Silvia Columbu, Valentina Mameli, Monica Musio, Philip Dawid, *The Hyvärinen scoring rule in Gaussian linear time series models* in *Journal of Statistical Planning and Inference*, 212 (2021), pp. 126-140.

The Hyvärinen scoring rule in Gaussian linear time series models

Silvia Columbu [a],
Valentina Mameli[b],
Monica Musio[a],
Philip Dawid [c]

Abstract

In this work we study stationary linear time-series models, and construct and analyse ''score-matching'' estimators based on the Hyvärinen scoring rule. We consider two scenarios: a single series of increasing length, and an increasing number of independent series of fixed length. In the latter case there are two variants, one based on the full data, and another based on a sufficient statistic.

We study the empirical performance of these estimators in three special cases, autoregressive (AR), moving average (MA) and fractionally differenced white noise (ARFIMA) models, and make comparisons with full and pairwise likelihood estimators. The results are somewhat model-dependent, with the new estimators doing well for MA and ARFIMA models, but less so for AR models.

Keywords:
Scoring rule estimators Hyvärinen scoring rule Gaussian linear time series

1. Introduction

Composite likelihoods methods have become an appealing tool, as alternative to the likelihood estimation method, in complex statistical models with interdependencies. The increasing importance of this methodology is due to its computational feasibility in a variety of applications. However, for the first order moving average model (MA(1)), the pairwise likelihood method, which is a special case of composite likelihood, has very poor asymptotic efficiency as the moving average parameter tends to the boundary of the parameter space (Davis and Yau, 2011). Composite likelihood estimation methods form a subset of a more general class of methods based on proper scoring rules, estimation being conducted by minimising the empirical score over distributions in the model (Dawid and Musio, 2014; Dawid et al., 2016). Some important proper scoring rules are the log-score (Good, 1952), which recovers the full (negative log) likelihood, the Brier score (Brier, 1950) and the Hyvärinen score (Hyvärinen, 2005). In the setting of MA(1) we consider alternatives to the pairwise likelihood approach, based on the theory of proper scoring rules, focusing on the Hyvärinen score. This score is a homogeneous proper scoring rule (see Ehm and Gneiting (2012) and Parry et al. (2012)), which is unchanged by applying a positive scale factor to the probability distribution. Homogeneous scoring rules have been characterised for continuous real variables (Parry et al., 2012) and for discrete variables (Dawid et al., 2012). In a Bayesian framework, Dawid and Musio (2015) have shown, for the case of continuous variables, how to handle Bayesian model selection with improper within-model prior distributions, by exploiting the use of homogeneous proper scoring rules. The discrete counterpart has been empirically studied by Dawid et al. (2017). In a recent contribution, Shao et al. (2019) consider the use of the Hyvärinen score for model comparison. Although the majority of contributions involving the use of Hyvärinen scoring

[a] *Department of Mathematics and Computer Science, University of Cagliari, Italy*
[b] *Department of Economics and Statistics, University of Udine, Italy*
[c] *Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, UK*

rules focus on Euclidean spaces, scholars have also investigated extensions to non-Euclidean spaces: for an early study see Dawid (2007). Recently, Mardia et al. (2016) proposed an extension of the Hyvärinen scoring rule to compact oriented Riemannian manifolds, and Takasu et al. (2018) constructed a novel class of homogeneous strictly proper scoring rules for statistical models on spheres.

Given the growing interest in the use of this scoring rule, in this paper we aim to derive an estimation method based on the Hyvärinen scoring rule not only for moving average model but in general for estimating linear Gaussian time series models.

We distinguish two separate cases: a first in which the length of a single time series increases to infinity, and a second in which the length of the time series is fixed and the number of series increases to infinity.

The consistency and asymptotic distribution of the Hyvärinen estimator are derived for the case of a single time series of increasing length. In particular, under some mild regularity conditions we derive consistency of the proposed estimator for linear Gaussian time series models, and its asymptotic distribution is found in the specific case of autoregressive moving average (ARMA) causal invertible models. For time series with fixed length and the number of time series increasing to infinity the performances of two estimators based on the Hyvärinen scoring rule, namely *the total Hyvärinen estimator* and *the matrix Hyvärinen estimator*, are compared through simulation studies with the full maximum likelihood and the pairwise maximum likelihood estimators. To evaluate the novel inferential procedure based on the Hyvärinen scoring rule we consider simple situations where the likelihood function is available. In particular, three simple time series models have been considered in the design of simulations: autoregressive (AR), moving average (MA) and fractionally differenced white noise (ARFIMA).

Different behaviours can be detected for the total Hyvärinen estimator among the settings examined. In particular, it outperforms the pairwise likelihood estimator in terms of efficiency for the MA and ARFIMA processes.

The paper unfolds as follows. Section 2 introduces basic notions on scoring rules. In Section 3 we introduce the Hyvärinen scoring rule for Gaussian linear time series. Some asymptotic results for the Hyvärinen estimator are given. In the specific case of n independent series we describe the total Hyvärinen estimator and the matrix Hyvärinen estimator. Section 4 summarises the results of the simulation studies on n time series of fixed length T. Section 5 presents a simulation study for a single time series model and a simple application in a real case study. Section 6 provides some concluding remarks. Technical details are postponed to the Appendix.

2. Scoring rules

A *scoring rule* is a loss function designed to measure the quality of a proposed probability distribution Q, for a random variable X, in light of the outcome x of X. Specifically, if a forecaster quotes a predictive distribution Q for X and the event $X = x$ realises, then the forecaster's loss will be $S(x, Q)$. The expected value of $S(X, Q)$ when X has distribution P is denoted by $S(P, Q)$.

The scoring rule S is *proper* (relative to the class of distributions $\mathcal{P}$) if

$$S(P, Q) \geq S(P, P), \text{ for all } P, \ Q \in \mathcal{P}. \tag{1}$$

It is *strictly proper* if equality obtains only when $Q = P$.

Any proper scoring rule gives rise to a general method for parameter estimation, based on an unbiased estimating equation: see Section 2.2.

2.1. Examples of proper scoring rules

Some important proper scoring rules are the log-score, $S(x, Q) = -\log q(x)$ (Good, 1952), where $q(\cdot)$ is the density function of Q, which recovers the full (negative log) likelihood; and the Brier score (Brier, 1950). A particularly interesting case, which avoids the need to compute the normalising constant, produces the *score matching* estimation method of Hyvärinen (2005), based on the following proper scoring rule:

$$S(\mathbf{x}, Q) = \Delta_{\mathbf{x}} \ln q(\mathbf{x}) + \frac{1}{2} \left\| \nabla_{\mathbf{x}} \ln q(\mathbf{x}) \right\|^2 , \tag{2}$$

where $\mathbf{X}$ ranges over the whole of $\mathbb{R}^p$ supplied with the Euclidean norm $\| \cdot \|$, $q(\cdot)$ is assumed twice continuously differentiable, and $\mathbf{x}$ is the realised value of $\mathbf{X}$. In (2), $\nabla_{\mathbf{x}}$ denotes the gradient operator, and $\Delta_{\mathbf{x}}$ the Laplacian operator, with respect to $\mathbf{x}$. For $p = 1$ we can express

$$S(x, Q) = \frac{q''(x)}{q(x)} - \frac{1}{2} \left(\frac{q'(x)}{q(x)} \right)^2 . \tag{3}$$

The scoring rule (2) is a *2-local homogeneous proper scoring rule* (see Parry et al. (2012)). It is homogeneous in the density function $q(\cdot)$, *i.e.* its value is unaffected by applying a positive scale factor to the density q, and so can be computed even if we only know the density function up to a scale factor. Inference performed using any homogeneous scoring rule does not require knowledge of the normalising constant of the distribution.

2.2. Estimation based on proper scoring rules

Let $(x_1, \ldots, x_n)$ be independent realisations of a random variable X, having distribution $P_{\boldsymbol{\theta}}$ depending on an unknown parameter $\boldsymbol{\theta} \in \Theta$, where Θ is an open subset of $\mathbb{R}^m$. Given a proper scoring rule S, let $S(x, \boldsymbol{\theta})$ denote $S(x, P_{\boldsymbol{\theta}})$.

Inference for the parameter $\boldsymbol{\theta}$ may be performed by minimising the *total empirical score*,

$$S(\boldsymbol{\theta}) = \sum_{p=1}^{n} S(x_p, \boldsymbol{\theta}), \tag{4}$$

resulting in the *minimum score estimator*, $\widehat{\boldsymbol{\theta}}_S = \arg\min_{\boldsymbol{\theta}} S(\boldsymbol{\theta})$.

Under broad regularity conditions on the model (see *e.g.* Barndorff-Nielsen and Cox (1994)), $\widehat{\boldsymbol{\theta}}_S$ satisfies the *score equation*:

$$s(\boldsymbol{\theta}) := \sum_{p=1}^{n} s(x_p, \boldsymbol{\theta}) = 0, \tag{5}$$

where $s(x, \boldsymbol{\theta}) := \nabla_{\boldsymbol{\theta}} S(x, \boldsymbol{\theta})$, the gradient vector of $S(x, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The score equation is an unbiased estimating equation (Dawid and Lauritzen, 2005). When S is the log-score, the minimum score estimator becomes the maximum likelihood estimator.

