\theta,3} & \partial_\sigma \zeta_{\theta,3} & \partial_\lambda \zeta_{\theta,3} \end{pmatrix}.$$

Solving (43) for $\hat{\theta} - \theta$ and multiplying afterwards by $\sqrt{n}$ gives

$$\sqrt{n}(\hat{\theta} - \theta) \approx \left(\frac{1}{n} \sum_{i=1}^{n} D\zeta_\theta(x_i, x_{i-1}) \right)^{-1} \left(-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \zeta_\theta(x_i, x_{i-1}) \right). \tag{44}$$

Since the explicit inversion of $\frac{1}{n} \sum_{i=1}^{n} D\zeta_\theta(x_i, x_{i-1})$ is too time-consuming, we proceed as follows: We first apply the ergodic theorem to the matrix $\frac{1}{n} \sum_{i=1}^{n} D\zeta_\theta(x_i, x_{i-1})$, then simplify by integration by parts, and invert the resulting matrix afterwards.

Since the $(x_i)_{i=1}^{\infty}$ are stationary and ergodic, also functions $g(x_i, x_{i+1}, \ldots)$, $i = 1, \ldots, \infty$, have these properties, see Breiman (1992), Proposition 6.6. Therefore r_i, $\psi(r_i)$, and $\chi(r_i)$, $i = 1, \ldots, \infty$, are also stationary and ergodic, and the ergodic theorem applies, such that

$$\frac{1}{n} \sum_{i=1}^{n} D\zeta_\theta(x_i, x_{i-1}) \xrightarrow{a.s.} \mathrm{E}(D\zeta_\theta(x_1, x_0)). \tag{45}$$

Further we may write this limiting matrix, assumed finite, as

$$\begin{aligned} \mathrm{E}(D\zeta_\theta(x_1, x_0)) &= \int D\zeta_\theta(x_1, x_0) f(x_1, x_0) d(x_1, x_0) \\ &= \int \left(\int D\zeta_\theta(x_1, x_0) f(x_1|x_0) dx_1 \right) f(x_0) dx_0. \end{aligned}$$

We continue with an integration by parts on the inner integral

$$\begin{aligned} \int D\zeta_\theta(x_1, x_0) f(x_1|x_0) dx_1 &= -\int \zeta_\theta(x_1, x_0) \frac{\partial_\theta f(x_1|x_0)}{f(x_1|x_0)} f(x_1|x_0) dx_1 \\ &= -\int \zeta_\theta(x_1, x_0) \Lambda'_{1,\theta} f(x_1|x_0) dx_1 \end{aligned}$$

with $\Lambda_{1,\theta} := \partial_\theta \log f(x_1|x_0)$.
Therefore we showed that

$$\begin{aligned} \mathrm{E}(D\zeta_\theta(x_1, x_0)) &= -\int \left(\int \zeta_\theta(x_1, x_0) \Lambda'_{1,\theta} f(x_1|x_0) dx_1 \right) f(x_0) dx_0 \\ &= -\mathrm{E}\left(\zeta_\theta(x_1, x_0) \Lambda'_{1,\theta} \right), \end{aligned}$$

and by (45) it is

$$\frac{1}{n}\sum_{i=1}^{n} D\zeta_\theta(x_i, x_{i-1}) \xrightarrow{a.s.} -\mathrm{E}\left(\zeta_\theta(x_1, x_0)\Lambda'_{1,\theta}\right). \tag{46}$$

Remark 4 In addition the CLT for stationary and ergodic martingale differences from McLeish (1974) may be applied on the second term in (44); this results in the asymptotic normality of $\sqrt{n}(\hat{\theta} - \theta)$.

The sequence $(\zeta_\theta(x_i, x_{i-1}))_{i=1}^{\infty}$ is stationary and ergodic. For the martingale difference property we show that

$$\mathrm{E}(\zeta_\theta(x_i, x_{i-1})|x_{<i}) = 0, \tag{47}$$

since conditioning once more and by the properties of the conditional expectation (47) implies $\mathrm{E}(\zeta_\theta(x_i, x_{i-1})|\zeta_\theta(x_{i-1}, x_{i-2}), \zeta_\theta(x_{i-2}, x_{i-3}), \ldots) = 0$, the desired martingale difference property.

