Statistica Sinica **21** (2011), 255-277

COMMENTS ON PAIRWISE LIKELIHOOD IN TIME SERIES MODELS

Richard A. Davis and Chun Yip Yau

Columbia University and Chinese University of Hong Kong

Abstract: This note is concerned with the asymptotic properties of pairwise likelihood estimation procedures for linear time series models. The latter includes ARMA as well as fractionally integrated ARMA processes, where the fractional integration parameter $d < 0.5$. In some cases, including AR(1) processes and long-memory processes with $d < 0.25$, the loss in efficiency in using pairwise likelihood methods is slight. On the other hand, for some models such as the MA(1), the loss in efficiency can be large, and for long-memory models with $d > 0.25$, the pairwise likelihood estimator is not even asymptotically normal. A comparison between using all pairs and consecutive pairs of observations in defining the likelihood is given. We also explore the application of pairwise likelihood to a popular nonlinear model for time series of counts. In this case, the likelihood based on the entire data set cannot be computed without resorting to simulation-based procedures. On the other hand, it is possible to numerically compute the pairwise likelihood precisely. We illustrate the good performance of pairwise likelihood in this case.

Key words and phrases: ARFIMA model, composite likelihood, linear time series, pairwise likelihood, Poisson autoregressive model.

1. Introduction

Although the likelihood principle plays an important role in the formal theory of statistical inference, it is often not feasible to use in applications such as genetic and spatial data where complex interdependences are present. For instance, the computation of the likelihood may require inversion of a large dimensional covariance matrix, or repeated evaluation of high-dimensional integrals over the distribution of a latent process, which are computationally prohibitive. For the latent-process specified models, one often has to resort to simulation-based methods to approximate the likelihood and the quality of the approximations can be difficult to assess.

To overcome the limitations in computing the exact likelihood, Lindsay (1988) proposed the composite likelihood as a pseudo-likelihood for inference. The pseudo-likelihood may take various forms such as combinations of likelihoods for small subsets of the data or combinations of conditional likelihoods. These procedures adopt some features of the full likelihood which are useful for inference while keeping the computation feasible.

Pairwise likelihood (PL) is one special case of a composite likelihood, in which the pseudo-likelihood is defined as the product of the bivariate likelihood of all possible pairs of observations. That is, for n observations $y_1, y_2, \ldots, y_n$ from a statistical model with parameter η, the pairwise likelihood is defined by $\prod_{i<j}^{n} f_{\eta,i,j}(y_i, y_j)$, where $f_{\eta,i,j}(y_i, y_j)$ is the joint density of the random variables y_i and y_j. The PL can be viewed as the likelihood of an imaginary data set of $n(n-1)/2$ independent samples of bivariate observations. The evaluation of pairwise likelihood requires $n(n-1)/2$ evaluations of a bivariate density function, which is computationally efficient when the bivariate density function can be computed quickly in comparison to the joint density of n observations. A general discussion of pairwise likelihood can be found in Cox and Reid (2004).

Due to the simplicity of pairwise likelihood, it has been applied in many fields in statistics, including image analysis (Nott and Ryden (1999)), longitudinal binary data (Kuk and Nott (2000)), multivariate survival data analysis (Parner (2001)), multilevel models (Renard, Molenberghs and Geys (2004)), frailty models for longitudinal count data (Henderson and Shimakura (2003)), Gaussian spatial data (Hjort and Omre (1994)), binary spatial data (Heagerty and Lele (1998)), spatial generalized linear mixed models (Varin and Vidoni (2009)), state space models (Varin, Høst, and Skare (2005), Joe and Lee (2009)) and vast dimensional time-varying covariance models (Engle, Shephard and Sheppard (2009)). Mardia et al. (2009) and Joe and Lee (2009) considered pairwise likelihood in longitudinal time series setting. That is, there are n independent observations from a p-dimensional multivariate vector, where each entry has a time series structure. The asymptotic setup is that p is fixed while $n \to \infty$. In this paper, we consider a traditional time series setting in which we have a single realization from a univariate time series. That is, we have a single observation from a p-dimensional random vector having a time series structure, and $p \to \infty$.

In this paper we provide some theory behind the application of PL in time series models. As time series observations are ordered in time with the bulk of the dependence occurring in adjacent observations, it may be more appropriate to consider a further simplification of pairwise likelihood based only on consecutive pairs of observations. We shall call this consecutive pairwise likelihood (CPL). We first develop theory about consistency and asymptotic distribution for maximum CPL estimators (MCPLE) in linear time series models. Our setting includes both short-memory models, where the autocorrelation function is summable, and long-memory models, where $\rho(k) \sim \lambda |k|^{2d-1}$ decays parabolically, for some constants, $\lambda > 0$ and $d < 0.5$. It is found that MCPLE is consistent for both short- and long-memory time series. The MCPLE is asymptotically normally distributed with $\sqrt{n}$ convergent rate when the series is short-memory or long-memory with $d \leq 1/4$. When $d > 1/4$, the MCPLE is no longer normally distributed and the convergent rate is slower than the standard $\sqrt{n}$ rate.

Secondly, we compare the performance of MCPLE and MLE using some simple time series models. First, for AR(1) processes, the asymptotic relative efficiency of the MCPLE to the MLE is one for all values of the AR(1) parameter. On the other hand, the asymptotic variance of PL estimates is much larger and in fact becomes infinite as the AR parameter converges to zero. In contrast, in a MA(1) model the asymptotic relative efficiency of the MCPLE to the MLE is disappointingly small. This result may seem surprising since the dependence in an MA(1) model does not extend beyond lag 1. At the other extreme, even for a long-memory ARFIMA(0,d,0) process, CPL performs extremely well when the memory parameter $d \in (0, 0.2)$. However, when $d > 0.25$, the procedure breaks down theoretically in the sense that the asymptotic relative efficiency of MCPLE to MLE becomes infinite. Based on empirical studies, the MCPLE still has decent performance in this extremely long-memory case. While it is tempting to believe that there may be severe information loss using only dependence between pairwise observations, the loss of information can be small even for complicated models. In some cases it is even fully efficient, as seen in the AR(1) example in Section 3; see also Mardia et al. (2009). It is not always clear in which cases the CPL procedure performs well. As the dependence structure in time series models are often explicit, we hope the comparison of CPL with the full likelihood in some standard time series models will shed some light on this issue.

As a comparison to the MCPLE, we study the MPLE, which is defined by the product of the bivariate densities of all distinct pairs of observations. Consistency of the MPLE is established under the condition that the autocovariance function of the time series is absolutely summable. We suspect that the MPLE is inconsistent for ARFIMA models. This is supported by a simulation example in Section 2.

We also demonstrate the inconsistency of MPLE via a simulation example using ARFIMA model where the autocovariance function is not absolutely summable.

Lastly we illustrate the good performance of the pairwise likelihood applied to a nonlinear model for time series of counts. This model, described in Zeger (1988) (see also Davis and Rodriguez-Yam (2005)) assumes that conditional on an AR(1) latent process, the observations are independent Poisson random variables. In this case, exact likelihood requires an n-fold integration over the joint distribution of the latent process, which is not feasible to compute. On the other hand, the CPL, which is computable since it only involves two dimensional integrals, gives performance comparable to the results found in Davis and Rodriguez-Yam (2005), which uses simulation-based methods to approximate the likelihood. The good performance suggests the promise of the CPL estimation technique for more complicated models.

The paper is organized as follows. In Section 2 we show consistency and weak convergence of estimators for CPL in linear time series models. In Section

3 we compare the performance of PL, CPL and the full likelihood, and illustrate the CPL technique with the nonlinear Poisson model.

2. Pairwise Likelihood in Linear Time Series Models

In this section we consider the linear time series model

$$X_t = \mu + \sum_{j=0}^{\infty} \psi_j Z_{t-j}\,, \quad t = 1, 2 \ldots, \tag{2.1}$$

where for $j \geq 0$, $\psi_j = \psi_j(\theta)$ satisfies $\psi_0 = 1$ and $\sum_{t=0}^{\infty} \psi_t^2 < \infty$, θ is a parameter belonging to a compact m-dimensional parameter space Θ, and $\{Z_t\}$ is IID$(0, \sigma^2)$ with finite fourth cumulant κ. Letting $\eta = (\sigma^2, \theta)$ denote the parameter vector, the auto-covariance function $E((X_{t+k}-\mu)(X_t-\mu)) = \sigma^2 \sum_{t=0}^{\infty} \psi_t \psi_{t+k} = \sigma^2 \gamma_\theta(k)$, where $\gamma_\theta(k) = \sum_{t=0}^{\infty} \psi_t \psi_{t+k}$. Further assume that $\gamma_\theta(k)$ is twice continuously differentiable for all k. This setting includes both short-memory models, such as ARMA models where $\gamma_\theta(k)$ is summable, and long-memory models where $\gamma_\theta(k) \sim \lambda |k|^{2d-1}$ decays parabolically, for some constants $\lambda > 0$, and $d < 0.5$. Without loss of generality we assume $\mu = 0$ in the discussion below.