From the general theory of unbiased estimating functions, under broad regularity conditions on the model the minimum score estimate $\widehat{\boldsymbol{\theta}}_S$ is asymptotically consistent and normally distributed: $\widehat{\boldsymbol{\theta}}_S \sim N(\boldsymbol{\theta}, \{nG(\boldsymbol{\theta})\}^{-1})$, where $G(\boldsymbol{\theta})$ denotes the *Godambe information matrix* $G(\boldsymbol{\theta}) := M(\boldsymbol{\theta})^{\mathrm{T}} V(\boldsymbol{\theta})^{-1} M(\boldsymbol{\theta})$, where $V(\boldsymbol{\theta}) = \mathrm{E}\left\{s(X, \boldsymbol{\theta}) s(X, \boldsymbol{\theta})^{\mathrm{T}}\right\}$ is the *variability matrix*, and $M(\boldsymbol{\theta}) = \mathrm{E}\left\{\nabla_{\boldsymbol{\theta}} s(X, \boldsymbol{\theta})^{\mathrm{T}}\right\}$ is the *sensitivity matrix*. In contrast to the case for the full likelihood, V and M are different in general: see Dawid and Musio (2014) and Dawid et al. (2016). We point out that estimation of the matrix $V(\boldsymbol{\theta})$, and (to a somewhat lesser extent) of the matrix $M(\boldsymbol{\theta})$, is not an easy task: see Varin (2008), Varin et al. (2011) and Cattelan and Sartori (2016).

3. Gaussian linear time series models

In this section we introduce some results based on the use of the Hyvärinen scoring rule in the setting of Gaussian linear time series models.

Let $\boldsymbol{\theta} = (\mu, \sigma^2, \boldsymbol{\lambda})$ be an m-dimensional parameter, where $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$ and $\boldsymbol{\lambda} \in \mathbb{R}^{m-2}$. Consider the Gaussian linear time series model (y_t) defined by

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j z_{t-j}, \quad t = 1, 2, \ldots, \tag{6}$$

where, for $j \geq 0$, $\psi_j = \psi_j(\boldsymbol{\lambda})$ satisfies $\psi_0 = 1$ and $\sum_{t=0}^{\infty} \psi_t^2 < \infty$. The (z_t) are i.i.d. Gaussian variables with mean 0 and variance σ^2. The auto-covariance function is $\mathrm{E}\{(y_{t+j}-\mu)(y_t-\mu)\} = \sigma^2 \sum_{t=0}^{\infty} \psi_t \psi_{t+j} = \sigma^2 \gamma_{\lambda}(j)$, where $\gamma_{\lambda}(j) = \sum_{t=0}^{\infty} \psi_t \psi_{t+j}$ is twice continuously differentiable for all j. Using basic differentiation rules, it is easy to find the Hyvärinen score based on the single time series $Y_T = (y_1, y_2, \ldots, y_T)$:

$$H(Y_T, \boldsymbol{\theta}) = -\frac{1}{\sigma^2} \sum_{i=1}^{T} \Gamma^{ii} + \frac{1}{2} \sum_{i=1}^{T} \left\{ \sum_{t=1}^{T} \frac{1}{\sigma^2} \Gamma^{it} (y_t - \mu) \right\}^2, \tag{7}$$

where the matrix Γ has (i, j) entry $\Gamma_{ij} = \gamma_{\lambda}(|i-j|)$ and Γ^{ij} is the (i, j) entry of Γ^{-1}. We will denote the Hyvärinen estimator based on a single series by $\widehat{\boldsymbol{\theta}}_{\mathrm{H}}$.

3.1. Asymptotic results for a single time series

In this section we analyse the asymptotic statistical properties of the Hyvärinen scoring rule estimator, based on (7), for a single time series.

The following theorem shows the consistency of the estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{H}}$ in the Gaussian linear time series setting. The proof of the Theorem is deferred to the Appendix and follows arguments similar to those used by Davis and Yau (2011) to prove the consistency of the pairwise likelihood estimator.

Theorem 3.1. *Suppose (y_t) is the linear process in (6) with $\mu = 0$ and parameter $\boldsymbol{\theta}_0 = (\sigma_0^2, \boldsymbol{\lambda_0})$. Let*

$$\widehat{\boldsymbol{\theta}}_{\mathrm{H}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, H(Y_T, \boldsymbol{\theta})$$

be the minimum score estimator, where $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\lambda})$ and $\boldsymbol{\lambda} \in \Lambda$, where Λ is a compact set. If the identifiability condition

$$\sigma_1^2 \gamma_{\lambda_1}(j) = \sigma_2^2 \gamma_{\lambda_2}(j) \text{ iff } (\sigma_1^2, \boldsymbol{\lambda}_1) = (\sigma_2^2, \boldsymbol{\lambda}_2) \tag{8}$$

is satisfied, then $\widehat{\boldsymbol{\theta}}_{\mathrm{H}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$ as $T \to \infty$.

Once consistency has been proved, we focus on the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_{\mathrm{H}}$. Its analytic form involves the elements Γ^{ij} of the inverse of the auto-covariance matrix. In order to guarantee its absolute summability, we restrict our attention to the case of ARMA causal invertible processes.

Defining $b_{ij} = \Gamma^{ij}/\sigma^2$, the gradient and the hessian with respect to $\widehat{\boldsymbol{\theta}}_{\mathrm{H}}$ are given, respectively, by the following two expressions:

$$J(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} H(Y_T, \boldsymbol{\theta}) = \left(\frac{\partial H(Y_T, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = -\sum_{i=1}^{T} \nabla_{\boldsymbol{\theta}}(b_{ii}) + \sum_{i,j,t=1}^{T} b_{it} \nabla_{\boldsymbol{\theta}}(b_{ij}) y_j y_t \tag{9}$$

$$K(\boldsymbol{\theta}) = \frac{\partial J(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial^2 H(Y_T, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} = -\sum_{i=1}^{T} \frac{\partial^2 b_{ii}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} + \sum_{i,j,t=1}^{T} \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}} \left(\frac{\partial b_{it}}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} y_j y_t + \sum_{i,j,t=1}^{T} \frac{\partial^2 b_{ij}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} b_{it} y_j y_t \tag{10}$$

where $\nabla_{\boldsymbol{\theta}} = \partial/\partial\boldsymbol{\theta}$ denotes differentiation with respect to the components of the vector $\boldsymbol{\theta}$.

Theorem 3.2. *Suppose that (y_t) is an ARMA(p, q) causal and invertible process. If the identifiability condition (8) holds, then*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_{m-1}\left(0, M(\boldsymbol{\theta}_0)^{-1} V(\boldsymbol{\theta}_0) M^T(\boldsymbol{\theta}_0)^{-1}\right),$$

where $M(\boldsymbol{\theta}_0)$ is invertible in a neighbourhood of $\boldsymbol{\theta}_0$ and equal to

$$M(\boldsymbol{\theta}_0) = \sum_{r,k=-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}_0} \frac{\gamma^{-1}(k)}{\sigma_0^2} \left(\frac{\partial \gamma^{-1}(k+r)}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \gamma(r)$$

and

$$V(\boldsymbol{\theta}_0) = \left(\frac{\partial}{\partial \boldsymbol{\theta}_0} \frac{\gamma^{-1}(0)}{\sigma_0^2}\right) \left(\frac{\partial}{\partial \boldsymbol{\theta}_0} \frac{\gamma^{-1}(0)}{\sigma_0^2}\right)^{\mathrm{T}} + \sum_{r,k=-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}_0} \frac{\gamma^{-1}(k)}{\sigma_0^2} \gamma^{-1}(0) \left(\frac{\partial \gamma^{-1}(k+r)}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \gamma(r).$$

Theorem 3.2 shows that the Hyvärinen scoring rule estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{H}}$, in the case that (y_t) is an ARMA causal invertible process, is asymptotically normally distributed with rate of decay $\sqrt{T}$. As is well known, the auto-covariance function of an ARMA process decays exponentially, which means that an ARMA process is a short memory process, and its auto-covariance function is absolutely summable (Brockwell and Davis, 1991). This property, together with the duality of ARMA models under causality and invertibility, allows us to prove asymptotic normality. For the complete proof refer to the Appendix. These results are based on the first-order approximations to the distribution of the Hyvärinen scoring rule estimator, providing a satisfactory approximation for large sample sizes, but may be unreliable for small values of T.

3.2. Estimation approaches for n independent time series

In the remainder of this section we discuss the case of n independent series of length T. We assume that T is fixed while n increases to infinity. We also specialise to the case that the common mean μ and innovation variance $\sigma^2 = \sigma_0^2$ are known; without loss of generality we take $\mu = 0$.