Now let $|r_i| \le k$, hence $\psi(r_i) = r_i$ and $\chi(r_i) = r_i^2 - 1$.
By definition r_i is only a function of x_i and x_{i-1}, further $\mathrm{E}(x_i|x_{i-1}) = \mu + e^{-\lambda d}(x_{i-1} - \mu)$, and then we have

$$\mathrm{E}(r_i|x_{<i}) = \mathrm{E}(r_i|x_{i-1}) = \mathrm{E}(x_i|x_{i-1}) - \mathrm{E}\left(\mu + e^{-\lambda d}(x_{i-1} - \mu)\right) = 0.$$

Applying this and $\mathrm{Var}(x_i|x_{i-1}) = \frac{\sigma^2}{2\lambda}(1 - e^{-2\lambda d})$ leads further to

$$\begin{aligned}\mathrm{E}(r_i^2|x_{<i}) &= \mathrm{Var}(r_i|x_{<i}) + \mathrm{E}(r_i|x_{<i})^2 \\ &= \frac{2\lambda}{\sigma^2(1 - e^{-2\lambda d})}(\mathrm{Var}(x_i|x_{i-1}) + e^{-2\lambda d}\,\mathrm{Var}(x_{i-1}|x_{i-1}) \\ &\quad -2e^{-\lambda d}\,\mathrm{Cov}(x_i, x_{i-1}|x_{i-1})) = 1,\end{aligned}$$

since $\mathrm{Var}(x_{i-1}|x_{i-1}) = 0$ and $\mathrm{Cov}(x_i, x_{i-1}|x_{i-1}) = 0$.

Therefore we have $\mathrm{E}(\psi(r_i)|x_{<i}) = 0$, $\mathrm{E}(\chi(r_i)|x_{<i}) = 0$, and we may deduce (47). Then the CLT of McLeish applies such that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \zeta_\theta(x_i, x_{i-1}) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathrm{E}\,\zeta_\theta\zeta_\theta'\right).$$

Now with (44) we have the following ICs of M-estimates.

Lemma 1 *Denoting ζ_θ by (42), the ICs $\psi_\theta : \mathbb{R}^2 \to \mathbb{R}^3$ of M-estimates are given by*

$$\psi_\theta(x_i, x_{i-1}) = \left(\mathrm{E}\left(\zeta_\theta(x_1, x_0)\Lambda'_{1,\theta}\right)\right)^{-1}\zeta_\theta(x_i, x_{i-1}).$$

For the graphical illustration of these curves we consider $\psi_\theta(x_1, x_0) = \zeta_\theta(x_1, x_0)$, that is, the components $\psi_\mu(x_1, x_0) = \psi(r_1)$, $\psi_\sigma(x_1, x_0) = \chi(r_1)$, and $\psi_\lambda(x_1, x_0) =$

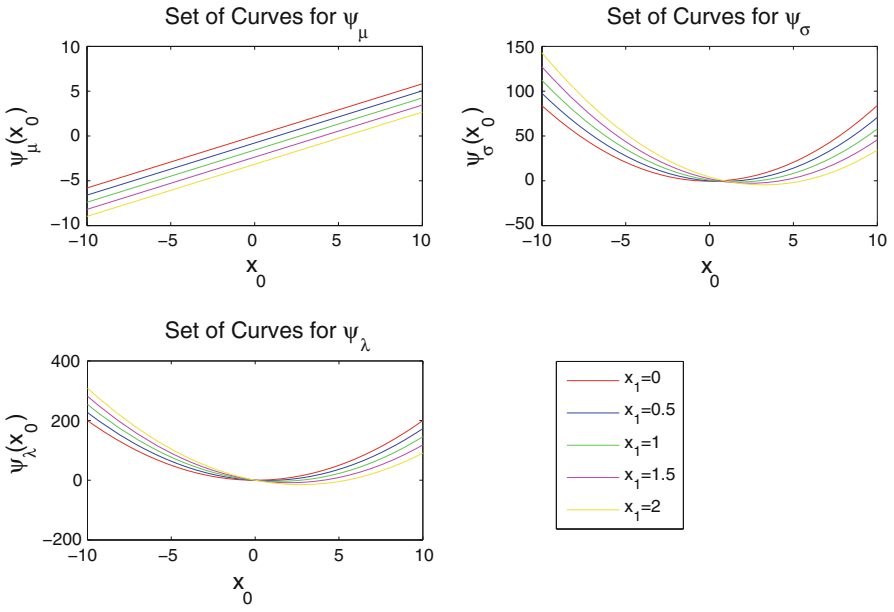

Fig. 1 Classical influence *curves*

$-\frac{\sqrt{2\lambda}}{\sigma}\psi\,(r_1)\,(x_0-\mu)+\frac{e^{-\lambda d}}{\sqrt{1-e^{-2\lambda d}}}\chi\,(r_1)$, where the functions ψ and χ are given by (8) and (14), respectively.