Estimation using the pairwise likelihood involves computing the joint likelihood (Gaussian likelihood in this case) for all pairs of observations. The bivariate density for any pair of observations $\{X_t, X_{t+k}\}$, $k \neq 0$, is

$$f_\eta(X_t, X_{t+k}) = \frac{1}{2\pi\sigma^2\sqrt{\Delta_\theta(k)}} \exp\left(-\frac{(X_t^2 + X_{t+k}^2)\gamma_\theta(0) - 2X_t X_{t+k}\gamma_\theta(k)}{2\sigma^2 \Delta_\theta(k)} \right),$$

where $\sigma^4 \Delta_\theta(k) = \sigma^4(\gamma_\theta^2(0) - \gamma_\theta^2(k))$ is the determinant of the covariance matrix of $\{X_t, X_{t+k}\}$. Note that we are not assuming that $\{X_t, X_{t+1}\}$ has a bivariate Gaussian density, rather, this is used only as the *objective function* for the estimation procedure.

Given a vector of observation $\mathbf{X}_n = (X_1, \ldots, X_n)$, the pairwise likelihood is defined to be the product of the bivariate densities of each distinct pair of observations whereas the consecutive likelihood is the product of the bivariate densities of consecutive pairs of observations at various lags. Specifically, we have

The Pairwise log-Likelihood (PL):

$$\begin{aligned} PL(\eta; \mathbf{X}_n) &= \sum_{j=1}^{n-1} \sum_{t=1}^{n-j} \log f_\eta(X_t, X_{t+j}) \\ &= -\frac{1}{2} \sum_{j=1}^{n-1} \sum_{t=1}^{n-j} \left[\frac{(X_t^2 + X_{t+j}^2)\gamma_\theta(0) - 2X_t X_{t+j}\gamma_\theta(j)}{\sigma^2 \Delta_\theta(j)} + \log\left(\sigma^4 \Delta_\theta(j)\right) \right]. \end{aligned} \tag{2.2}$$

The kth order Consecutive Pairwise log-Likelihood (CPL_k):

$$CPL_k(\eta; \mathbf{X}_n) = \sum_{j=1}^{k} \sum_{t=1}^{n-j} \log f_\eta(X_t, X_{t+j})$$
$$= -\frac{1}{2} \sum_{j=1}^{k} \sum_{t=1}^{n-j} \left[\frac{(X_t^2 + X_{t+j}^2)\gamma_\theta(0) - 2X_t X_{t+j}\gamma_\theta(j)}{\sigma^2 \Delta_\theta(j)} + \log\left(\sigma^4 \Delta_\theta(j)\right) \right] . \quad (2.3)$$

In many time series models, most of the dependence occurs in adjacent observations while the dependence diminishes as the time lag between observations increases. Therefore the use of pairwise likelihood may lose efficiency since too many redundant pairs of observations can skew the information confined in pairs of adjacent observations. We will focus on consecutive likelihood. Examples in Section 3 show that CPL is superior to PL. On the other hand, there is an identifiability issue relative to CPL. By (2.3), CPL_k depends on the model only through the autocovariances $\sigma^2\gamma_\theta(0)$, $\sigma^2\gamma_\theta(1), \ldots, \sigma^2\gamma_\theta(k)$. The model is not identifiable through these autocovariances for more complicated models such as an $AR(k+1)$ model. Generally, when θ is m-dimensional, one takes $k \geq m$ to ensure identifiability of the parameters through the autocovariances . If $k = m$, it can be shown that one recovers the methods of moment of estimator for θ. In other words, one solves for θ by matching the theoretical and sample autocorrelation functions up to lag m. When $k > m$, CPL is in general different from the method of moments. Note that $CPL_{n-1}(\eta; \mathbf{X}_n) = PL(\eta; \mathbf{X}_n)$. Thus the kth order CPL bridges CPL_1 and PL.

Another way to resolve identification issues is to work with the consecutive k-tuple log-likelihood

$$CTL_k(\eta; \mathbf{X}_n) = \sum_{t=1}^{n-k} \log f_\eta(X_t, X_{t+1}, \ldots, X_{t+k}) . \quad (2.4)$$

Since CTL_k involves joint densities with higher dimensions, it may be more computationally expensive to use. As an intermediate step, one can view CPL_k as a further simplification of CTL_k. While we will focus on CPL_k in this paper, the theory of CTL_k follows in a similar way as that of CPL_k. Caragea and Smith (2007) considered the small blocks method which is similar to CTL_K, where a k-tuple of observations $(X_t, X_{t+1}, \ldots, X_{t+k})$ is regarded as a block. But the small blocks method considers only separated blocks, while CTL_K can be viewed as a moving block method.

As every $\log f_\eta(X_t, X_{t+j})$ is the exact log-likelihood of the imaginary sample of two observations $\{X_t, X_{t+j}\}$, CPL naturally shares similar important properties as the exact log-likelihood, such as that the expectation of the derivative of

the log-likelihood function is zero and that the maximum of the expected value of the log-likelihood function is attained at the true parameter. These properties lead to the consistency of maximum consecutive likelihood estimation, given in the following theorem.

Theorem 2.1. *Suppose* $\{X_t\}$ *is the linear process specified in* (2.1) *with* $\mu = 0$ *and parameter* $\eta_o = (\sigma_o^2, \theta_o)$. *Let*

$$\hat{\eta}_n = \arg\max_{\eta} CPL_k(\eta; \mathbf{X}_n)$$

be the maximum consecutive likelihood estimator (*MCPLE*). *If the identifiability condition*

$$\sigma_1^2 \gamma_{\theta_1}(j) = \sigma_2^2 \gamma_{\theta_2}(j) \quad \text{for} \quad j = 0, 1, \ldots, k \qquad \text{iff} \qquad (\sigma_1^2, \theta_1) = (\sigma_2^2, \theta_2) \tag{2.5}$$

is satisfied, then $\hat{\eta}_n \xrightarrow{a.s.} \eta_o$.

Proof of Theorem 2.1. We only prove the result for the $k = 1$ case, the other cases being similar. Let $E_\eta(\cdot)$ be the expectation evaluated under the P_η, the probability measure induced by $\{X_t\}$, which follows the model (2.1) with parameter $\eta = (\sigma^2, \theta)$. For true parameter η_o, by the Ergodic Theorem, we have

$$\sum_{t=1}^{n-1} \frac{X_t^2 + X_{t+1}^2}{n} \xrightarrow{a.s.} 2\mathrm{E}_{\eta_o}(X_1^2) = 2\sigma_o^2 \gamma_{\theta_o}(0)$$

and

$$\sum_{t=1}^{n-1} \frac{X_t X_{t+1}}{n} \xrightarrow{a.s.} \mathrm{E}_{\eta_o}(X_1 X_2) = \sigma_o^2 \gamma_{\theta_o}(1)\,. \tag{2.6}$$

Thus we have

$$\begin{aligned} \frac{1}{n} CPL(\eta; \mathbf{X}_n) \;&=\; \frac{1}{n} \sum_{t=1}^{n-1} \log f_\eta(X_t, X_{t+1}) \\ &\xrightarrow{a.s.} \mathrm{E}_{\eta_o}(\log f_\eta(X_1, X_2))\,. \end{aligned}$$

We claim that the maximum of $\mathrm{E}_{\eta_o}(\log f_\eta(X_1, X_2))$ over η is attained uniquely at $\eta = \eta_o$. For Gaussian distributed $\{X_t\}$, Jensen's Inequality implies that

$$E_{\eta_o}\left(\log \frac{f_{\eta_1}(X_t, X_{t+1})}{f_{\eta_o}(X_t, X_{t+1})}\right) \le \log E_{\eta_o}\left(\frac{f_{\eta_1}(X_t, X_{t+1})}{f_{\eta_o}(X_t, X_{t+1})}\right) = \log(1) = 0\,.$$

It follows that

$$\mathrm{E}_{\eta_o}(\log f_{\eta_o}(X_t, X_{t+1})) \ge \mathrm{E}_{\eta_o}(\log f_{\eta_1}(X_t, X_{t+1})) \tag{2.7}$$

for any η_1, and the equality holds if and only if $f_{\eta_o}(X_t, X_{t+1}) = f_{\eta_1}(X_t, X_{t+1})$ almost surely, which holds if and only if (2.5) is satisfied. From (2.3), it can be seen that the expectation $E_{\eta_o}\big(\log(f_{\eta_1}(X_t, X_{t+1})/f_{\eta_o}(X_t, X_{t+1}))\big)$ depends only on the autocovariance function of the model regardless of the distribution assumption, and hence (2.7) in fact holds for any distribution of $\{X_t\}$ with finite second moments. This proves the claim.