Consider now n independent and identically distributed processes $Y_1, \ldots, Y_n$, where $Y_p = (y_{p1}, \ldots, y_{pT})$, each having the T-variate normal distribution with mean-vector 0 and variance covariance matrix $\sigma^2\Gamma$, with unknown parameter $\boldsymbol{\lambda}$. Let the $(n \times T)$ random matrix Y have the vector Y_p as its pth row. We define the *total Hyvärinen score* (HT) as the sum of n single Hyvärinen scores in (7):

$$\mathrm{HT}(\boldsymbol{\lambda}) = \sum_{p=1}^{n} H_p(Y_p, \boldsymbol{\lambda}), \tag{11}$$

where

$$H_p(Y_p, \boldsymbol{\lambda}) = -\frac{1}{\sigma^2} \sum_{i=1}^{T} \Gamma^{ii} + \frac{1}{2} \sum_{i=1}^{T} \left\{ \sum_{t=1}^{T} \frac{1}{\sigma^2} \Gamma^{it} y_{pt} \right\}^2. \tag{12}$$

The estimate of $\boldsymbol{\lambda}$ minimising the total Hyvärinen score will be denoted by $\widehat{\boldsymbol{\lambda}}_{\mathrm{HT}}$.

An alternative approach is to consider as basic observable the sum-of-squares-and-products matrix $\mathrm{SSP} = Y^T Y$, which is a sufficient statistic for the multivariate normal model, having the Wishart distribution with n degrees of freedom and scale matrix $\sigma^2 \Gamma$. Then inference for the parameter $\boldsymbol{\lambda}$ can be performed by resorting to the Hyvärinen score based directly on the Wishart model. We will call this scoring rule the *matrix Hyvärinen score*.

Assuming $n \geq T$, so that the joint distribution of the upper triangle $\left(s_{ij} : 1 \leq i \leq j \leq T\right)$ of the sum-of-squares-and-products random matrix SSP (which has a Wishart distribution with parameters n and $\sigma^2\Gamma$) has a density, and taking into consideration all of the properties of the derivatives of traces and determinants, it can be shown that the Hyvärinen score based on this joint density is

$$\mathrm{HW}(\mathrm{SSP}, \Gamma) = -\frac{(n-T-1)}{2} \sum_{i=1}^{T} (s^{ii})^2 + \frac{1}{2} \sum_{i,j=1}^{T} \left\{ \frac{(n-T-1)}{2} s^{ij} - \frac{1}{2\sigma^2} \Gamma^{ij} \right\}^2 , \tag{13}$$

where s^{ij} and Γ^{ij} are the elements of the inverse matrices SSP^{-1} and Γ^{-1}, respectively. The matrix Hyvärinen estimator for $\boldsymbol{\lambda}$, minimising $\mathrm{HW}(\mathrm{SSP}, \Gamma)$ with respect to $\boldsymbol{\lambda}$, will be denoted by $\widehat{\boldsymbol{\lambda}}_{\mathrm{HW}}$.

The derivative of $\mathrm{HW}(\mathrm{SSP}, \Gamma)$ with respect to $\boldsymbol{\lambda}$ is

$$\mathrm{HW}_{\boldsymbol{\lambda}}(\mathrm{SSP}, \Gamma) = -\frac{1}{2\sigma^2} \sum_{i,j=1}^{T} \left\{ \frac{(n-T-1)}{2} s^{ij} - \frac{1}{2\sigma^2} \Gamma^{ij} \right\} \frac{\partial \Gamma^{ij}}{\partial \boldsymbol{\lambda}}, \tag{14}$$

and $\mathrm{E}\left\{\mathrm{HW}_{\boldsymbol{\lambda}}(\mathrm{SSP}, \Gamma)\right\} = 0$ since $\mathrm{E}\left(s^{ij}\right) = \Gamma^{ij}/(\sigma^2(n-T-1))$ (see Von Rosen (1997)).

Moreover, $M(\boldsymbol{\lambda}) = \mathrm{E}\left\{\mathrm{HW}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\mathrm{SSP}, \Gamma)\right\} = (1/(4\sigma^4)) \sum_{i,j=1}^{T} \left(\partial \Gamma^{ij}/\partial \boldsymbol{\lambda}\right) \left(\partial \Gamma^{ij}/\partial \boldsymbol{\lambda}\right)^{\mathrm{T}}$. The function $V(\boldsymbol{\lambda})$, calculated after taking account of (14),

$$V(\boldsymbol{\lambda}) = \frac{(n-T-1)^2}{16\sigma^4} \sum_{i,j,k,l=1}^{T} \frac{\partial \Gamma^{ij}}{\partial \boldsymbol{\lambda}} \left(\frac{\partial \Gamma^{kl}}{\partial \boldsymbol{\lambda}} \right)^{\mathrm{T}} \mathrm{cov}\left(s^{ij}, s^{kl}\right), \tag{15}$$

involves calculations requiring the covariance matrix of the random matrix SSP^{-1}, which has an Inverse Wishart distribution with scale matrix $\frac{1}{\sigma^2}\Gamma^{-1}$: see Von Rosen (1997) for details on the components of the covariance matrix.

In general, the Godambe information needed to estimate the standard error of $\widehat{\boldsymbol{\lambda}}_{\mathrm{HW}}$ is not easy to derive analytically due to the form of the matrix Γ. It should be pointed out that this approach cannot be used if we have a single time series of length T with T increasing to ∞, since for non-singularity of the Wishart distribution we need to assume $n \geq T$.

4. Numerical assessment on *n* time series of fixed length *T*

In this section we report simulation studies designed to assess and compare the behaviours of the estimators obtained by using the total and the matrix Hyvärinen estimators. We refer to the case described in paragraph 3.2 in which T is fixed and n increases to ∞. For comparison, we will consider also the full and pairwise maximum likelihood estimators (Davis and Yau, 2011). We discuss three examples: the first order autoregressive AR(1), the first order moving average MA(1) and the fractionally differenced white noise ARFIMA(0, d, 0). Various parameter settings are considered in all simulation studies. All calculations have been done in the statistical computing environment R (R Core Team, 2019). The summary statistics shown are: average estimates of the parameters, asymptotic standard deviations (*sd*) and the asymptotic relative efficiency (ARE) with respect to the maximum likelihood estimator.

4.1. First order autoregressive models

The stationary univariate autoregressive process of order 1, denoted by AR(1), is defined by

$$\begin{aligned} y_1 &= \mu + \frac{1}{\sqrt{1-\phi^2}} z_1 \\ y_t &= \mu + \phi(y_{t-1} - \mu) + z_t, \quad (t = 2, \ldots, T), \end{aligned}$$

where (z_t) is a Gaussian white noise process with mean 0 and variance σ^2. Let $\boldsymbol{\theta} = (\sigma^2, \lambda) = (\sigma^2, \phi)$, where λ is represented by the scalar parameter ϕ. Here ϕ, with $|\phi| < 1$, is the *autoregressive parameter*. Then $y_1, \ldots, y_T$ are jointly normal with mean vector $\mu 1_T$ (where 1_T is the T-dimensional unit vector), and covariance matrix $\sigma^2\Gamma$, with Γ having components $\Gamma_{lm} = \phi^{|l-m|}/(1-\phi^2)$ $(l, m = 1, \ldots, T)$.

For comparison purposes we consider also the numerical performance of a class of pairwise likelihood estimators. Since, in the time series considered, dependence decreases in time, as in Davis and Yau (2011) we shall restrict to the *first order consecutive pairwise likelihood*, rather than the complete pairwise likelihood, so that adjacent observations are more closely related than the others. This choice is motivated also by the loss in efficiency incurred in using the kth order consecutive pairwise likelihood as k increases (see Davis and Yau (2011), Joe and Lee (2009). Note that, when it is known that $\mu = 0$ but σ^2 is unknown, the pairwise likelihood estimator of ϕ is $\widehat{\phi}_{\mathrm{PL}} = 2\sum_{t=2}^{T} y_t y_{t-1} / \sum_{t=2}^{T} (y_t^2 + y_{t-1}^2)$, which is also the Yule–Walker estimator (Davis and Yau, 2011).

Table 1
Simulation 1. Estimated mean (*Est.*), asymptotic standard deviation (*sd*), and Asymptotic Relative Efficiency (ARE) of estimators of the parameter ϕ in the AR(1) model, for $n = 200$, $T = 50$, and varying values of ϕ. We denote by $\widehat{\phi}$ the maximum likelihood estimate, by $\widehat{\phi}_{\text{PL}}$ the pairwise likelihood estimate, and by $\widehat{\phi}_{\text{HT}}$ and $\widehat{\phi}_{\text{HW}}$ the total and the matrix Hyvärinen estimates, respectively.