In addition, the parameters of the process are fixed to $\mu = 0$, $\sigma = 1$, $\lambda = 1$, and the time lag to $d = 1$. The ICs $\psi_\theta(x_1, x_0)$ are plotted as functions of x_0 for fixed x_1, $x_1 = 0, 0.5, 1, 1.5, 2$ and for two cases: The classical case (for $k = 10$), and in a second case the cut-off point k is set to $k = 2$.

Figure 1 shows the set of curves in the classical case: We obtain a straight line for ψ_μ^*, a parabola for ψ_σ^*, and we also get a parabola for ψ_λ^*, however with a smaller curvature than ψ_σ^*. We remark that in the classical case all curves are unbounded.

Figure 2 summarizes the results for $k = 2$: The IC for μ is a cut-off line, both ψ_σ and ψ_λ are parabolas, the λ-component is however not bounded.

Remark 5 Choice of the time lag: For the estimation of the parameters of a general diffusion process sampled at discrete time, different schemes of observation are possible, see Iacus (2008). Moreover, this choice influences the estimation results, see Rao (1999).

In this article, we focus on the large sample scheme, where d is fixed and $n \to \infty$, in particular, we set $d = 1$.

However, one should keep in mind that different choices of the time lag are possible; the effect of this choice on the ICs and the estimation results is not investigated in this article but would be worthwhile to study in future research.

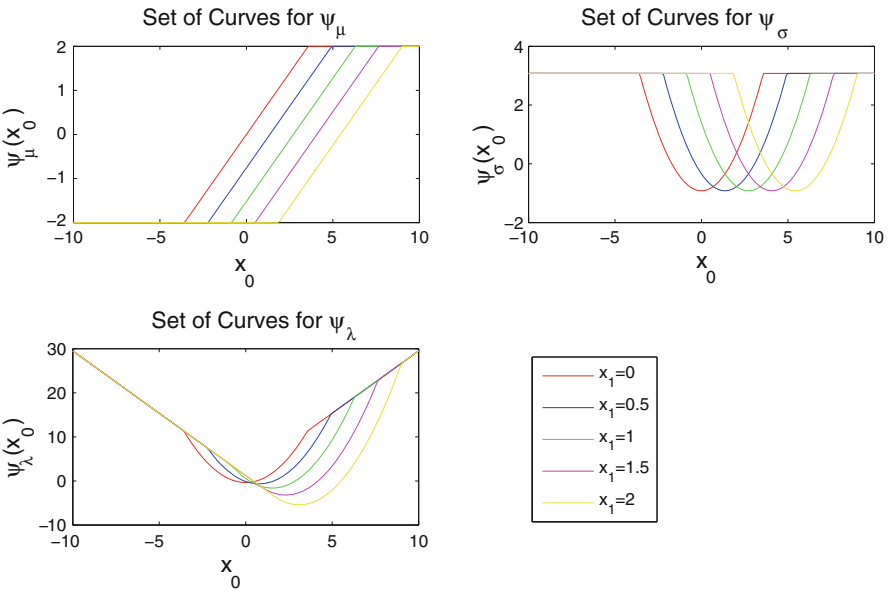

Fig. 2 Influence *curves* of M-estimates

5.2 Algorithm for optimal influence curves

For the optimally robust ICs, we have to solve the equations given in Theorem 1 numerically. Therefore, we transform (29) into a fixed-point equation.

With some basic calculus and introducing the function $z\colon \mathbb{R} \to \mathbb{R}^3$, (29) is equivalent to $a(x_0) = Az(x_0)$, where

$$z(x_0) = \frac{\mathrm{E}(\Lambda_\theta(x_1, x_0)w(x_1, x_0)|x_0)}{\mathrm{E}(w(x_1, x_0)|x_0)}. \tag{48}$$

For the numerical solution of (31), (30), and (48), we start with the initial values $b_0 \in (0, \infty)$, $A_0 = I_\theta^{-1}$, $a_0 = (0, 0, 0)'$, and

$$w_0(x_1, x_0) = \min\left\{1, b_0|I_\theta^{-1}\Lambda_\theta(x_1, x_0)|^{-1}\right\}.$$

Then, in a first iteration step,

$$\begin{aligned} z_1(x_0) &= \tfrac{\mathrm{E}(\Lambda_\theta(x_1,x_0)w_0(x_1,x_0)|x_0)}{\mathrm{E}(w_0(x_1,x_0)|x_0)}, \\ A_1^{-1} &= \mathrm{E}(\Lambda_\theta(x_1, x_0) - z_1(x_0))(\Lambda_\theta(x_1, x_0) - z_1(x_0))'w_0(x_1, x_0), \\ a_1(x_0) &= A_1 z_1(x_0), \end{aligned}$$

and

$$r^2 b_1 - \mathrm{E}(|A_1 \Lambda_\theta(x_1, x_0) - a_1(x_0)| - b_1)_+ = 0.$$

In a second step, we may calculate

$$w_1(x_1, x_0) = \min\left\{1, b_1 |A_1 \Lambda_\theta(x_1, x_0) - a_1(x_0)|^{-1}\right\}.$$