To make use of the compactness property of Θ we profile out σ^2. For a fixed θ, $CPL(\eta; \mathbf{X}_n) \equiv CPL((\sigma^2, \theta); \mathbf{X}_n)$ can be shown to be maximized by

$$\hat{\sigma}_n^2(\theta) = \frac{\gamma_\theta(0)}{\Delta_\theta(1)} \sum_{t=1}^{n-1} \frac{X_t^2 + X_{t+1}^2}{2n} - \frac{2\gamma_\theta(1)}{\Delta_\theta(1)} \sum_{t=1}^{n-1} \frac{X_t X_{t+1}}{2n}\,. \tag{2.8}$$

Thus maximizing $CPL(\eta; \mathbf{X}_n)$ over η is equivalent to maximizing $CPL((\hat{\sigma}_n^2(\theta), \theta); \mathbf{X}_n)$ over θ. Let $\hat{\theta}_n$ be this maximizer and B^c be a null set of the probability space Ω such that (2.6) holds for all $\omega \in B$. We show that for each $\omega \in B$, $\hat{\theta}_n \to \theta_o$. For any $\omega \in B$, suppose on the contrary that $\hat{\theta}_n \nrightarrow \theta_o$. By the compactness of Θ there exists a subsequence $\{n_k\}$ such that $\hat{\theta}_{n_k} \to \theta^*$ for some $\theta^* \neq \theta_o$. Then

$$\hat{\sigma}_{n_k}^2(\hat{\theta}_{n_k}) \to \frac{\gamma_{\theta^*}(0)}{\Delta_{\theta^*}(1)} \sigma_o^2 \gamma_{\theta_o}(0) - \frac{\gamma_{\theta^*}(1)}{\Delta_{\theta^*}(1)} \sigma_o^2 \gamma_{\theta_o}(1) =: \sigma_*^2\,.$$

Now

$$\begin{aligned}
&\lim_{k\to\infty} \frac{1}{n_k} CPL((\hat{\sigma}_{n_k}^2(\hat{\theta}_{n_k}), \hat{\theta}_{n_k}); \mathbf{X}_{n_k}) \\
&\quad \geq \lim_{k\to\infty} \frac{1}{n_k} CPL((\sigma_o^2, \theta_o); \mathbf{X}_{n_k}) = \mathrm{E}_{\eta_o}(\log f_{\eta_o}(X_1, X_2)) \\
&\quad > \mathrm{E}_{\eta_o}(\log f_{(\theta^*, \sigma_*^2)}(X_1, X_2)) = \lim_{k\to\infty} \frac{1}{n_k} CPL((\sigma_*^2, \theta^*); \mathbf{X}_{n_k}) \\
&\quad = \lim_{k\to\infty} \frac{1}{n_k} CPL((\sigma_{n_k}^2(\hat{\theta}_{n_k}), \hat{\theta}_{n_k}); \mathbf{X}_{n_k})\,,
\end{aligned}$$

which is a contradiction. Hence $\hat{\theta}_n \to \theta_o$ for each $\omega \in B$. It then follows from (2.8) that $\hat{\sigma}_n^2(\hat{\theta}_n) \to \sigma_o^2$ for each $\omega \in B$, which establishes the strong consistency.

Once the consistency of the MCPLE has been established, the asymptotic distribution can be derived using a Taylor series expansion of the pseudo-likelihood around the true value. It turns out that asymptotical normality with rate $\sqrt{n}$ only holds in the short-memory and long-memory cases with $d < 0.25$. This is the content of the following theorem.

Theorem 2.2. *For $k \geq 1$, let*

$$a_k(\eta) = \frac{\gamma_\theta(0)}{\sigma^2 \Delta_\theta(k)}\,, \quad b_k(\eta) = -\frac{\gamma_\theta(k)}{\sigma^2 \Delta_\theta(k)}\,, \quad c_k(\eta) = \log \sigma^4 \Delta_\theta(k)\,.$$

Denote the k-dimensional column vector with one in every entry by $\mathbf{1}$, the derivative with respect to η by $'$ and matrix transpose by ^{T}. Set

$$H(\eta) = 2\sigma^2\gamma_{\theta_o}(0)\sum_{j=1}^{k} a_j''(\eta) + 2\sigma^2\sum_{j=1}^{k} b_j''(\eta)\gamma_{\theta_o}(j) + \sum_{j=1}^{k} c_j''(\eta)\,,$$

$$\tau_{i,j} = \sigma^4\left(\sum_{k=-\infty}^{\infty}\gamma_{\theta_o}(k)\gamma_{\theta_o}(k+i-j) + \sum_{k=-\infty}^{\infty}\gamma_{\theta_o}(k)\gamma_{\theta_o}(k+i+j) + \kappa\gamma_{\theta_o}(i)\gamma_{\theta_o}(j)\right),$$

$$\Sigma_1 = \tau_{0,0}\mathbf{1}\mathbf{1}^T\,,\quad \Sigma_2 = \mathbf{1}(\tau_{0,1}\ \tau_{0,2}\ \ldots\ \tau_{0,k})\,,\quad \Sigma_3 = (\tau_{i,j})_{i,j=1,\ldots,k}\,,$$

$$M = \begin{pmatrix}\Sigma_1 & \Sigma_2\\ \Sigma_2^T & \Sigma_3\end{pmatrix},$$

$$V = (a_1'(\eta_o)\ \ldots\ a_k'(\eta_o)\ \ b_1'(\eta_o)\ \ldots\ b_k'(\eta_o))\,,$$

where $H(\eta_o)$ is assumed to be invertible in a neighborhood of η_o. If (2.5) holds and $\{X_t\}$ is

- *short-memory or long-memory with $d < 1/4$, then*

$$\sqrt{n}(\hat{\eta}_n - \eta_o) \to N\left(0, 4H(\eta_o)^{-1}VMV^TH^T(\eta_o)^{-1}\right).$$

- *long-memory with $d = 1/4$, then*

$$\sqrt{n/\log n}(\hat{\eta}_n - \eta_o) \to N\left(0, 16\lambda^2\sigma_o^2H(\eta_o)^{-1}V\mathbf{1}\mathbf{1}^TV^TH^T(\eta_o)^{-1}\right).$$

- *long-memory with $d > 1/4$, then $n^{1-2d}(\hat{\eta}_n - \eta_o)$ converges to a non-Gaussian centered random variable with variance $[16\lambda^2\sigma_o^2/(-1+4d)]H(\eta_o)^{-1}V\mathbf{1}\mathbf{1}^TV^T H^T(\eta_o)^{-1}$.*

Proof of Theorem 2.2. Define

$$\begin{aligned}H_n(\eta) &= \frac{1}{n}CPL_k''(\eta;\mathbf{X}_n)\\ &= \frac{1}{n}\sum_{j=1}^{k}\sum_{t=1}^{n-j}\left[(X_t^2 + X_{t+j}^2)a_j''(\eta) + 2X_tX_{t+j}b_j''(\eta) + c_j''(\eta)\right]\end{aligned}$$

and

$$\begin{aligned}J_n(\eta) &= \frac{1}{\sqrt{n}}CPL_k'(\eta;\mathbf{X}_n)\\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{k}\sum_{t=1}^{n-j}\left[(X_t^2 + X_{t+j}^2)a_j'(\eta) + 2X_tX_{t+j}b_j'(\eta) + c_j'(\eta)\right]. \qquad (2.9)\end{aligned}$$

A Taylor series expansion of $J_n(\eta)$ and Theorem 2.1 give

$$J_n(\eta_o) = H_n(\eta_n^+)\sqrt{n}(\eta_o - \hat{\eta}_n)\,,$$

where η_n^+ is between η_o and $\hat{\eta}_n$. To derive the asymptotic distribution of $\hat{\eta}_n$, it suffices to look at the asymptotic properties of $H_n(\eta_n^+)$ and $J_n(\eta_o)$. By the Ergodic Theorem and the fact that $\eta_n^+ \stackrel{a.s.}{\to} \eta_o$, we have

$$H_n(\eta_n^+) \stackrel{a.s.}{\to} 2\sigma_o^2\gamma_{\theta_o}(0)\sum_{j=1}^{k} a_j''(\eta_o) + 2\sigma_o^2\sum_{j=1}^{k} b_j''(\eta_o)\gamma_{\theta_o}(j) + \sum_{j=1}^{k} c_j''(\eta_o) = H(\eta_o)\,. \tag{2.10}$$

Next we compute the asymptotic mean and variance of $J_n(\eta_o)$. Note that $J_n(\eta_o)$ has expectation zero as it is a sum of derivatives of log-likelihood functions. For the variance of $J_n(\eta_o)$, it suffices to consider only the non-deterministic terms

$$A_j := \frac{a_j'(\eta_o)}{\sqrt{n}}\sum_{t=1}^{n-j}(X_t^2 + X_{t+j}^2) \quad \text{and} \quad B_j := \frac{2b_j'(\eta_o)}{\sqrt{n}}\sum_{t=1}^{n-j} X_tX_{t+j}\,,$$