ϕ	$\widehat{\phi}$		$\widehat{\phi}_{\text{PL}}$			$\widehat{\phi}_{\text{HT}}$			$\widehat{\phi}_{\text{HW}}$		
	Est.	*sd*	*Est.*	*sd*	ARE	*Est.*	*sd*	ARE	*Est.*	*sd*	ARE
−0.9	−0.8997	0.0041	−0.8997	0.0045	0.8625	−0.9008	0.0150	0.0738	−0.9004	0.0244	0.0278
−0.8	−0.8000	0.0059	−0.7999	0.0064	0.8340	−0.8007	0.0146	0.1602	−0.8007	0.0236	0.0613
−0.7	−0.7002	0.0071	−0.7001	0.0079	0.8087	−0.7007	0.0139	0.2599	−0.7005	0.0226	0.0979
−0.6	−0.6002	0.0080	−0.6002	0.0089	0.7986	−0.6008	0.0130	0.3794	−0.6008	0.0216	0.1367
−0.5	−0.5001	0.0087	−0.4999	0.0097	0.8069	−0.5009	0.0122	0.5060	−0.5011	0.0202	0.1853
−0.4	−0.4002	0.0092	−0.4000	0.0101	0.8351	−0.4006	0.0115	0.6466	−0.4001	0.0184	0.2505
−0.3	−0.2997	0.0096	−0.2997	0.0102	0.8808	−0.2998	0.0109	0.7773	−0.2995	0.0164	0.3438
−0.2	−0.2003	0.0099	−0.2002	0.0102	0.9347	−0.2005	0.0104	0.8991	−0.2007	0.0143	0.4780
−0.1	−0.0997	0.0100	−0.0997	0.0101	0.9813	−0.0997	0.0102	0.9776	−0.0999	0.0125	0.6493
0	0.0002	0.0101	0.0002	0.0101	0.9998	0.0002	0.0101	1.0077	0.0003	0.0117	0.7401
0.1	0.1005	0.0100	0.1005	0.0101	0.9810	0.1005	0.0101	0.9810	0.1007	0.0125	0.6506
0.2	0.1997	0.0099	0.1997	0.0102	0.9350	0.1998	0.0104	0.8980	0.1995	0.0143	0.4802
0.3	0.2997	0.0096	0.2997	0.0102	0.8808	0.2998	0.0109	0.7774	0.2995	0.0164	0.3433
0.4	0.3993	0.0092	0.3993	0.0101	0.8355	0.3997	0.0115	0.6451	0.3995	0.0184	0.2506
0.5	0.5002	0.0087	0.5003	0.0097	0.8071	0.5006	0.0122	0.5077	0.5004	0.0201	0.1867
0.6	0.5997	0.0080	0.5997	0.0089	0.7985	0.5998	0.0130	0.3757	0.5990	0.0215	0.1376
0.7	0.6992	0.0071	0.6992	0.0079	0.8087	0.6997	0.0138	0.2630	0.6993	0.0227	0.0977
0.8	0.8001	0.0058	0.8001	0.0064	0.8343	0.8006	0.0146	0.1605	0.8002	0.0235	0.0618
0.9	0.8998	0.0041	0.8998	0.0044	0.8622	0.8999	0.0150	0.0734	0.8987	0.0244	0.0278

Simulation 1. The values of the model parameters are $\mu = 0$ and $\sigma = 1$, with the autoregressive parameter $\phi \in \{-0.9, -0.8, \ldots, 0.8, 0.9\}$. In the simulation study, 1000 replicates are generated of $n = 200$ processes of length $T = 50$. Results are summarised in Table 1. The numerical results in Table 1 and in the panel (a) of Fig. 1 suggest that $\widehat{\phi}_{\text{HT}}$ and $\widehat{\phi}_{\text{HW}}$ do not have high efficiency as ϕ approaches 1: in particular, the asymptotic efficiency of $\widehat{\phi}_{\text{HW}}$ tends to 0 for large values of $|\phi|$. In contrast, under the same model setting, there is only a modest loss of efficiency for the pairwise likelihood-based estimator $\widehat{\phi}_{\text{PL}}$.

4.2. First order moving average models

The univariate moving average process of order 1, denoted by MA(1), is defined by

$$y_t = \mu + \alpha z_{t-1} + z_t, \qquad (t = 1, \ldots, T), \tag{16}$$

where $|\alpha| < 1$ and $z_0, \ldots, z_T$ are independent Gaussian random variables with 0 mean and variance σ^2. Let $\boldsymbol{\theta} = (\sigma^2, \lambda) = (\sigma^2, \alpha)$, where λ is represented by the scalar parameter α. Then the random variables $y_1, \ldots, y_T$ are jointly normal, each having mean μ and variance $\sigma^2(1 + \alpha^2)$. The variables y_t and y_{t+k} are independent for $|k| > 1$, while y_t and y_{t+1} have covariance $\sigma^2\alpha$ $(t = 1, \ldots, T-1)$. Hence, the covariance matrix $\sigma^2\Gamma$ of $Y = (y_1, y_2, \ldots, y_T)$ has components $\sigma^2\Gamma_{ss} = \sigma^2(1+\alpha^2)$, $\sigma^2\Gamma_{st} = \sigma^2\alpha$ if $|s-t| = 1$, $\sigma^2\Gamma_{st} = 0$ otherwise.

As before, we consider the first order consecutive pairwise likelihood since the use of a higher order consecutive pairwise likelihood is unrealistic due to the independence of y_t and y_{t+k} for $k \geq 2$. For $t = 1, \ldots, T-1$, the pair (y_t, y_{t+1}) has a bivariate Gaussian distribution, in which the two components both have mean μ and variance $\sigma^2(1+\alpha^2)$, and have covariance $\sigma^2\alpha$.

Simulation 2. The values of the model parameters are $\mu = 0$ and $\sigma = 1$, with the moving average parameter $\alpha \in \{-0.9, -0.8, \ldots, 0.8, 0.9\}$. In the simulation study, 1000 replicates are generated of $n = 200$ processes of length $T = 50$. Results are summarised in Table 2. The simulation shows that the total Hyvärinen estimator $\widehat{\alpha}_{\text{HT}}$ achieves the same efficiency as the MLE in the MA(1) model for values of the moving average parameter near 0; see Table 2 and panel (b) of Fig. 1. However, the loss in efficiency of the total Hyvärinen estimator $\widehat{\alpha}_{\text{HT}}$ is modest even when the absolute value of the moving average parameter reaches 1. In contrast, the pairwise likelihood estimator $\widehat{\alpha}_{\text{PL}}$ shows very poor performances in terms of asymptotic relative efficiency: the ARE ranges from 1 to 0.1 as $|\alpha|$ increases.

4.3. Fractionally differenced white noise

The fractionally differenced white noise, ARFIMA(0, d, 0), model is defined by

$$(1 - \Pi)^d y_t = z_t, \text{ with } t = 1, \ldots, T,$$

where Π is the lag operator and $d \in (0, 0.5)$, and $z_1, \ldots, z_T$ are independent Gaussian random variables with 0 mean and variance σ^2. Let $\boldsymbol{\theta} = (\sigma^2, \lambda) = (\sigma^2, d)$, where λ is represented by the scalar parameter d. Then the random variables