The n-th iteration step, $n \geq 2$, is given by

$$z_n(x_0) = \frac{\mathrm{E}(\Lambda_\theta(x_1,x_0) w_{n-1}(x_1,x_0)|x_0)}{\mathrm{E}(w_{n-1}(x_1,x_0)|x_0)}, \tag{49}$$

$$A_n^{-1} = \mathrm{E}(\Lambda_\theta(x_1, x_0) - z_n(x_0))(\Lambda_\theta(x_1, x_0) - z_n(x_0))' w_{n-1}(x_1, x_0), \tag{50}$$

$$a_n(x_0) = A_n z_n(x_0),$$

and

$$r^2 b_n - \mathrm{E}(|A_n \Lambda_\theta(x_1, x_0) - a_n(x_0)| - b_n)_+ = 0. \tag{51}$$

Then we have $w_n(x_1, x_0) = \min\{1, b_n |A_n \Lambda_\theta(x_1, x_0) - a_n(x_0)|^{-1}\}$, and the n-step numerical approximation of the optimal IC is

$$\psi^*_{\theta,n}(x_1, x_0) = (A_n \Lambda_\theta(x_1, x_0) - a_n(x_0)) w_n(x_1, x_0).$$

Remark 6 Solving the equations given in Theorem 1 numerically, the n-th approximation of the optimal IC is not yet an IC. Conditional centering and Fisher-consistency however, may be obtained easily by standardization. A further issue, not treated here, would be to investigate how close one gets this way to the minimum maxasyMSE. So far we rely on $\psi^*_{\theta,n}$ approximating the optimal IC.

We solved the equations (49), (50), and (51) with Matlab. For the numerical integration we applied the method quad, based on the recursive adaptive Simpson quadrature method.

The exact numerical evaluation of the conditional expectation in (49) was however too time-consuming, and we proceeded as follows: First, we evaluated the function $z_n(x_0)$ on a grid, afterwards we plugged in the interpolation in (50).

For the numerical solution of (49), (50), and (51), the parameters of the process were fixed to $\mu = 0$, $\sigma = 1$, $\lambda = 1$, and the time lag to $d = 1$.

We solved the system of equations for the radius $r = 0.02$. The algorithm converged fast, when going from the third to the fourth iteration we had a change of 10^{-6} with respect to b and the matrix A.

We plotted the optimal ICs ψ^*_θ as a function of x_0 for fixed values of x_1, $x_1 = 0, 0.5, 1, 1.5, 2$.

Figure 3 summarizes the set of curves with the choice $b_0 = 1$ and $r = 0.02$. Even for this small radius, we see already a remarkably big deviation from the classical case: ψ^*_μ is a cut-off line and redescending, and both ψ^*_λ and ψ^*_σ become bounded parabolas.