$j = 1, \ldots, k$. Note that for $i, j \le k$, we have the approximation

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{t=1}^{n-i}(X_t^2 + X_{t+i}^2), \sum_{t=1}^{n-j}(X_t^2 + X_{t+j}^2)\right)\\
&\quad = 4\sum_{t=1}^{n}\sum_{k=1}^{n}\mathrm{Cov}(X_t^2, X_k^2)\\
&\quad = 4\sum_{t=1}^{n}\sum_{k=1}^{n}\left[\mathrm{Cum}(X_t, X_t, X_{t+k}, X_{t+k}) + 2\mathrm{Cov}(X_t, X_{t+k})^2\right]\\
&\quad = 4\sigma_o^4\sum_{t=1}^{n}\sum_{k=1}^{n}\kappa\sum_{p\ge 1}\psi_p^2\psi_{p+k}^2 + 8\sigma_o^4\sum_{t=1}^{n}\sum_{k=1}^{n}\gamma_{\theta_o}^2(k)\\
&\quad \sim 4n\sigma_o^4\left(2\sum_{k=-n}^{n}\gamma_{\theta_o}^2(k) + \kappa\gamma_{\theta_o}^2(0)\right)\,,
\end{aligned} \tag{2.11}$$

where $x_n \sim y_n$ means $x_n/y_n \to 1$ as $n \to \infty$. Similarly, we have

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{t=1}^{n-i} X_tX_{t+i}, \sum_{t=1}^{n-j} X_tX_{t+j}\right)\\
&\quad \sim n\sigma_o^4\left(\sum_{k=-n}^{n}\gamma_{\theta_o}(k)(\gamma_{\theta_o}(k+i-j) + \gamma_{\theta_o}(k+i+j)) + \kappa\gamma_{\theta_o}(i)\gamma_{\theta_o}(j)\right),
\end{aligned} \tag{2.12}$$

$$\begin{aligned}
&\mathrm{Cov}\left(\sum_{t=1}^{n-i}(X_t^2 + X_{t+i}^2), \sum_{t=1}^{n-j} X_tX_{t+j}\right)\\
&\quad \sim 2n\sigma_o^4\left(2\sum_{k=-n}^{n}\gamma_{\theta_o}(k)\gamma_{\theta_o}(k+j) + \kappa\gamma_{\theta_o}(0)\gamma_{\theta_o}(j)\right).
\end{aligned} \tag{2.13}$$

When X_t is short-memory or long-memory with $d < 1/4$,

$$\sum_{k=-n}^{n} \gamma_{\theta_o}(k+i)\gamma_{\theta_o}(k+j)$$

converges for all $i, j \leq k$ as $n \to \infty$. From (2.11)−(2.13), we have

$$\mathrm{Cov}(A_i, A_j) \to 4\tau_{0,0} a_i' a_j^T,\ \mathrm{Cov}(A_i, B_j) \to 4\tau_{0,j} a_i' b_j^T,\ \mathrm{Cov}(B_i, B_j) \to 4\tau_{i,j} b_i' b_j^T \quad (2.14)$$

which, combined with (2.9) and (2.14), gives

$$\mathrm{Var}(J_n(\eta_o)) = \mathrm{Var}(\sum_{j=1}^{k} A_j + \sum_{j=1}^{k} B_j) \to 4VMV^T. \quad (2.15)$$

For long-memory processes, $\gamma_{\theta_o}(k) \sim \lambda k^{2d-1}$, it can be shown that for any integer j,

$$\lim_{n\to\infty} \frac{\sum_{k=-n}^{n} \gamma_{\theta_o}(k)\gamma_{\theta_o}(k+j)}{\log n} = 2\lambda^2 \quad (2.16)$$

for $d = 1/4$, and

$$\lim_{n\to\infty} \frac{\sum_{k=-n}^{n} \gamma_{\theta_o}(k)\gamma_{\theta_o}(k+j)}{n^{-1+4d}} = \frac{2\lambda^2}{-1+4d} \quad (2.17)$$

for $d > 1/4$. Putting (2.9), (2.11)−(2.13) and (2.16)−(2.17) together, we have

$$\frac{1}{\log n}\mathrm{Var}(J_n(\eta_o)) \to 16\lambda^2\sigma_o^2 V 11^T V^T, \quad \text{if } d = \frac{1}{4}, \quad (2.18)$$

and

$$\frac{1}{n^{-1+4d}}\mathrm{Var}(J_n(\eta_o)) \to \frac{16\lambda^2\sigma_o^2}{-1+4d} V 11^T V^T, \quad \text{if } d > \frac{1}{4}. \quad (2.19)$$

Next we consider the limiting distribution of the term $J_n(\eta_o)$. Note that $J_n(\eta_o)$ is a linear combination of the A_j's and B_j's, which involve the sample autocovariance functions. If $\{X_t\}$ is short memory, it is well known that (e.g., Brockwell and Davis (1991, Thm. 7.2.1) the sample autocovariances are asymptotically multivariate normal. When $\{X_t\}$ is long-memory, the distributional properties of the sample autocovariances have been studied by Hosking (1996). For the case $d \leq 1/4$, the sample autocovariances are asymptotically normal. Therefore, for the short-memory or long-memory cases with $d \leq 1/4$, $J_n(\eta_o)$ is asymptotically normal with mean zero and variance given in (2.15) and (2.18), respectively. If $d > 1/4$, the sample autocovariances converge to a non-Gaussian distribution related to the Rosenblatt process (Rosenblatt (1961) and $J_n(\eta_o)$ is

not asymptotically normal. Since $\sqrt{n}(\hat{\eta}_n - \eta_o) = -H_n(\eta_n^+)^{-1}J_n(\eta_o)$, the result now follows from (2.10), (2.15) and (2.18)−(2.19).

In the following we discuss some aspects of the pairwise likelihood. Since the dependence between two observations decreases with the time lag, the bivariate density $f_\eta(X_t, X_{t+k})$ is close to the product of marginals $f_\eta(X_t)f_\eta(X_{t+k})$, when k is large. Therefore, PL is dominated by the marginal densities when n is large. Indeed, it can be shown that when normalized by n^2, the PL converges to the expected value of the marginal log-likelihoods, $E(\log f_\eta(X_t))$, where $f_\eta(X_t) = 1/\sqrt{2\pi\sigma^2\gamma_\theta(0)}\exp(-X_t^2/2\sigma^2\gamma_\theta(0))$ is the marginal density of X_t. As dependence parameters are not identifiable from the marginal distribution in most time series models, the proof of consistency of CPL does not extend to the case of PL. Nevertheless, when the autocovariance function $\gamma_\theta(\cdot)$ is absolutely summable, the PL produces consistent estimators. This is the content of Theorems 2.3 and 2.4.

Theorem 2.3. *For each $\eta \in R^+ \times \Theta$,*

$$\lim_{n\to\infty}\left|\frac{1}{n^2}PL(\eta;\mathbf{X}_n) - \mathrm{E}(\log f_\eta(X_1))\right| = 0 \quad \textit{almost surely}\,.$$

Proof of Theorem 2.3. For any $\eta \in R \times \Theta$,

$$\begin{aligned}\frac{1}{n^2}PL(\eta;\mathbf{X}_n) &= \frac{1}{n^2}\sum_{k=1}^{n-1}\sum_{t=1}^{n-k}\log f_\eta(X_t, X_{t+k})\\ &= \frac{1}{2n^2}\sum_{t=1}^{n}\sum_{u\neq t}^{n}(\log f_\eta(X_t, X_u) - \log f_\eta(X_t)f_\eta(X_u))\\ &\quad + \frac{1}{2n^2}\sum_{t=1}^{n}\sum_{u\neq t}^{n}\log f_\eta(X_t)f_\eta(X_u)\\ &= A_{n,\eta} + B_{n,\eta}\,, \text{say}\,.\end{aligned}$$

Note that for any fixed η, $B_{n,\eta} = [(n-1)/n^2]\sum_{t=1}^{n}\log f_\eta(X_t) \overset{a.s.}{\to} \mathrm{E}(\log f_\eta(X_1))$ as $n \to \infty$ by the Ergodic Theorem. The proof is completed if $A_{n,\eta} \overset{a.s.}{\to} 0$. To show this, note that

$$\begin{aligned}&\log f_\eta(X_t, X_u) - \log f_\eta(X_t)f_\eta(X_u)\\ &= \frac{(X_t^2 + X_u^2)\gamma_\theta^2(|t-u|) - 2X_tX_u\gamma_\theta(|t-u|)\gamma_\theta(0)}{2\sigma^2\gamma_\theta(0)\Delta_\theta(|t-u|)} + \log\left(1 - \frac{\gamma_\theta(|t-u|)^2}{\gamma_\theta^2(0)}\right)\,.\end{aligned}$$

Thus, for C being some generic constant, $|A_{n,\eta}|$ is bounded by

$$\begin{aligned}
&\frac{C}{n^2}\sum_{t}^{n}\sum_{u\neq t}^{n}\frac{(X_t^2+X_u^2)\gamma_\theta^2(|u-t|)}{2\sigma^2\gamma_\theta(0)\Delta_\theta(|u-t|)}+\frac{C}{n^2}\sum_{t}^{n}\sum_{u\neq t}^{n}\frac{(X_t^2+X_u^2)|\gamma_\theta(|u-t|)|}{2\sigma^2\Delta_\theta(|u-t|)}\\
&\quad+\frac{C\sum_{t=1}^{n}\sum_{u\neq t}^{n}\gamma_\theta^2(|u-t|)}{n^2\gamma_\theta^2(0)}\\
&\leq\frac{C}{\gamma_\theta(0)}\frac{\sum_{t=1}^{n}X_t^2}{n}\frac{1}{n}\sum_{k=0}^{n}\frac{\gamma_\theta^2(k)}{\Delta_\theta(k)}+\frac{C\sum_{t=1}^{n}X_t^2}{n}\frac{1}{n}\sum_{k=0}^{n}\frac{|\gamma_\theta(k)|}{\Delta_\theta(k)}\\
&\quad+\frac{C}{n\gamma_\theta^2(0)}\sum_{k=1}^{n}\gamma_\theta^2(k)\,,
\end{aligned}\tag{2.20}$$

where we have used the fact that $2|X_tX_u|\leq X_t^2+X_t^2$ and $\log(1-x)<x$ for positive x. Since $\gamma_\theta(k)\to 0$ as $k\to\infty$ in both the short and long-memory cases, it follows by Cesaro summability that $\sum_{k=0}^{n}\gamma_\theta(k)=o(n)$ and $\sum_{k=0}^{n}\gamma_\theta^2(k)=o(n)$. Hence $A_{n,\eta}$ indeed coverges to 0 almost surely.