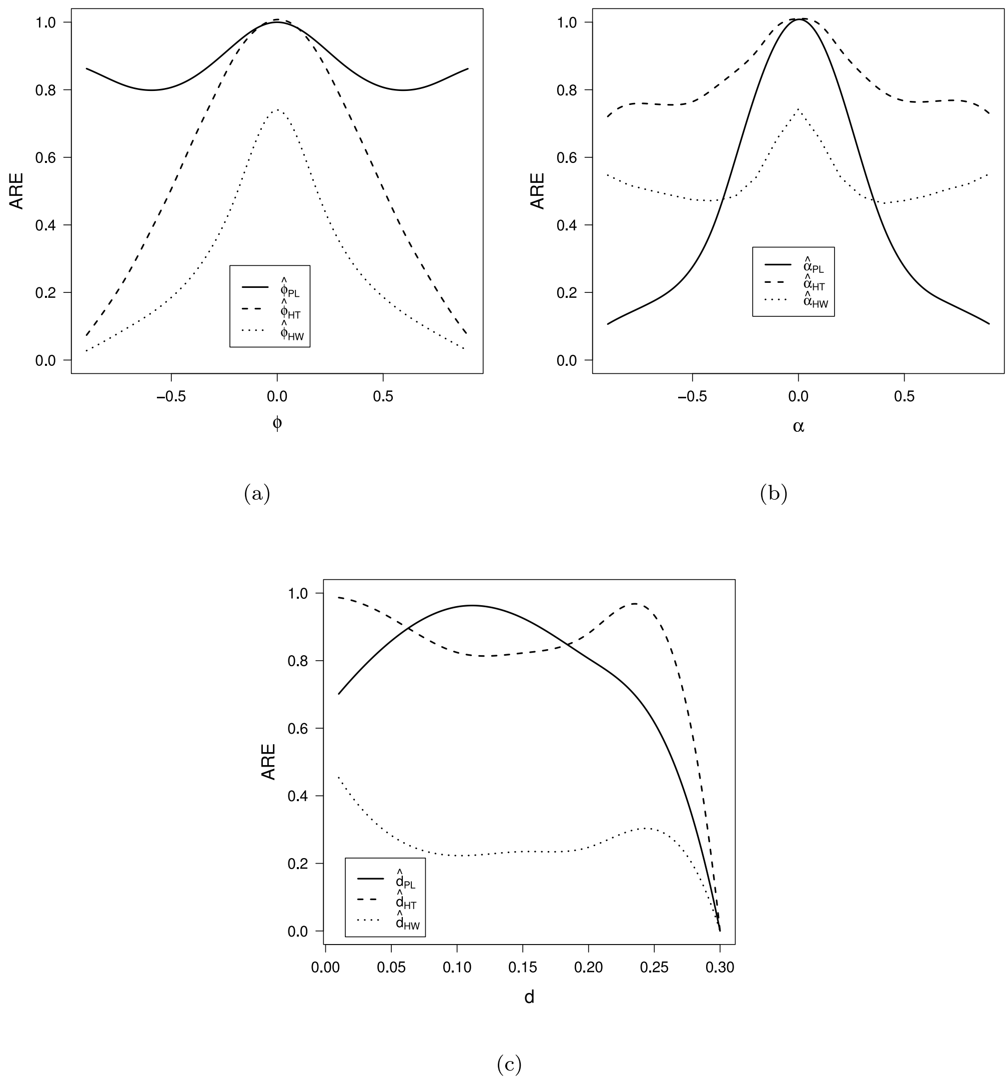


Fig. 1. Asymptotic Relative Efficiency (ARE) of estimators for the AR(1) model (Panel (a)), for the MA(1) model (Panel (b)) and for the ARFIMA(0, d, 0) model (Panel (c)).

$y_1, \ldots, y_T$ are jointly normal, with covariance matrix $\sigma^2 \Gamma$ whose components (see Hosking (1981)) are

$$\sigma^2 \Gamma_{ij} = \frac{(-1)^{|k|}\sigma^2 \Gamma(1-2d)}{\Gamma(|k|-d+1)\Gamma(-|k|-d+1)} \qquad (k = i - j) \tag{17}$$

(where in the right-hand side of (17), Γ denotes the gamma function.)

As before, we consider the first order consecutive pairwise likelihood since no great improvement can be detected by using a higher order consecutive pairwise likelihood: see the results of Davis and Yau (2011). For $t = 1, \ldots, T-1$, the pair (y_t, y_{t+1}) has a bivariate Gaussian distribution, in which the two components both have mean μ and variance $\sigma^2\Gamma(1-2d)/\Gamma(1-d)^2$, and have covariance $-\sigma^2\Gamma(1-2d)/\Gamma(2-d)\Gamma(-d)$.

Simulation 3. The values of the model parameters are $\mu = 0$ and $\sigma = 1$, with the fractional parameter $d \in \{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$. In the simulation study, 1000 replicates are generated of $n = 100$ processes of length $T = 50$. Results are summarised in Table 3. Simulation 3 shows that the total Hyvärinen estimator $\widehat{d}_{\mathrm{HT}}$ achieves the same efficiency as the MLE in the ARFIMA(0, d, 0) model near 0 and near 0.3; see Table 3 and panel (c) of Fig. 1. The loss in efficiency of the total Hyvärinen estimator $\widehat{d}_{\mathrm{HT}}$ is very slight when $d \in (0, 0.3)$. The efficiency of $\widehat{d}_{\mathrm{HW}}$ is poor with ARE values ranging from 0 to 0.45. For all the estimators considered the ARE is 0 when $d \in (0.3, 0.5)$. The pairwise estimator $\widehat{d}_{\mathrm{PL}}$ performs better than $\widehat{d}_{\mathrm{HW}}$, however the values of ARE range from 0.6 to 0.96, reaching a maximum when $d = 0.1$, with a major loss of efficiency with respect to the total Hyvärinen estimator.

Table 2
Simulation 2. Estimated mean (*Est.*), asymptotic standard deviation (*sd*), and Asymptotic Relative Efficiency (ARE) of estimators of the parameter α in the MA(1) model, for $n = 200$, $T = 50$, and varying values of α. We denote by $\widehat{\alpha}$ the maximum likelihood estimate, by $\widehat{\alpha}_{\text{PL}}$ the pairwise likelihood estimate, and by $\widehat{\alpha}_{\text{HT}}$ and $\widehat{\alpha}_{\text{HW}}$ the total and the matrix Hyvärinen estimates, respectively.

α	$\widehat{\alpha}$		$\widehat{\alpha}_{\text{PL}}$			$\widehat{\alpha}_{\text{HT}}$			$\widehat{\alpha}_{\text{HW}}$		
	Est.	*sd*	*Est.*	*sd*	ARE	*Est.*	*sd*	ARE	*Est.*	*sd*	ARE
−0.9	−0.8998	0.0055	−0.8996	0.0167	0.1064	−0.8999	0.0064	0.7208	−0.8993	0.0074	0.5471
−0.8	−0.7997	0.0066	−0.7996	0.0176	0.1390	−0.7998	0.0075	0.7566	−0.7992	0.0091	0.5177
−0.7	−0.6997	0.0075	−0.6996	0.0183	0.1692	−0.6997	0.0086	0.7583	−0.6993	0.0106	0.5020
−0.6	−0.6004	0.0083	−0.6005	0.0182	0.2080	−0.6007	0.0095	0.7553	−0.6003	0.0119	0.4878
−0.5	−0.5004	0.0089	−0.4999	0.0169	0.2757	−0.5007	0.0101	0.7646	−0.5002	0.0129	0.4743
−0.4	−0.4000	0.0093	−0.3997	0.0148	0.3984	−0.4003	0.0104	0.8038	−0.4001	0.0136	0.4713
−0.3	−0.3003	0.0097	−0.3000	0.0126	0.5905	−0.3006	0.0105	0.8527	−0.3006	0.0139	0.4838
−0.2	−0.2000	0.0099	−0.2002	0.0111	0.7926	−0.2001	0.0104	0.9119	−0.1999	0.0135	0.5408
−0.1	−0.1003	0.0101	−0.1004	0.0103	0.9456	−0.1004	0.0101	0.9882	−0.1006	0.0124	0.6557
0	0.0001	0.0101	0.0001	0.0101	1.0082	0.0001	0.0101	1.0101	0.0005	0.0117	0.7429
0.1	0.1000	0.0101	0.1000	0.0103	0.9526	0.1001	0.0101	0.9933	0.0997	0.0124	0.6554
0.2	0.2000	0.0099	0.2000	0.0111	0.7932	0.2000	0.0104	0.9171	0.1994	0.0135	0.5402
0.3	0.2994	0.0097	0.2996	0.0126	0.5853	0.2994	0.0105	0.8475	0.2992	0.0139	0.4835
0.4	0.4000	0.0093	0.4006	0.0148	0.3979	0.4000	0.0105	0.7938	0.3994	0.0137	0.4639
0.5	0.5002	0.0089	0.5000	0.0169	0.2760	0.5004	0.0101	0.7672	0.5000	0.0129	0.4721
0.6	0.6001	0.0083	0.6000	0.0182	0.2075	0.6001	0.0095	0.7643	0.5993	0.0119	0.4850
0.7	0.6999	0.0075	0.6997	0.0182	0.1707	0.6999	0.0086	0.7682	0.6996	0.0106	0.5047
0.8	0.7999	0.0066	0.7997	0.0175	0.1402	0.8000	0.0075	0.7639	0.7995	0.0091	0.5209
0.9	0.8999	0.0055	0.8997	0.0167	0.1072	0.9000	0.0064	0.7300	0.8995	0.0074	0.5504

Table 3
Simulation 3. Estimated mean (*Est.*), asymptotic standard deviation (*sd*), and Asymptotic Relative Efficiency (ARE) of estimators of the parameter d in the ARFIMA model, for $n = 200$, $T = 50$, and varying values of d. We denote by $\widehat{d}$ the maximum likelihood estimate, by $\widehat{d}_{\text{PL}}$ the pairwise likelihood estimate, and by $\widehat{d}_{\text{HT}}$ and $\widehat{d}_{\text{HW}}$ the total and the matrix Hyvärinen estimates, respectively.

d	$\widehat{d}$		$\widehat{d}_{\text{PL}}$			$\widehat{d}_{\text{HT}}$			$\widehat{d}_{\text{HW}}$		
	Est.	*sd*	*Est.*	*sd*	ARE	*Est.*	*sd*	ARE	*Est.*	*sd*	ARE
0.01	0.0121	0.0059	0.0101	0.007	0.7015	0.0101	0.0059	0.9866	0.0105	0.0087	0.4537
0.05	0.0526	0.0062	0.0499	0.0067	0.8585	0.0499	0.0065	0.9257	0.0504	0.0117	0.2827
0.1	0.1034	0.006	0.0997	0.0062	0.9593	0.1	0.0067	0.8241	0.1001	0.0128	0.223
0.15	0.1545	0.0052	0.15	0.0054	0.9258	0.1503	0.0058	0.8226	0.1497	0.0108	0.2349
0.20	0.2041	0.0038	0.1999	0.0043	0.8061	0.2	0.0041	0.8809	0.1997	0.0077	0.2475
0.25	0.2587	0.0021	0.2499	0.0026	0.6173	0.2499	0.0021	0.9339	0.2495	0.0038	0.3005
0.3	0.3032	0	0.3	0.0009	0	0.3	0.0001	0	0.3	0.0043	0

4.4. Discussion

It should be noted that for the MA(1) and the ARFIMA(0, d, 0) models no analytic expressions for the derivatives of (7) are available. The standard deviations of $\widehat{\phi}_{\text{HT}}$, $\widehat{\alpha}_{\text{HT}}$ and $\widehat{d}_{\text{HT}}$ are empirical estimates of the square root of the Godambe information function, which is obtained by compounding the empirical estimates of V and M. The standard deviations of the pairwise maximum likelihood estimator and the maximum likelihood estimator are obtained using the analytic expressions (see Pace and Salvan (1997)) for the AR(1) model and the empirical counterparts for the MA(1) model. Numerical evaluation of scoring rule derivatives has been carried out using the `R` package `numDeriv`; see Gilbert and Varadhan (2012).