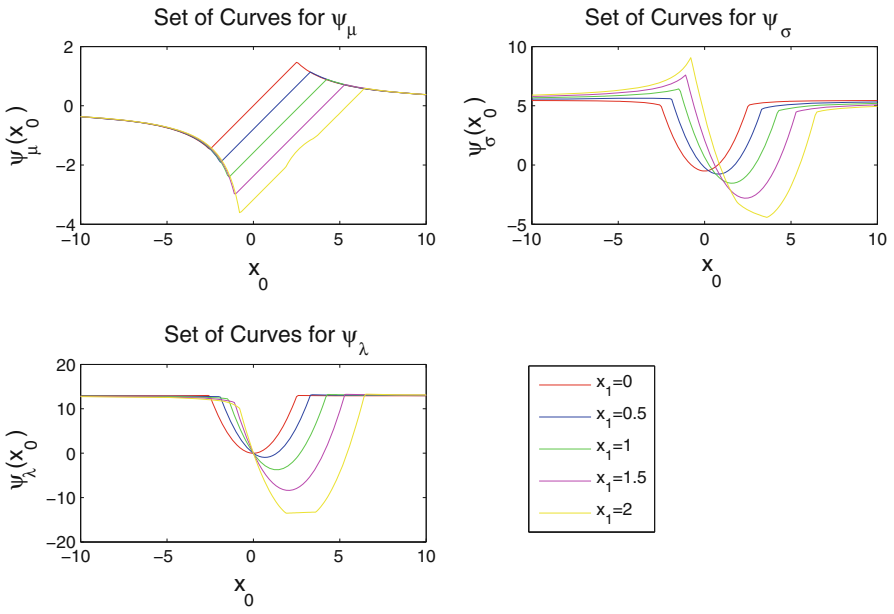

Fig. 3 Optimal influence *curves* with $r = 0.02$

This effect is similar to the one described in Kohl et al. (2010): There the normal location and scale model was studied, that is $P_\theta = \mathcal{N}(\mu, \sigma^2)$. Our plot of ψ_μ^* is consistent with the location part in Kohl et al. (2010), ψ_λ^* with the scale part, and ψ_σ^* is between the two.

5.3 Comparison

We derived the optimally robust IC and the IC of M-estimates, both extending the classical IC. From Figs. 1, 2 and 3 we see: The classical IC is not bounded, the IC of M-estimates is not bounded either, as seen from the λ-component in Fig. 2.

Comparing the IC with respect to μ of M-estimates with the optimally robust IC, there is a surprising difference: The optimally robust IC for μ is redescending, contrary to the clipping in the M-approach.

Moreover, the graph of the optimally robust ICs is more refined than the graph of the ICs of M-estimates.

6 Estimator construction, simulations, and open problems

In this section, we construct, given the optimally robust IC ψ_θ^*, one for each $\theta \in \mathbb{R}^3$, an estimator $S^* = (S_n^*)$ that should be AL at each θ, with IC ψ_θ^*. In addition, when passing from the submodel to full neighborhoods, the maxasyMSE should not increase. In the independent case, this construction was treated with considerable technical effort and complexity in Rieder (1994, Chapter 6) using M-estimates, minimum

distance (MD)-estimates and One-Step estimates based on suitable initial estimates. However, our approach is not to get into this theoretical complexity, but to carry out in our context the One-Step construction dating back to LeCam (1986), and to analyze it by means of simulations. This part of the article is therefore rather intuitively and theoretically incomplete; the remaining theoretical problems are summarized in Sect. 6.2 for future research.

On the theoretical side, Staab (1984) reports a counterexample for an $AR(1)$-process in which the risk does increase when passing from the submodel to the full model. This is due to the instability of the asymptotic variance, and contrary to the independent case. Therefore, in the dependent case, the estimator construction is more demanding and one might have to settle down on some neighborhood submodels.

6.1 Estimator construction and simulations

In this subsection, we carry out the One-Step construction based on the optimal IC and compare it afterwards with MoM, MLE, and M-estimates in a simulation study for different contamination models.

More precisely, we consider the following one-step estimator S

$$S_n = \theta_0 + \frac{1}{n}\sum_{i=1}^{n} \psi^*_{\theta_0}(x_i, x_{i-1}), \tag{52}$$

where $\psi^*_{\theta_0}$ denotes the optimally robust IC, numerically obtained from Theorem 1, and $\theta_0 = \theta_0(x_0, x_1, \ldots, x_n)$ is an initial estimate.

Remark 7 (a) Choice of the initial estimate: The One-Step estimator in (52) heavily depends on the choice of the initial estimator θ_0, which is required $\sqrt{n}$-consistent, uniformly on the neighborhoods. In the independent case, this influence disappears asymptotically and minimum distance (MD)-estimators were employed as initial estimates. However, in the dependent case, the uniformly consistency of those estimators is not yet studied on neighborhoods. In our simulation study we consider therefore an intuitively robustified MoM estimator as initial estimator. In addition, the M-estimator given by the system of Eqs. (10), (12), and (13) is used as initial estimator.

(b) AL: It remains to show that the One-Step estimator in (52) is asymptotically linear. In the independent case this has been shown in Rieder (1994), and for ARMA time series this has been shown in Staab (1984).

For the simulation study, as contamination models, we consider an additive outlier (AO)-model, and a contamination of the ideal transition distributions by Dirac measures. As estimates we calculate MoM, MLE, M-estimates, and the One-Step estimator from (52) based on $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\sigma}_0, \hat{\lambda}_0)'$. The components $\hat{\mu}_0$ and $\hat{\sigma}_0$ are obtained by replacing the mean $\bar{x}$ and the standard deviation s in the MoM estimates by med and mad, the median and the median absolute deviation, that is

Table 1 Different estimators for an ideal Ornstein–Uhlenbeck process with $k = b_0 = 1$, and rad $= 0$