Theorem 2.4. *Suppose* $\{X_t\}$ *satisfies* (2.1) *with parameter* $\eta_o=(\sigma_o^2,\theta_o)$. *For* $\mathbf{X}_n=(X_1,\ldots,X_n)$, *let*

$$\tilde{\eta}_n=\arg\max_{\eta}PL(\eta;\mathbf{X}_n)$$

be the maximum pairwise likelihood estimator (*MPLE*). *If* $\sum_{k=0}^{\infty}|\gamma_\theta(k)|<\infty$, *then* $\tilde{\eta}_n\overset{a.s.}{\to}\eta_o$.

Proof of Theorem 2.4. Similar to the proof of Theorem 2.1, for any fixed θ we can find a profile likelihood estimator $\tilde{\sigma}_n^2(\theta)$, where maximizing $PL(\eta;\mathbf{X}_n)$ over η is equivalent to maximizing $PL((\tilde{\sigma}_n^2(\theta),\theta);\mathbf{X}_n)$ over θ. Let B^c be a null set of the probability space Ω such that $\sum_{t=1}^{n}X_t^2/n\overset{a.s.}{\to}E_{\eta_o}(X_1^2)$ and, for all integer k, $\sum_{t=1}^{n}X_tX_{t+k}^2/n\overset{a.s.}{\to}E_{\eta_o}(X_1X_{1+k})$ hold for all $\omega\in B$. We show that for each $\omega\in B$, $\tilde{\theta}_n\to\theta_o$. For any $\omega\in B$, suppose on the contrary that $\tilde{\theta}_n\nrightarrow\theta_o$. By the compactness of Θ, there exists a subsequence $\{n_k\}$ such that $\tilde{\theta}_{n_k}\to\theta^*$ for some $\theta^*\neq\theta_o$. Let $\sigma_*^2=\lim_{k\to\infty}\tilde{\sigma}_{n_k}^2(\theta^*)$, $\eta^*=(\sigma_*^2,\theta^*)$, and $\check{\gamma}_\eta(k)=\sigma^2\gamma_\theta(k)$. We first show that

$$\frac{1}{n}PL(\eta_o;\mathbf{X}_n)-\frac{1}{n}PL(\eta^*;\mathbf{X}_n)>0\tag{2.21}$$

almost surely, for all sufficiently large n. We only consider the case where $f_\eta(X_t)$ is unidentifiable in the sense of (2.5), i.e., $\check{\gamma}_{\eta_o}(0)=\check{\gamma}_{\eta^*}(0)$. Otherwise the proof is obvious by Theorem 2.3 and the fact that $E_{\eta_o}(f_{\eta_o}(X_1))>E_{\eta_o}(f_{\eta^*}(X_1))$ when

$\check{\gamma}_{\eta_o}(0) \neq \check{\gamma}_{\eta^*}(0)$. Note that, for any integer $q < n-1$,

$$\begin{aligned}
&\frac{1}{n}PL(\eta_o;\mathbf{X}_n) - \frac{1}{n}PL(\eta^*;\mathbf{X}_n) \\
&\quad = \frac{1}{n}\left(CPL_q(\eta_o;\mathbf{X}_n) - CPL_q(\eta^*;\mathbf{X}_n)\right) \\
&\qquad + \frac{1}{n}\sum_{k=q+1}^{n-1}\sum_{t=1}^{n-k}\left(\log f_{\eta_o}(X_t,X_{t+k}) - \log f_{\eta^*}(X_t,X_{t+k})\right) \\
&\quad = C_{n,q,\eta_o,\eta^*} + D_{n,q,\eta_o,\eta^*}\,, \text{say.}
\end{aligned}$$

We show (2.21) by arguing that C_{n,q,η_o,η^*} is positive and D_{n,q,η_o,η^*} is of smaller order than C_{n,q,η_o,η^*} for sufficiently large q and n. By the Ergodic Theorem,

$$\begin{aligned}
C_{n,q,\eta_o,\eta^*} &= \sum_{k=1}^{q}\frac{1}{n}\sum_{t=1}^{n-k}\left(\log f_{\eta_o}(X_t,X_{t+k}) - \log f_{\eta^*}(X_t,X_{t+k})\right) \\
&\to \sum_{k=1}^{q}\left(\mathrm{E}_{\eta_o}(\log f_{\eta_o}(X_t,X_{t+1})) - \mathrm{E}_{\eta_o}(\log f_{\eta^*}(X_t,X_{t+1}))\right)\,.
\end{aligned}$$

Using (2.7) and its straightforward generalization to the pairs $\{X_t, X_{t+k}\}$ for $k = 2,\ldots,q$, it can be seen that C_{n,q,η_o,η^*} is strictly positive for sufficiently large n. Next, after some algebra, we have

$$\begin{aligned}
D_{n,q,\eta_o,\eta^*} &= \frac{\check{\gamma}_{\eta_o}(0)}{n}\sum_{k=q+1}^{n-1}\sum_{t=1}^{n-k}\frac{(X_t^2+X_{t+k}^2)(\check{\gamma}^2_{\eta^*}(k) - \check{\gamma}^2_{\eta_o}(k))}{\Delta_{\eta^*}(k)\Delta_{\eta_o}(k)} \\
&\quad - \frac{2}{n}\sum_{k=q+1}^{n-1}\sum_{t=1}^{n-k}\frac{(X_tX_{t+k})(\check{\gamma}_{\eta_o}(k) - \check{\gamma}_{\eta^*}(k))(\check{\gamma}^2_{\eta_o}(0) + \check{\gamma}_{\eta^*}(k)\check{\gamma}_{\eta_o}(k))}{\Delta_{\eta^*}(k)\Delta_{\eta_o}(k)} \\
&\quad + \frac{1}{n}\sum_{k=q+1}^{n-1}(n-k)(\log\Delta_{\eta_o}(k) - \log\Delta_{\eta^*}(k)) \\
&= D_1 + D_2 + D_3\,, \quad \text{say.}
\end{aligned}$$

Similar to the calculations in (2.20), we have

$$\begin{aligned}
|D_1| &\le C\frac{\sum_{t=1}^n X_t^2}{n}\sum_{k=q+1}^{n-1}\frac{|\check{\gamma}^2_{\eta^*}(k) - \check{\gamma}^2_{\eta_o}(k)|}{\Delta_{\eta^*}(k)\Delta_{\eta_o}(k)}\,, \\
|D_2| &\le C\frac{\sum_{t=1}^n X_t^2}{n}\sum_{k=q+1}^{n-1}\frac{|\check{\gamma}_{\eta_o}(k) - \check{\gamma}_{\eta^*}(k)|}{\Delta_{\eta^*}(k)\Delta_{\eta_o}(k)}\,, \\
|D_3| &\le C\sum_{k=q+1}^{n-1}|\check{\gamma}^2_{\eta_o}(k) - \check{\gamma}^2_{\eta^*}(k)|\,.
\end{aligned}$$

Table 1. Sample variances of MCPLE and MPLE for d in the ARFIMA(0,0.2,0) model for various sample size n over 500 replications.

n	100	400	1,600	6,400
MCPLE	0.1630	0.0758	0.0371	0.0187
MPLE	0.3086	0.2539	0.3188	0.3186

For $q \geq 1$ and all sufficiently large n, C_{n,q,η_o,η^*} is positive and thus there exists an $\epsilon > 0$ such that $C_{n,q,\eta_o,\eta^*} > \epsilon$. By the assumption that $\gamma_\theta(\cdot)$ is summable, for sufficiently large n there exists an integer q^* such that $|D_i| \leq \epsilon/3$ for $i = 1, 2, 3$. Thus $|D_{n,q^*,\eta_o,\eta^*}| \leq \epsilon$. It follows that for sufficiently large n, $C_{n,q^*,\eta_o,\eta^*} + D_{n,q^*,\eta_o,\eta^*} > 0$ and (2.21) follows. In fact, if E is a compact neighborhood of θ^* that does not cover θ_o, then (2.21) holds uniformly on $\theta^* \in E$. Therefore, for a sufficiently large n_k, $\tilde{\theta}_{n_k} \in E$ and

$$\frac{1}{n_k} PL((\tilde{\sigma}^2_{n_k}(\tilde{\theta}_{n_k}), \tilde{\theta}_{n_k}); \mathbf{X}_{n_k}) \geq \frac{1}{n_k} PL((\sigma_o^2, \theta_o); \mathbf{X}_{n_k}) > \frac{1}{n_k} PL((\tilde{\sigma}^2_{n_k}(\tilde{\theta}_{n_k}), \tilde{\theta}_{n_k}); \mathbf{X}_{n_k}),$$

which is a contradiction. Hence $\tilde{\theta}_n \to \theta_o$ for each $\omega \in B$. It can then be shown that $\tilde{\sigma}^2_n(\tilde{\theta}_n) \to \sigma_o^2$, which establishes the strong consistency.