Results from simulations reveal that the estimators considered produce estimates very close to the true values of the parameters.

Results not reported here show that, as expected, the differences in terms of bias among the estimators fade out as the length of the time series increases. Panels (a), (b) and (c) of Fig. 1 depict the asymptotic relative efficiency as a function of ϕ for the AR(1) model, as a function of α for the MA(1) model, and as a function of d for the ARFIMA(0, d, 0) model, respectively.

All the results of the simulation studies are in agreement with the findings of Davis and Yau (2011) who focus on pairwise likelihood-based methods for linear time series.

5. Experiments on a single time series

5.1. Numerical assessment

We consider also a simulation study designed to assess and compare the behaviours of the estimators obtained by using the total Hyvärinen estimators in the case of a single time series with T increasing to ∞. We investigate the case of a

Table 4
Simulation 4. Estimated mean (*Est.*), asymptotic standard deviation (*sd*), and 95% Empirical Coverage Probabilities (ECP) of the parameters α and σ in the MA(1) model, for $T = 400$, varying values of α and $\sigma = 1$. We denote by $\widehat{\alpha}$ and $\widehat{\sigma}$ the maximum likelihood estimates, by $\widehat{\alpha}_{\mathrm{PL}}$ and $\widehat{\sigma}_{\mathrm{PL}}$ the pairwise likelihood estimates, and by $\widehat{\alpha}_{\mathrm{HT}}$ and $\widehat{\sigma}_{\mathrm{HT}}$ the total Hyvärinen estimates.

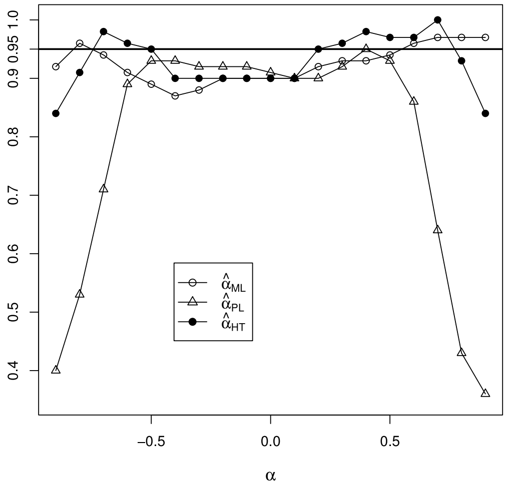
α	Estimates of α									Estimates of σ								
	$\widehat{\alpha}$			$\widehat{\alpha}_{\mathrm{PL}}$			$\widehat{\alpha}_{\mathrm{HT}}$			$\widehat{\sigma}$			$\widehat{\sigma}_{\mathrm{PL}}$			$\widehat{\sigma}_{\mathrm{HT}}$		
	Est.	*sd*	ECP	*Est.*	*sd*	ECP	*Est.*	*sd*	ECP	*Est.*	*sd*	ECP	*Est.*	*sd*	ECP	*Est.*	*sd*	ECP
−0.9	−0.902	0.023	0.92	−0.856	0.128	0.40	−0.913	0.032	0.84	0.998	0.035	0.95	1.013	0.072	0.76	1.074	0.143	0.92
−0.8	−0.799	0.035	0.96	−0.819	0.192	0.53	−0.815	0.042	0.91	0.999	0.035	0.95	0.983	0.099	0.73	1.058	0.109	0.96
−0.7	−0.696	0.037	0.94	−0.746	0.144	0.71	−0.710	0.068	0.98	0.998	0.035	0.95	0.972	0.070	0.78	1.023	0.121	0.99
−0.6	−0.594	0.041	0.91	−0.630	0.135	0.89	−0.602	0.051	0.96	0.998	0.035	0.95	0.981	0.061	0.91	1.005	0.046	0.93
−0.5	−0.493	0.044	0.89	−0.509	0.09	0.93	−0.498	0.051	0.95	0.998	0.035	0.95	0.991	0.043	0.95	1.002	0.040	0.94
−0.4	−0.392	0.046	0.87	−0.402	0.069	0.93	−0.395	0.053	0.9	0.998	0.035	0.95	0.994	0.037	0.94	1	0.04	0.94
−0.3	−0.292	0.048	0.88	−0.298	0.06	0.92	−0.293	0.051	0.9	0.998	0.035	0.95	0.996	0.036	0.95	0.998	0.036	0.94
−0.2	−0.192	0.049	0.9	−0.196	0.054	0.92	−0.193	0.050	0.9	0.998	0.035	0.95	0.997	0.035	0.95	0.998	0.035	0.95
−0.1	−0.093	0.045	0.9	−0.095	0.051	0.92	−0.093	0.05	0.9	0.998	0.036	0.95	0.998	0.0.035	0.95	0.998	0.035	0.95
0	0.006	0.049	0.90	0.007	0.050	0.91	0.005	0.049	0.90	0.998	0.035	0.95	0.998	0.035	0.95	0.998	0.035	0.95
0.1	0.104	0.049	0.90	0.108	0.051	0.90	0.103	0.049	0.90	0.998	0.035	0.95	0.998	0.035	0.95	0.998	0.034	0.95
0.2	0.203	0.048	0.92	0.211	0.054	0.90	0.201	0.050	0.95	0.998	0.035	0.95	0.997	0.035	0.95	0.997	0.035	0.95
0.3	0.302	0.047	0.93	0.314	0.060	0.92	0.300	0.051	0.96	0.998	0.035	0.95	0.995	0.036	0.95	0.997	0.036	0.97
0.4	0.401	0.045	0.93	0.95	0.072	0.95	0.399	0.049	0.98	0.998	0.035	0.95	0.992	0.038	0.95	0.996	0.036	0.96
0.5	0.4998	0.043	0.94	0.534	0.093	0.93	0.498	0.049	0.97	0.998	0.035	0.95	0.986	0.044	0.94	0.996	0.039	0.96
0.6	0.599	0.04	0.96	0.661	0.145	0.86	0.598	0.05	0.97	0.998	0.035	0.95	0.972	0.066	0.89	0.997	0.044	0.96
0.7	0.698	0.036	0.97	0.777	0.152	0.64	0.697	0.055	1.00	0.998	0.035	0.95	0.963	0.076	0.74	0.997	0.072	0.98
0.8	0.797	0.030	0.97	0.843	0.107	0.43	0.798	0.040	0.93	0.998	0.035	0.95	0.976	0.065	0.72	1.006	0.078	0.91
0.9	0.898	0.022	0.97	0.873	0.113	0.36	0.908	0.037	0.84	0.998	0.035	0.95	1.009	0.066	0.79	1.12	0.142	0.82

MA(1) model. For comparison, as before, we will consider also the full and pairwise maximum likelihood estimators (Davis and Yau, 2011). Moreover, we suppose that the parameter μ is known and equal to 0 and that the two parameters σ and α are unknown. We consider the moving average parameter $\alpha \in \{-0.9, -0.8, \ldots, 0.8, 0.9\}$. In the simulation study, 100 replicates of a single process of length $T = 400$ are generated. Results are summarised in Table 4. Moreover, Table 4 shows the Empirical Coverage Probability of the 95% confidence intervals based on the full and the pairwise likelihood methods, and the Hyvärinen scoring rule. The behaviour of the Empirical Coverage Probability is also summarised in Fig. 2. The simulation shows that the total Hyvärinen estimator $\widehat{\alpha}_{\mathrm{HT}}$ performs similarly to the MLE in the MA(1) model for values of the moving average parameter near 0, the first one tends to show a slightly higher standard deviation; see Table 4. The gap in terms of standard deviation increases when the absolute value of the moving average parameter approaches 1. When looking at the coverage probability, we notice, see Fig. 2 panel (a), that for both estimators the nominal value is never reached in the central part of the distribution of α, whereas, sometimes, when approaching the tails, the Hyvärinen estimator tends to outperform the maximum likelihood one. At this regard, it is important to recall that asymptotic properties of the estimators considered are derived to first order. In contrast, the pairwise likelihood estimator $\widehat{\alpha}_{\mathrm{PL}}$ shows very poor performances in terms of standard deviation, coverage probability and bias as $|\alpha|$ increases. The situation is different if we focus on the estimates of the variability parameter σ. In this case all the three estimators perform well when $\alpha \in [-0.5, 0.5]$. The scenario worsens in the tails. Considering the pairwise estimator, as shown in Fig. 2(b), we observe a dramatic decrease in terms of coverage probability. If we focus on our proposal we can note that the standard deviation of the variability parameter σ increases as the moving average parameter approaches the boundaries of the parameter space.