	Mean	Median	SD	RMSE
MoM				
$\mu = 0$	−0.0425	−0.0531	0.1167	0.3751
$\sigma = 1$	2.6276	2.7605	0.3786	5.2709
$\lambda = 1$	7.0969	7.2824	1.8052	20.0263
MLE				
$\mu = 0$	−0.0440	−0.0503	0.1162	0.3752
$\sigma = 1$	0.9730	1.0018	0.1024	0.3189
$\lambda = 1$	0.9893	0.9857	0.2256	0.6776
M-estimate				
$\mu = 0$	−0.0453	−0.0255	0.1131	0.3685
$\sigma = 1$	0.9578	0.9563	0.1017	0.3331
$\lambda = 1$	0.9566	0.9766	0.2011	0.6786
One-step(M)				
$\mu = 0$	−0.0413	−0.0511	0.1186	0.3791
$\sigma = 1$	0.9748	0.9975	0.1046	0.3238
$\lambda = 1$	0.9778	0.9695	0.2274	0.6858
$\hat{\theta}_0$				
$\mu = 0$	−0.0386	−0.0364	0.1156	0.3676
$\sigma = 1$	0.5483	0.5716	0.0745	1.4459
$\lambda = 1$	0.4800	0.4980	0.0921	1.6674
One-step ($\hat{\theta}_0$)				
$\mu = 0$	−0.0417	−0.0475	0.1201	0.3836
$\sigma = 1$	0.8891	0.9014	0.1031	0.4675
$\lambda = 1$	0.8911	0.9029	0.1770	0.6328

$$\hat{\mu}_0 = \mathrm{med}(x), \quad \hat{\sigma}_0 = \sqrt{2\hat{\lambda}}\,\mathrm{mad}(x), \tag{53}$$

whereas the component $\hat{\lambda}_0$ is obtained by minimizing numerically the l_1 distance between empirical and theoretical autocorrelation of time lag k, that is by minimizing the sum $\sum |\exp(-\lambda k d) - \widehat{\mathrm{ACF(k)}}|$.

This estimator and in addition the M-estimator are used as initial estimate for the construction in (52), resulting in One-Step estimates denoted by One-Step ($\hat{\theta}_0$), and One-Step (M), respectively.

Table 1 summarizes the estimation results in case of the ideal law with cut-off points set to $k = b_0 = 1$, as well as the estimators mean, median, standard deviation, and root MSE (RMSE).

The σ and λ components of the MLE being close to the exact values and having a small RMSE, the MLE is preferable to the other estimates in the uncontaminated case. The estimator One-Step (M) performs also well, in particular it improves the M-estimate. The initial estimator $\hat{\theta}_0$, however, heavily underestimates each of the parameter

Table 2 Different estimators for an AO-model (AO $\sim \mathcal{N}(3, 1)$)

	Mean	Median	SD	RMSE
MoM				
$\mu = 0$	0.1583	0.1918	0.1122	0.6032
$\sigma = 1$	3.6925	3.7336	0.5562	8.6764
$\lambda = 1$	7.2224	6.4953	2.3949	20.9475
MLE				
$\mu = 0$	0.1577	0.1913	0.1126	0.6023
$\sigma = 1$	1.8789	1.7734	0.5837	3.2849
$\lambda = 1$	1.8622	1.5919	0.7969	3.6262
M-estimate				
$\mu = 0$	0.0548	0.0678	0.1096	0.3716
$\sigma = 1$	1.2757	1.2259	0.2907	1.2332
$\lambda = 1$	1.7532	1.5270	0.7356	3.2470
One-step (M)				
$\mu = 0$	0.1215	0.0975	0.0998	0.4870
$\sigma = 1$	1.4286	1.2306	0.4566	1.9269
$\lambda = 1$	1.0097	0.9481	0.4569	1.3712
$\hat{\theta}_0$				
$\mu = 0$	0.0655	0.0627	0.1316	0.4459
$\sigma = 1$	0.7388	0.7292	0.0964	0.8752
$\lambda = 1$	0.5931	0.5947	0.1066	1.3260
One-step ($\hat{\theta}_0$)				
$\mu = 0$	0.0778	0.0927	0.1053	0.4003
$\sigma = 1$	1.0256	1.0282	0.1173	0.3611
$\lambda = 1$	0.9184	0.8944	0.1606	0.5466

components, whereas the One-Step construction based on $\hat{\theta}_0$ improves $\hat{\theta}_0$ and leads to better results. The MoM estimate always leads to poor results under the ideal law, resulting from the small autocorrelation in the data, which leads to large values for the λ-estimate.

For the contamination models the percentage of contamination is set to $p = 0.04$. The cut-off point of the M-estimates is $k = 1$, for the calculation of the optimally robust ICs we consider the radius $r = p\sqrt{n} = 0.4$, and the initial cut-off point $b_0 = 1$.