When $\sum_{k=0}^{\infty} |\gamma_\theta(k)| = \infty$, the remainder term D_{n,q,η_o,η^*} is of order larger than $O(1)$, meaning that (2.21) may not hold in general. Thus we do not expect PL to be consistent in this case. Table 1 compares the performance of CPL and PL for the long-memory $ARFIMA(0, d, 0)$ model with $d = 0.2$. While the MCPLE shows a $\sqrt{n}$ rate of consistency, the sample variances of the MPLE stay around the same level as the sample size increases, demonstrating the possibility of inconsistency of PL when $\sum_{k=0}^{\infty} |\gamma_\theta(k)| = \infty$.

3. Examples

In this section we compare the performance of the maximum likelihood estimators (MLE) and the maximum consecutive pairwise likelihood estimators (MCPLE) through three simple time series models: AR(1), MA(1) and ARFIMA(0,d,0).

Example 3.1. Suppose $\{X_t\}$ follows the AR(1) model,

$$X_t = \phi X_{t-1} + Z_t \,,$$

$\phi \in (-1, 1)$, $Z_t \sim$ IID(0,σ^2), and let $\eta = (\sigma^2, \phi)$. The exact log-likelihood

$L(\eta; \mathbf{X}_n)$ and the consecutive pairwise log-likelihood $CPL_1(\eta; \mathbf{X}_n)$ are given by

$$\begin{aligned} L(\eta; \mathbf{X}_n) &= \sum_{t=1}^{n-1} \log f_\eta(X_{t+1}|X_s, s \le t) + \log f_\eta(X_1) \\ &= \sum_{t=1}^{n-1} \log f_\eta(X_{t+1}|X_t) + \log f_\eta(X_1)\,, \\ CPL_1(\eta; \mathbf{X}_n) &= \sum_{t=1}^{n-1} \log f_\eta(X_t, X_{t+1}) = \sum_{t=1}^{n-1} \log f_\eta(X_{t+1}|X_t) + \sum_{t=1}^{n-1} \log f_\eta(X_t)\,. \end{aligned}$$

Note the similarity between the log-likelihood and the CPL_1. Interestingly, MCPLE has a closed form expression

$$\hat{\phi}_n = \hat{\rho}(1) \quad \text{and} \quad \hat{\sigma}^2 = \hat{\gamma}^2(0)(1 - \hat{\phi}_n^2)\,,$$

where $\hat{\gamma}(0) = \sum_{t=1}^{n-1}(X_t^2 + X_{t+1}^2)/2(n-1)$, $\hat{\gamma}(1) = \sum_{t=1}^{n-1} X_t X_{t+1}/(n-1)$, and $\hat{\rho}(1) = \hat{\gamma}(1)/\hat{\gamma}(0)$.

As is well known, (see Brockwell and Davis (1991)), $\hat{\phi}_n = \hat{\rho}(1)$ is the Yule-Walker estimator and is asymptotically efficient for estimating ϕ, so the asymptotic relative efficiency (ARE) of the MCPL_1 to MLE is identically 1. The excellent performance of MCPLE can be explained by the fact that both MLE and MCPLE are asymptotically using the complete and sufficient statistics $\sum_{t=1}^{n} X_t^2$ and $\sum_{t=2}^{n} X_t X_{t-1}$ to obtain an asymptotically unbiased estimator. This argument can be extended to a general AR(p) model, thus it is best to use CPL_p to estimate an $AR(p)$ model. In the AR(1) case, as the CPL_1 is efficient, there is no reason to consider CPL_k, for $k > 1$. It is interesting to note, however, that there is a decrease in efficiency as k increases (see Table 2 and Figure 1a). The ARE of the CPL_k estimator relative to the MLE is smallest for $|\phi|$ around 0.5 and approaches 1 as $|\phi| \to 0$ or 1.

Example 3.2. Let $\{X_t\}$ follow the MA(1) model,

$$X_t = Z_t + \theta Z_{t-1}\,,$$

$\theta \in (-1, 1)$, $Z_t \sim \text{IID}(0, \sigma^2)$. The asymptotic variance of the MLE is $\sigma^2(1 - \theta^2)$ (e.g., Brockwell and Davis (1991)). For this model, it does not make sense to use PL or higher order CPL since X_t and X_{t+k} are independent for $k \ge 2$. Therefore we compare CPL_1 with CTL_2 and CTL_3. Figure 1b shows the asymptotic variance of the four estimators across the parameter $\theta \in (-1, 1)$. Surprisingly, in contrast to Example 3.1, the MCPLE and MCTLE gives relatively higher asymptotic variance than that of the MLE. The asymptotic variance even approaches infinity near $\theta = \pm 1$. Although for the MA(1) model the dependence

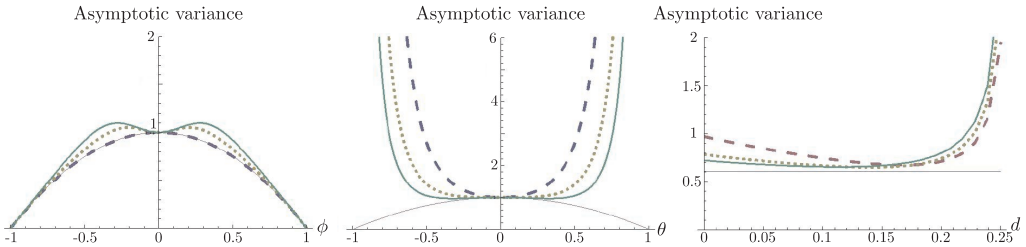

Figure 1. Asymptotic variances of MLE and consecutive likelihood estimators for various models. Fig 1a): ϕ in AR(1) model. Fig 1b): θ in MA(1) model. Fig 1c): d in ARFIMA(0,d,0) model. Thin-solid-line: MLE. Dash-line: MCPLE. For Fig 1a) and 1c), Dotted-line: estimator of CPL_2. Thick-solid-line: estimator of CPL_3. For Fig 1b), Dotted-line: estimator of CTL_2. Thick-solid-line: estimator of CTL_3.

Table 2. Asymptotic relative efficiency between CPL and CTL estimators and MLE for ϕ in AR(1), θ in MA(1) and d in ARFIMA(0,d,0) models.

					AR(1)					
ϕ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CPL_1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
CPL_2	1.000	0.966	0.913	0.889	0.893	0.911	0.933	0.953	0.972	0.987
CPL_3	1.000	0.963	0.890	0.831	0.808	0.816	0.848	0.888	0.929	0.966
					MA(1)					
θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
CPL_1	1.000	0.961	0.845	0.671	0.468	0.278	0.135	0.051	0.013	0.001
CTL_2	1.000	0.989	0.951	0.890	0.728	0.530	0.313	0.137	0.038	0.004
CTL_3	1.000	0.995	0.979	0.940	0.859	0.711	0.491	0.250	0.078	0.009
					ARFIMA$(0,d,0)$					
d	0.01	0.05	0.10	0.15	0.20	0.24	$d \geq 0.25$			
CPL_1	0.627	0.706	0.812	0.894	0.831	0.313	0			
CPL_2	0.776	0.840	0.901	0.932	0.786	0.264	0			
CPL_3	0.841	0.890	0.927	0.907	0.725	0.231	0			

only exists in consecutive pairs, in contrast to the AR(1) model, MCPLE and MCTLE are much inferior to the MLE. This example shows the potential large loss in efficiency using pairwise likelihood, even for short memory models.

Example 3.3. Let $\{X_t\}$ be the long-memory ARFIMA(0,d,0) model defined by

$$(1-B)^d X_t = Z_t\,, \quad \{Z_t\} \sim IID(0,\sigma^2)\,,$$

where B is the lag operator and $d \in (0, 0.5)$. The asymptotic variance of the MLE of d is $6/\pi^2$ and the convergence rate is $\sqrt{n}$ (e.g., Beran (1994)). Since the convergence rate of MCPLE is $\sqrt{n}$ only when $d \in (0, 0.25)$, it only makes sense to

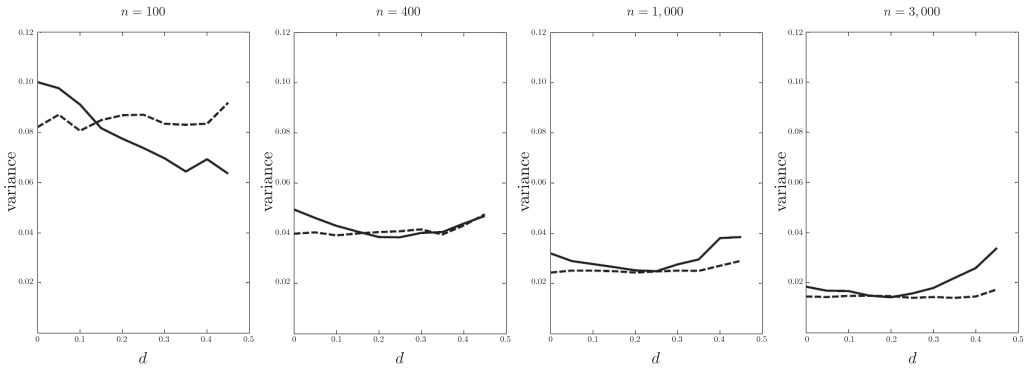

Figure 2. Sample variances of $\hat{d}$ for MLE and CPL_1 of ARFIMA(0,d,0) models for various n, based on 500 replications. See Example 3.3. Solid-line: CPL_1. Dash-line: MLE.

compare the asymptotic variance of the three estimators in this range (see Figure 1c). Even though the dependence of the process is long and the log-likelihood function is not at all similar to the CPL function, the asymptotic variances of MLE and MCPLE are very close in the range $d \in (0, 0.2)$.