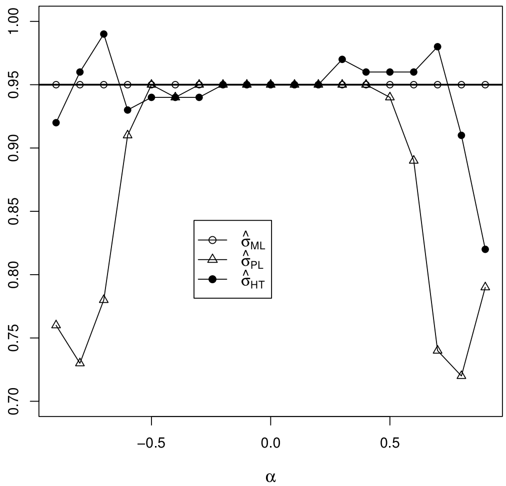
For both parameters, the Hyvärinen scoring rule overestimates the real coverage, on average, it produces longer confidence intervals, especially when $|\alpha|$ approaches 1. The confidence intervals based on the Hyvärinen scoring rule exhibit more reliable coverage than the confidence intervals obtained from the pairwise likelihood.

5.2. Real data example

In order to illustrate the behaviour of the Hyvärinen scoring rule in the context of the linear Gaussian time series models, we consider the well known Box & Jenkins `AirPassengers` time series dataset (Box et al., 1976) available on the `R` base package. The dataset concerns the number of international air travellers in the US between 1949 and 1960. This data set consists of $T = 144$ observations. The data are illustrated in Fig. 3(a): this figure suggests that there is a linear increasing trend of the series and a seasonal component of period 12: in fact, as is well known, there is an increase in the number of travellers during the summer periods. This series is clearly non-stationary, we therefore transform it to achieve stationarity. In order to remove both components of trend and seasonality, we consider a first and a seasonal differencing (see Fig. 3(b)). As suggested by the correlogram of the transformed series in Fig. 3(c), we estimate a moving average model of order 1. We fit the model by the full likelihood and the Hyvärinen scoring rule. The estimates of the parameters μ, α and σ based on the full likelihood and the Hyvärinen scoring rule are reported in Table 5. Moreover, 95% confidence intervals for μ, α and σ are reported in Table 6. Tables 5 and 6 reveal that the two methods perform similarly,



(a)



(b)

Fig. 2. Empirical Coverage Probabilities (ECP) of the 95% confidence intervals for the MA(1), single series, with $T = 400$ for different values of α with $\sigma = 1$. Panel (a) reports empirical coverages for the estimates of α, panel (b) coverages for the estimates of σ for various values of α.

Table 5
Estimates of the parameters (μ, α and σ) based on the full likelihood and the Hyvärinen scoring rule.

	μ	α	σ
Likelihood	0.1934	−0.3196	11.712
Hyvärinen	0.2126	−0.3426	11.806

Table 6
95% confidence intervals for the three parameters (μ, ϕ and σ) based on the full likelihood and the Hyvärinen scoring rule.

	μ	α	σ
Likelihood	(−1.18,1.56)	(−0.49,−0.15)	(10.29,13.13)
Hyvärinen	(−1.59,2.02)	(−0.58,−0.10)	(9.99,13.62)

although the confidence intervals obtained with the Hyvärinen scoring rule are wider than the one obtained with the likelihood method confirming the results of the simulation studies.

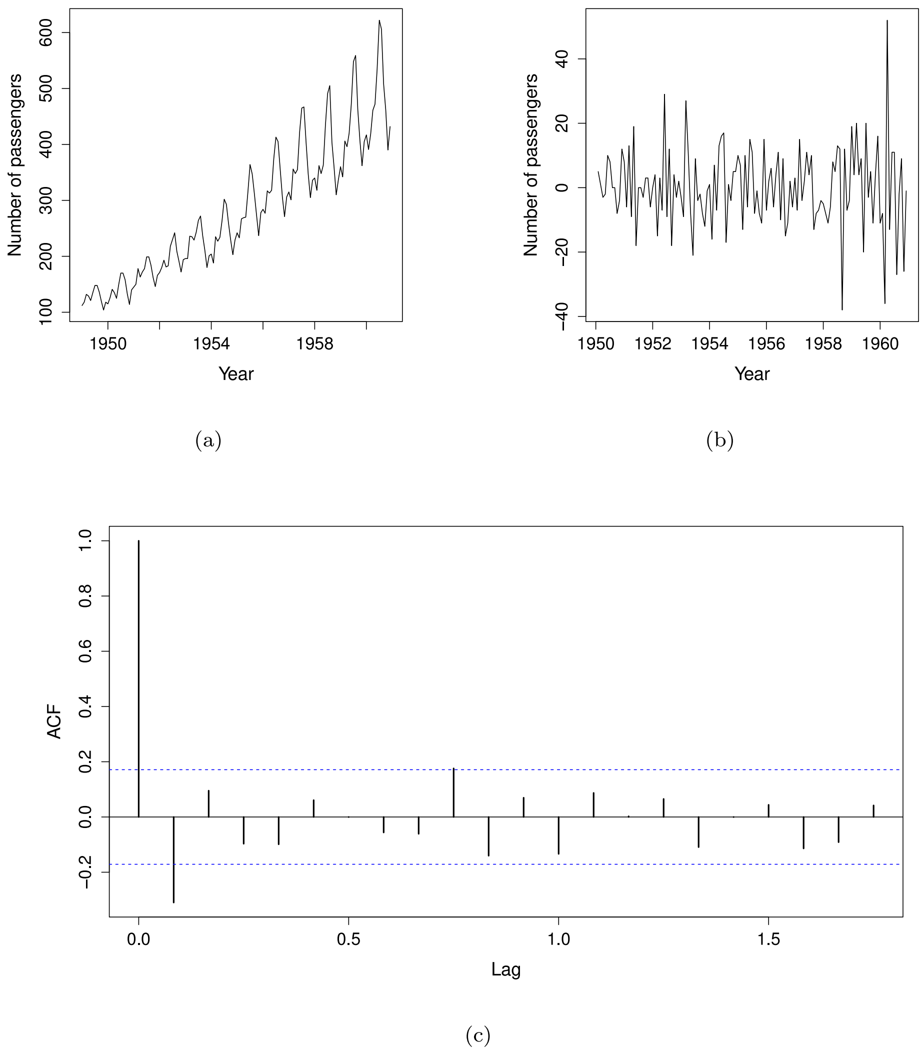


Fig. 3. `AirPassengers` time series description. Panel (a) represents the original time series, panel (b) the time series with first and seasonal differences, panel (c) reports the correlogram of the differenced `AirPassengers` time series.

6. Conclusions

In this paper we have considered the use of Hyvärinen scoring rules in linear time series estimation under different conditions. We have established the consistency of the Hyvärinen scoring rule estimator for a single times series under some general conditions and its asymptotic normality in an ARMA time series context.

We have investigated, for n independent time series, the performances of two estimators based on the Hyvärinen scoring rule, which can be regarded as a surrogate for a complex full likelihood. The properties of the estimators found using this scoring rule are compared with the full and pairwise maximum likelihood estimators. Three simple models are discussed: the first a stationary first order autoregressive model, the second a first order moving average model and the third a fractionally differenced white noise. In the first case the total Hyvärinen method leads to poor estimators; in contrast, in the second and third this method produces good estimators. The opposite behaviour is observed for the pairwise estimators. For the moving average process and the fractionally differenced white noise, there can be a large gain in efficiency, as compared to the pairwise likelihood method, by using the total or the matrix Hyvärinen scoring rule estimators. For the autoregressive model, in contrast, the total Hyvärinen score methods suffer a loss of efficiency as $|\phi|$ approaches 1.

The Hyvärinen and the pairwise estimators may work well for many time series models, but it is clear that the loss of efficiency incurred in using the Hyvärinen scoring rules or pairwise likelihood can be substantial. This depends on the underlying model (for both short-memory and long-memory), and no overall general principle has emerged that might offer guidance for cases not yet considered.

In all examples, results not reported here show that there is a great improvement in the performances of the matrix Hyvärinen estimator based on the Wishart model as the ratio T/n becomes negligible. The matrix Hyvärinen estimator has the apparent advantage over the other estimators (apart from full maximum likelihood) of being based on the sufficient statistic of the model; nevertheless the total Hyvärinen estimator shows good performance in terms of efficiency.

Although examples illustrated in Section 4 focus on simple linear time series models, they are classical examples of application of full and pairwise likelihood based methods in this framework (Davis and Yau, 2011), which highlight that the total Hyvärinen score may offer a viable and useful approach to estimation in linear time series models.