In a first contamination, we select randomly times and add an additive outlier being distributed according to $\mathcal{N}(3, 1)$, that is we consider for $i = 1, \ldots, n$ the model

$$y_i = x_i + w_i \cdot z_i,$$

where $W_i \sim \mathcal{B}(1, p)$, a Bernoulli random variable with $p = 0.04$, independent of $Z_i \sim \mathcal{N}(3, 1)$, and x_i, $i = 1, \ldots, n$, follows an ideal Ornstein–Uhlenbeck process.

Table 3 Different estimators for the transitions being contaminated with $H(dx_j|x_{<j}) = x_{j-1} + 5$

	Mean	Median	SD	RMSE
MoM				
$\mu = 0$	0.1289	0.1021	0.1287	0.5615
$\sigma = 1$	3.8810	3.8659	0.5967	9.2847
$\lambda = 1$	6.3045	6.0899	1.3021	17.2233
MLE				
$\mu = 0$	0.1292	0.1034	0.1277	0.5601
$\sigma = 1$	2.1761	2.0589	0.4810	3.9894
$\lambda = 1$	1.9825	1.9567	0.4974	3.4467
M-estimate				
$\mu = 0$	−0.0012	−0.0095	0.1516	0.4548
$\sigma = 1$	1.3105	1.3461	0.2213	1.1851
$\lambda = 1$	1.6970	1.6515	0.4472	2.5803
One-step (M)				
$\mu = 0$	0.0193	−0.0328	0.1785	0.5391
$\sigma = 1$	1.1271	1.0989	0.1290	0.5581
$\lambda = 1$	0.3652	0.6647	0.9162	3.4034
$\hat{\theta}_0$				
$\mu = 0$	0.0341	0.0155	0.1827	0.5587
$\sigma = 1$	0.8187	0.7808	0.1103	0.6619
$\lambda = 1$	0.6757	0.6555	0.0939	1.0635
One-step ($\hat{\theta}_0$)				
$\mu = 0$	−0.0026	−0.0384	0.1771	0.5313
$\sigma = 1$	1.0513	1.0430	0.1185	0.3908
$\lambda = 1$	0.9409	0.9085	0.1748	0.5566

Table 2 shows the different estimators as well as their mean, median, standard deviation, and RMSE. The MLE behaves very poorly, the M-estimate improves the MLE, but especially the estimation for the λ-component is not satisfying.

The One-Step estimators behave well, in particular they improve the initial estimates ($\hat{\theta}_0$ and M-estimates) used for their calculation. The estimator $\hat{\theta}_0$ performs however better as initial estimate in the One-Step construction than the M-estimator. Even if the estimator for the λ-component is not so close to the exact value, One-Step ($\hat{\theta}_0$) has less standard deviation and a smaller RMSE than One-Step(M).

In a second contamination study, we consider $H(dx_j|x_{<j}) = x_{j-1} + 5$ as Markov kernel in the convex contamination (24). The estimation results are summarized in Table 3: MLE and M-estimate give very poor results, as well as the estimate for the λ-component of One-Step(M-estimate). However, the One-Step estimator based on $\hat{\theta}_0$, One-Step ($\hat{\theta}_0$), leads again to very satisfying results, therefore we may conclude that $\hat{\theta}_0$ is preferable to the M-estimator as initial estimator in the One-Step construction.

6.2 Open problems

The following theoretical problems remain for future research:

(a) The asymptotic linearity of the One-Step estimator from (52).
(b) The limes normality in the submodel of simple perturbations remains to be shown uniformly in q, that is the convergence in (25) for $q \in Z_\infty$ with $q(x_{\le j}) \ge -\varepsilon(x_{<j})\, P_\theta(dx_j|x_{<j})^{(n;j|<j)}$-a.s.
(c) From the counterexample in Staab (1984) we know that the estimator construction does not extend from the submodel to the full neighborhood system in the dependent case. However, once the uniform asymptotic normality is shown in the submodel, see (b), one could try to show this property in a larger submodel. That is, to find a submodel which is large enough such that the uniform asymptotic normality extends. Under so-called φ-mixing conditions, this was attempted in Ruckdeschel (2002).

7 Conclusion

We have derived elementary M- and optimally robust estimators for the parameters of a discretely observed Ornstein–Uhlenbeck process. The M-estimator has been obtained by adopting Huber's ψ and χ functions from Huber (1981), whereas the ψ- function of the second approach has been determined by an optimality criterion (minimizing the maxasyMSE on shrinking neighborhoods).