As the convergence rate of MCPLE is slower than $\sqrt{n}$ when $d \in [0.25, 0.5)$, MCPLE is less efficient than the MLE. Nevertheless, we explored the empirical variance of the estimators via simulation. Figure 2 shows the sample variance of the MLE and MCPLE for the cases $d = \{0.05, 0.1, 0.15, \ldots, 0.4, 0.45\}$ and sample sizes $n = \{100, 400, 1{,}000, 3{,}000\}$, each based on 500 replications. From the figures, the sample variances decrease as n increases for all three estimators. This demonstrates the consistency of the estimators. As explained by Theorem 2.2, the curve bends upward when $d \geq 0.25$ as n increases as a result of the different convergent rates between $d < 0.25$ and $d \geq 0.25$. Note that even for large sample size $n = 3{,}000$, the variance of MCPLE is only about two times higher than that of MLE in the range $d = [0.25, 0.5)$. Since the computation of MCPLE can be done much more efficiently than the MLE, the MCPLE might be used as a preliminary estimator.

From the three examples shown above, we note that MCPLE achieves the same efficiency as MLE in the AR(1) model. For the MA(1) model, the efficiency of MCPLE is poor with ARE values ranging from 1 to 0. On the other hand, for long-memory models such as ARFIMA models with $d \in (0, 0.25)$, the loss in efficiency of MCPLE is slight with ARE values around 0.8. For cases of ARFIMA models with $d \in (0.25, 0.5)$, the ARE is 0. So while MCPLE may work well for many time series models, its performance can suffer for both short-memory (e.g., MA(1)) and long-memory (e.g., ARFIMA with $d \in (0.25, 0.5)$) models.

Example 3.4. In this example we consider a nonlinear time series model for time series of counts. The Poisson-autoregressive (PAR) model assumes that there is a latent autoregressive process α_t underlining the observation Y_t such that given the α_t process, the Y_t are independent and Poisson-distributed with mean $\lambda_t = e^{\beta+\alpha_t}$. The conditional probability density of Y_t is described by

$$p(y_t|\alpha_t;\psi) = e^{-e^{\beta+\alpha_t}} \frac{e^{(\beta+\alpha_t)y_t}}{y_t!}$$
$$\alpha_t = \phi\alpha_{t-1} + \eta_t\,,$$

where $\{\eta_t\} \sim IIDN(0,\sigma^2)$, $|\phi| < 1$, and $\psi = \{\beta, \phi, \sigma^2\}$ is the parameter vector. Let the observed data be $\mathbf{y}_n = (y_1, \ldots, y_n)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$. The exact likelihood is given by

$$L(\psi;\mathbf{y_n}) = \int \prod_{\mathbf{t=1}}^{\mathbf{n}} \mathbf{p}(\mathbf{y_t}|\alpha_\mathbf{t};\psi)\mathbf{f}_\psi(\boldsymbol{\alpha})\mathbf{d}\boldsymbol{\alpha}, \tag{3.1}$$

where $f_\psi(\boldsymbol{\alpha})$ is the joint distribution of $\boldsymbol{\alpha}$. Since the integral in (3.1) is high dimensional, it is infeasible to compute the exact likelihood directly. Computational intensive simulation methods such as importance sampling or MCMC have been used to estimate the likelihood. (see Davis and Rodriguez-Yam (2005)). We compare the result of MCPLE to the estimation using likelihood approximation in Davis and Rodriguez-Yam (2005). The consecutive pairwise log-likelihood is given by

$$CPL(\psi;\mathbf{y}_n) = \log\left[\prod_{t=1}^{n-1}\int\int p(y_t|\alpha_t;\psi)p(y_{t+1}|\alpha_{t+1};\psi)f_\psi(\alpha_t,\alpha_{t+1})d\alpha_t d\alpha_{t+1}\right].$$

Here the computation reduces to $n-1$ double integrals, which can be efficiently evaluated by numerical methods such as Gauss-Hermite quadrature.

Davis and Rodriguez-Yam (2005) studied the performance of a likelihood-approximation based estimator called AIS, which works as follows. After computing a posterior mode $\boldsymbol{\alpha}^*$, the likelihood $L(\psi, \mathbf{y_n})$ can be expressed as

$$L(\psi;\mathbf{y_n}) = L_a(\psi;\mathbf{y_n},\boldsymbol{\alpha}^*)Er_a(\psi)\,,$$

where $L_a(\psi;\mathbf{y_n},\boldsymbol{\alpha}^*)$ is an approximation of $L(\psi;\mathbf{y_n})$ with closed form expression and $Er_a(\psi)$ is the associated approximation error. One can estimate ψ by maximizing $L_a(\psi;\mathbf{y_n},\boldsymbol{\alpha}^*)$ numerically. Alternatively, $Er_a(\psi)$ can be linearized around the maximizer of $L_a(\psi;\mathbf{y_n},\boldsymbol{\alpha}^*)$ and then estimated using importance sampling. Putting these two pieces together we can produce a quick simulation procedure for optimizing an approximation to $L(\psi;\mathbf{y_n})$.

Table 3. Comparison of AIS and CL estimates for PAR models. $n = 500$ and 500 replications. rmse stands for root mean square errors.

	β	ϕ	σ	β	ϕ	σ	β	ϕ	σ
TRUE	-0.613	-0.500	1.236	-0.613	0.500	1.236	-0.613	0.900	0.622
AIS-biases	-0.031	0.022	0.052	-0.056	-0.005	0.094	-0.001	0.010	0.010
AIS-rmse	0.093	0.063	0.100	0.143	0.065	0.120	0.294	0.029	0.058
CPL_1-biases	0.105	-0.005	-0.148	0.120	0.004	-0.134	0.842	0.019	-0.389
CPL_1-rmse	0.095	0.065	0.134	0.142	0.057	0.128	0.199	0.050	0.141
CPL_2-biases	0.101	-0.001	-0.137	0.110	-0.005	-0.110	0.483	-0.016	-0.105
CPL_2-rmse	0.130	0.077	0.174	0.184	0.079	0.173	0.466	0.061	0.184
CPL_3-biases	0.096	0.004	-0.128	0.104	-0.006	-0.129	0.477	-0.016	-0.109
CPL_3-rmse	0.134	0.074	0.162	0.187	0.080	0.164	0.470	0.064	0.182
TRUE	0.150	-0.500	0.619	0.150	0.500	0.619	0.150	0.900	0.312
AIS-biases	-0.004	0.000	0.009	-0.001	0.012	0.010	-0.001	0.010	0.001
AIS-rmse	0.049	0.089	0.059	0.073	0.091	0.061	0.148	0.037	0.048
CPL_1-biases	-0.001	-0.004	-0.003	-0.001	-0.003	-0.001	-0.029	-0.021	0.023
CPL_1-rmse	0.041	0.061	0.079	0.060	0.055	0.073	0.421	0.076	0.199
CPL_2-biases	-0.001	0.004	-0.003	-0.002	-0.014	0.005	0.031	0.006	-0.055
CPL_2-rmse	0.055	0.077	0.098	0.090	0.071	0.103	0.256	0.076	0.129
CPL_3-biases	0.005	0.003	-0.003	0.003	-0.012	-0.001	0.035	-0.000	-0.034
CPL_3-rmse	0.058	0.075	0.099	0.096	0.072	0.096	0.252	0.043	0.109
TRUE	0.373	-0.500	0.220	0.373	0.500	0.220	0.373	0.900	0.111
AIS-biases	0.005	-0.065	0.012	0.005	0.123	0.003	0.003	0.078	-0.016
AIS-rmse	0.039	0.383	0.088	0.045	0.393	0.083	0.060	0.231	0.062
CPL_1-biases	0.009	-0.108	-0.053	0.006	0.076	-0.042	0.146	0.061	-0.102
CPL_1-rmse	0.042	0.228	0.110	0.051	0.218	0.106	0.123	0.032	0.013
CPL_2-biases	0.006	-0.071	-0.037	-0.001	0.057	-0.032	0.113	0.067	-0.010
CPL_2-rmse	0.050	0.233	0.110	0.065	0.222	0.101	0.224	0.036	0.015
CPL_3-biases	0.002	-0.045	-0.022	0.006	0.049	-0.026	0.109	0.068	-0.098
CPL_3-rmse	0.051	0.209	0.094	0.066	0.219	0.097	0.259	0.029	0.021

A simulation experiment was conducted to compare CPL with AIS in the same setting as Table 5 in Davis and Rodriguez-Yam (2005). The results are shown in Table 3. In computing the double integrals in CPL, Gauss-Hermite quadrature with 10 nodes in each dimension was used. From the table, we notice that CPL_1 is comparable to AIS. In particular, for the cases where $\phi = 0.5$ or -0.5, CPL_1 has a smaller root-mean-square error (rmse) than AIS for the estimates of ϕ. On the other hand, CPL_2 and CPL_3 do not appear to outperform CPL_1; they have higher rmse than CPL_1 in most of the estimates in the cases where $\phi = 0.5$ or -0.5. However, MCPLE has poor performance for the cases where $\phi = 0.9$.

In the Poisson-Autoregressive model, where computation of the likelihood is not possible without resorting to simulation methods, the pairwise likelihood procedure is an attractive alternative. The performance of the MCPLE is competitive and the theory behind the CPL (consistency and asymptotic normality) are relatively straightforward to derive using standard arguments with mixing conditions. Moreover, it is possible to give estimates of the standard error of the MCPLE. Let

$$\begin{aligned} cpl_t(\psi) &= cpl(\psi; y_t, y_{t+1}) \\ &= \log\left[\int\int p(y_t|\alpha_t;\psi)p(y_{t+1}|\alpha_{t+1};\psi)f_\psi(\alpha_t,\alpha_{t+1})d\alpha_t d\alpha_{t+1}\right]. \end{aligned}$$

Note that $CPL_1(\psi;\mathbf{y_n}) = \sum_{t=1}^{n-1} cpl_t(\psi)$. Let ψ_o and $\hat{\psi}$ be the true value and the CPL_1 estimator of the parameter, respectively. Using a Taylor's series expansion on $CPL_1'(\hat{\psi};\mathbf{y_n},)$ around ψ_o shows that $\sqrt{n}(\hat{\psi}-\psi_o)$ is asymptotically equivalent to

$$-\Big(\frac{1}{n}\sum_{t=1}^{n-1} cpl_t''(\psi_o)\Big)^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n-1} cpl_t'(\psi_o)\,. \tag{3.2}$$

Since $\{cpl_t''(\psi_o)\}$ is an ergodic sequence, we have

$$\frac{1}{n}\sum_{t=1}^{n-1} cpl_t''(\psi_o) \xrightarrow{a.s.} E(cpl_1''(\psi_o))\,.$$

Also, since $\{cpl_t'(\psi_o)\}$ is a stationary, strongly mixing sequence, $(1/\sqrt{n})\sum_{t=1}^{n-1} cpl_t'(\psi_o)$ is asymptotically normal with covariance matrix given by

$$\sum_{n=-\infty}^{\infty}\boldsymbol{\gamma}(n)\,,$$

where $\boldsymbol{\gamma}(n)$ is the auto-covariance matrix of $\{cpl_t'(\psi_o)\}$. A consistent estimator of this quantity is

$$\sum_{k=-r_n}^{r_n}(1-\frac{|k|}{r_n})\hat{\boldsymbol{\gamma}}(k)\,,$$

where $r_n\to\infty$, $r_n/n\to 0$ and

$$\hat{\boldsymbol{\gamma}}(k) = \frac{1}{n}\sum_{t=k+1}^{n-1} cpl_t'(\hat{\psi})cpl_{t-k}'^{T}(\hat{\psi})\,.$$

Table 4. MCPLE and AIS estimates for the polio data. Std err stands for the standard error of the estimates.

Parameter	β_1	β_2	β_3	β_4	β_5	β_6	ϕ	σ^2
AIS	0.239	-3.746	0.161	-0.480	0.414	-0.011	0.661	0.272
Std err	0.285	2.867	0.151	0.164	0.122	0.127	0.209	0.112
CPL_1	0.303	-4.738	0.135	-0.492	0.397	-0.016	0.492	0.372
Std err	0.229	2.531	0.116	0.134	0.101	0.147	0.206	0.136

Thus the asymptotic variance of $\sqrt{n}(\hat{\psi} - \psi_o)$ can be estimated by

$$\left(\frac{1}{n}\sum_{t=1}^{n-1} cpl_t''(\hat{\psi})\right)^{-1} \left(\sum_{k=-r_n}^{r_n} (1 - \frac{|k|}{n})\hat{\gamma}(k)\right) \left(\frac{1}{n}\sum_{t=1}^{n-1} cpl_t''(\hat{\psi})\right)^{-1}. \qquad (3.3)$$

Example 3.5. In this example we apply the MCPLE to the Poisson AR model to the well-known Polio data set consisting of the monthly number of U.S. cases of poliomyelitis from 1970 to 1983. We compare CPL and AIS using the same model as in Davis and Rodriguez-Yam (2005), in which the distribution of Y_t, given the state α_t is Poisson with rate $\lambda_t = e^{\alpha_t + x_t^T \boldsymbol{\beta}}$. Here $\boldsymbol{\beta}^T := (\beta_1, \ldots, \beta_6)$, $\mathbf{x_t}$ is the vector of covariates given by

$$\mathbf{x_t}^T = \Big(1, \frac{t}{1{,}000}, \cos(2\pi\frac{t}{12}), \sin(2\pi\frac{t}{12}), \cos(2\pi\frac{t}{6}), \sin(2\pi\frac{t}{6})\Big),$$

and the state process $\{\alpha_t\}$ is assumed to follow an AR(1) model. The vector of parameters is $\boldsymbol{\psi} = (\beta_1, \ldots, \beta_6, \phi, \sigma^2)$. Table 4 shows the parameter and standard errors estimates of AIS and CPL_1. While the asymptotic variance of the AIS estimates are computed by bootstrap, the asymptotic variance of CPL_1 can be computed efficiently using (24) with $r_n = \sqrt{n} \approx 13$. The estimates of CPL_1 are comparable to that of AIS. In particular, only $\hat{\beta}_4$ and $\hat{\beta}_5$ are significantly different from zero for the regression coefficient estimates under both methods. Besides, the estimated variances of the latent process $\{\alpha_t\}$ for AIS and CPL_1 are in close agreement, $0.272/(1 - 0.661^2) = 0.483$ and $0.372/(1 - 0.492^2) = 0.491$, respectively. The standard error for the parameter estimates in AIS and CPL_1 are also comparable. CPL_2 and CPL_3 produce similar parameter estimates as CPL_1, but their standard error estimates are considerably more difficult to compute.

Acknowledgements

The research of Richard Davis was supported in part by NSF grant DMS-0743459.

References

Beran, J. (1994). *Statistics for Long-Memory Processes*. Chapman & Hall, New York.

Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Method*. Springer, New York.

Caragea, P. and Smith, R. (2007). Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. *J. Multivariate Anal.* **98**, 1417-1440.

Cox, D. R. and Reid, N. (2004). Miscellanea: A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-37.

Davis, R. A. and Rodriguez-Yam, G. (2005). Estimation for state-space models based on a likelihood approximation *Statist. Sinica* **15**, 381-406.

Engle, R. F., Shephard, N. and Sheppard, K. (2009). Fitting vast dimensional time-varying covariance models. *NYU Working Paper No. FIN*-08-009. Available at SSRN: http://ssrn.com/abstract=1354497

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.* **93**, 1099-111.

Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**, 355-66.

Hjort, N. L. and Omre, H. (1994). Topics in spatial statistics. *Scand. J. Statist.* **21**, 289-357.

Hosking, J. R. M. (1996). Asymptotic distributions of the sample mean, autocovariances, and autocorelations of long-memory time series. *J. Econometrics* **73**, 261-284.

Joe, H. and Lee, Y. (2009). On weighting of bivariate margins in pairwise likelihood. *J. Multivariate Anal.* **100**, 670-685.

Kuk, A. Y. and Nott, D. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statist. Probab. Lett.* **47**, 329-35.

Lindsay, B. (1988). Composite likelihood methods. In *Statistical Inference from Stochastic Processes,* (Edited by N. U. Prabhu), 221-39. American Mathematical Society, Providence, RI.

Mardia, K. V., Kent, J. T., Hughes, G. and Taylor, C. C. (2009). Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika* **96**, 975-982.

Nott, D. J. and Ryden, T. (1999). Pairwise likelihood methods for inference in image models. *Biometrika* **86**, 661-76.

Parner, E. T. (2001). A composite likelihood approach to multivariate survival data. *Scand. J. Statist.* **28**, 295-302.

Renard, D., Molenberghs, G. and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Comp. Statist. Data Anal.* **44**, 649-67.

Rosenblatt, M. (1961). Independence and dependence. In *Proceedings of the Fourth Berkeley symposium on Mathematical Statistics and Probability*, **2**, 431-43. University of California Press, Berkeley, CA.

Varin, C. Høst, G. and Skare, Ø. (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Comp. Statist. Data Anal.* **49**, 1173-91.

Varin, C. and Vidoni,P. (2009). Pairwise likelihood inference for general state space models. *Econometric Rev.* **28**, 170-85.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika* **75**, 621-29.

Department of Statistics, Columbia University, New York, NY 10027, U.S.A.
E-mail: rdavis@stat.columbia.edu

Department of Statistics, Chinese University of Hong Kong, Hong Kong.
E-mail: cyyau@sta.cuhk.edu.hk

(Received October 2009; accepted April 2010)