Comparing the ICs of the two approaches graphically showed that the more general robust approach is clearly preferable to the first, rather intuitive approach; we obtained for all parameter components bounded ICs.

Concerning the parameter estimation, we have carried out a One-Step construction based on the optimal IC and compared this One-Step estimator with MoM, MLE, and M-estimator for different contamination models. The One-Step estimates improved the initial estimates ($\hat{\theta}_0$ from (53) and M-estimates) used for their construction. Moreover, the One-Step estimate based on $\hat{\theta}_0$ led to very satisfying results under our contamination models.

We summarized remaining problems for future research in Sect. 6.2; a real data application seems worthwhile to study as well.

Acknowledgments Prof. Stephan Morgenthaler suggested the problem of robust parameter estimation for the Ornstein–Uhlenbeck process to me. I thank him and also Prof. Helmut Rieder for generous guidance and encouragement. I also thank the referees for their careful reading of the manuscript and their valuable comments.

References

Bednarski T (2010) Fréchet differentiability in statistical inference for time series. Stat Methods Appl 19:517–528
Bickel P (1984) Robust regression based on infinitesimal neighbourhoods. Ann Stat 12:1349–1368
Breiman L (1992) Probability. Society for Industrial and Applied Mathematics (SIAM), Philadelphia

Bustos O (1982) General M-estimates for contaminated pth-order autoregressice processes: consistency and asymptotic normality. Z Wahrsch verw Gebiete 59:491–504
Croux C, Dehon C (2010) Influence functions of the Spearman and Kendall correlation measures. Stat Methods Appl 19:497–515
Gobet E (2002) LAN property for ergodic diffusions with discrete observations. Annales de l'IHB (B) Probablités et statistiques 38(5):711–737
Hájek J (1972) Local asymptotic minimax and admissibility in estimation. In: Proceedings of 6th Berkeley symposium mathematics statistics probability, vol 1. University of California, 1970, pp 175–194
Hallin M, Koell C, Werker B (2000) Optimal inference for discretely observed semiparametric Ornstein–Uhlenbeck processes. J Stat Plan Inference 91:323–340
Hampel F, Ronchetti E, Rousseeuw P, Stahel W (1986) Robust statistics. The approach based on influence functions. Wiley, New York
Huber P (1981) Robust statistics. Wiley, New York
Iacus S (2008) Simulation and inference for stochastic differential equations. Springer, New York
Jeganathan P (1982) On the asymptotic theory of estimation when the limit of the loglikelihood ratios is mixed normal. Sankhyā Ser A 44:173–212
Kohl M, Ruckdeschel P, Rieder H (2010) Infinitesimally robust estimation in general smoothly parametrized models. Stat Methods Appl 19:333–354
LeCam L (1986) Asymptotic methods in statistical decision theory. Springer, NewYork
Mancini L, Ronchetti E, Trojani F (2005) Optimal conditionally unbiased bounded-influence inference in dynamic location and scale models. J Am Stat Assoc 100:628–641
McLeish D (1974) Dependent central limit theorems and invariance principles. Ann Probab 2:620–628
Muler N, Yohai V (2002) Robust estimates for arch processes. J Time Ser Anal 23:79–109
Muler N, Yohai V (2008) Robust estimates for garch models. J Stat Plan Inference 138:2918–2940
Øksendal B (2002) Stochastic differential equations. Springer, New York
Ornstein L, Uhlenbeck G (1930) On the theory of brownian motion. Phys Rev 36:823–841
Rao P (1999) Statistical inference for diffusion type processes. Oxford University Press, New York
Rieder H (1994) Robust asymptotic statistics. Springer, New York
Rieder H (2003) Robust estimation for time series models based on infinitesimal neighborhoods. Talk presented at EPF Lausanne. Slides available under http://www.stoch.uni-bayreuth.de/de/pdfFiles/zzk12Jun03.pdf
Ruckdeschel P (2002) Failure of construction of arma-estimators for full neighborhoods (tax.report)
Ruckdeschel P (2006) A motivation for $1/\sqrt{n}$-shrinking neighborhoods. Metrika 63:295–307
Sakamoto Y, Yoshida N (1998) Asymptotic expansion of M-estimator over Wiener space. Stat Inference Stoch Process 1:85–103
Staab M (1984) Robust parameter estimation for arma models. Ph.D. thesis, University of Bayreuth
Swensen A (1985) The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. J Multivar Anal 16:54–70
Vasicek O (1977) An equilibrium characterization of the term structure. J Financial Econ 5:177–188
Yoshida N (1990) Asymptotic behavior of M-estimator and related random field for diffusion process. Ann Inst Stat Math 42:221–251