A promising future line of research appears to be the investigation of the Hyvärinen scoring rule for more complex models, where the evaluation of the exact full likelihood may be difficult or even impossible, entailing multidimensional integration of the full joint density for each value of the parameter, which is likely to occur for instance in spatial statistics and non linear time series frameworks.

CRediT authorship contribution statement

Silvia Columbu: Methodology, Software, Formal analysis, Writing. **Valentina Mameli:** Methodology, Software, Investigation, Writing. **Monica Musio:** Writing, Reviewing, Supervision. **Philip Dawid:** Conceptualization, Editing, Supervision.

Acknowledgements

Monica Musio was supported by the project STAGE of the Fondazione di Sardegna (grant number F74I19000890005) and Regione Autonoma di Sardegna. Valentina Mameli carried out some of this work while at Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Italy. The authors would like to thank the editor and two anonymous referees for useful comments that lead to an improved version of the paper.

Appendix

Proof of Theorem 3.1. Let $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\lambda})$ and let $\mathrm{E}_{\boldsymbol{\theta}}$ denote the expectation with respect to the probability distribution for (y_t) defined in Eq. (6). Let $\boldsymbol{\theta}_0 = (\sigma_0^2, \boldsymbol{\lambda_0})$ denote the true parameter value. From the ergodicity of (y_t), it follows that $H(Y_T, \boldsymbol{\theta})$ is ergodic and stationary and therefore

$$\frac{1}{T}H(Y_T, \boldsymbol{\theta}) \xrightarrow{a.s.} H(\boldsymbol{\theta}_0, \boldsymbol{\theta}) := \mathrm{E}_{\boldsymbol{\theta}_0} H(y_1, \boldsymbol{\theta}). \tag{18}$$

Since the Hyvärinen score is strictly proper we have

$$H(\boldsymbol{\theta}_0, \boldsymbol{\theta}) \geq H(\boldsymbol{\theta}_0, \boldsymbol{\theta}_0) \tag{19}$$

with equality if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, by the identifiability condition (8). The approach used to derive the consistency of the total Hyvärinen estimator now follows the same general argument used to derive the consistency of the pairwise likelihood estimator in Davis and Yau (2011).

In particular, the compactness of Λ and the inequality (19) are used as devices for proving the claim.

Proof of Theorem 3.2. Define the sample gradient and Hessian as

$$J_T(\boldsymbol{\theta}) := -\frac{1}{T}\sum_{i=1}^{T} \nabla_{\boldsymbol{\theta}}(b_{ii}) + \frac{1}{T}\sum_{i,j,t=1}^{T} b_{it} \nabla_{\boldsymbol{\theta}}(b_{ij}) y_j y_t$$

and

$$K_T(\boldsymbol{\theta}) := -\frac{1}{T}\sum_{i=1}^{T} \frac{\partial^2 b_{ii}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} + \frac{1}{T}\sum_{i,j,t=1}^{T} \frac{\partial b_{ij}}{\partial\boldsymbol{\theta}}\left(\frac{\partial b_{it}}{\partial\boldsymbol{\theta}}\right)^{\mathrm{T}} y_j y_t + \frac{1}{T}\sum_{i,j,t=1}^{T} \frac{\partial^2 b_{ij}}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} b_{it} y_j y_t.$$

Using a Taylor expansion of $J_T(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$ and the consistency of Hyvärinen scoring rule estimator, it can be proved that, for some $\boldsymbol{\theta}_T^+$ between $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_T$,

$$J_T(\boldsymbol{\theta}_0) = K_T(\boldsymbol{\theta}_T^+)\sqrt{T}(\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\theta}}_T). \tag{20}$$

The asymptotic distribution of $\widehat{\boldsymbol{\theta}}_T$ can be derived by exploiting the asymptotic properties of $K_T(\boldsymbol{\theta}_T^+)$ and $J_T(\boldsymbol{\theta}_0)$, together with the fact that $\boldsymbol{\theta}_T^+ \xrightarrow{a.s.} \boldsymbol{\theta}_0$.

Writing

$$\frac{\partial}{\partial \boldsymbol{\theta}_0} = \left.\frac{\partial}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$
$$\Gamma = \Gamma(\boldsymbol{\lambda}_0)$$

it can be shown that

$$\mathrm{E}_{\boldsymbol{\theta}_0}(K_T) = \frac{1}{T}\sum_{i,j,t=1}^{T} \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0}\left(\frac{\partial b_{it}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \sigma_0^2 \Gamma_{jt} \xrightarrow{T\to\infty} M(\boldsymbol{\theta}_0). \tag{21}$$

The expectation in (21) can be rewritten as

$$\mathrm{E}_{\boldsymbol{\theta}_0}(K_T) = \frac{1}{T}\sum_{i,j,t=1}^{T} \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0}\left(\frac{\partial b_{it}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \sigma_0^2 \Gamma_{jt} = \frac{1}{T}\sum_{i,j,t=1}^{T} \frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(i-j)}{\sigma_0^2}\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(i-t)}{\sigma_0^2}\right)^{\mathrm{T}} \sigma_0^2 \gamma(j-t),$$

where $\gamma(j-t) = \Gamma_{jt}$ and $\gamma^{-1}(i-j) = \Gamma^{ij}$. Let $k = i-j$ and $r = j-t$. Without lose of generality, we assume that $\gamma(h) = 0$ if $|h| > T-1$. Then the previous expression and consequently the first term in (24) simplifies to

$$\frac{1}{T}\sum_{k,r=-T}^{T} (T - \max\{|k|, |k+r|, |r|\}) \frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k)}{\sigma_0^2}\left(\frac{\partial \gamma^{-1}(k+r)}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \gamma(r). \tag{22}$$

The absolute summability of the auto-covariance and the duality properties of autocorrelation and of its inverse for causal invertible autoregressive-moving average processes (see Cleveland (1972), Chatfield (1979) and Hosking (1980)) guarantee the following holds:

$$\begin{aligned}
&\lim_{T\to\infty}\sum_{r,k=-T}^{T} \frac{(T - \max\{|k|, |k+r|, |r|\})}{T} \\
&\quad \times \frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k)}{\sigma_0^2}\left(\frac{\partial \gamma^{-1}(k+r)}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \gamma(r) \\
&= \sum_{r,k=-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k)}{\sigma_0^2}\left(\frac{\partial \gamma^{-1}(k+r)}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \gamma(r) \\
&= \sum_{r,k=-\infty}^{\infty} M(r, k, \boldsymbol{\theta}_0) = M(\boldsymbol{\theta}_0).
\end{aligned} \tag{23}$$

In order to calculate the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_T$ we need to calculate the expectation and the variance of $J_T(\boldsymbol{\theta}_0)$. The calculation of the expectation of $J_T(\boldsymbol{\theta}_0)$ follows easily from the unbiasedness of the scoring rule estimating equation (Dawid and Lauritzen, 2005). However, calculation of the variance of $J_T(\boldsymbol{\theta}_0)$ is challenging due to the presence of the non deterministic term

$$B_i = \sum_{j,t=1}^{T} \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0} b_{it} y_j y_t.$$

It relies on the following calculation:

$$\begin{aligned}
\mathrm{var}(J_T(\boldsymbol{\theta}_0)) &= \frac{1}{T}\sum_{i=1}^{T} \mathrm{var}(B_i) \\
&= \frac{1}{T}\sum_{i,j,t,\ell,h=1}^{T} \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0} b_{it}\left(\frac{\partial b_{i\ell}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} b_{ih}\mathrm{cov}(y_j y_t, y_\ell y_h) \\
&= \frac{1}{T}\sum_{i,j,t,\ell,h=1}^{T} \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0}\frac{\Gamma^{it}}{\sigma_0^2}\left(\frac{\partial b_{i\ell}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}} \frac{\Gamma^{ih}}{\sigma_0^2}\{\mathrm{cov}(y_j, y_\ell)\mathrm{cov}(y_t, y_h) \\
&\quad + \mathrm{cov}(y_j, y_h)\mathrm{cov}(y_t, y_\ell) + \mathrm{cum}_4(y_j, y_t, y_\ell, y_h)\} \\
&= \frac{1}{T}\sum_{i,j,t,\ell,h=1}^{T} A_{j\ell th} + C_{jht\ell} + D_{it\ell h},
\end{aligned}$$

where

$$
\begin{aligned}
A_{j\ell th} &= \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0}\frac{\Gamma^{it}}{\sigma_0^2}\left(\frac{\partial b_{i\ell}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}}\frac{\Gamma^{ih}}{\sigma_0^2}\mathrm{cov}(y_j, y_\ell)\mathrm{cov}(y_t, y_h)\\
C_{jht\ell} &= \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0}\frac{\Gamma^{it}}{\sigma_0^2}\left(\frac{\partial b_{i\ell}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}}\frac{\Gamma^{ih}}{\sigma_0^2}\mathrm{cov}(y_j, y_h)\mathrm{cov}(y_t, y_\ell)\\
D_{it\ell h} &= \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0}\frac{\Gamma^{it}}{\sigma_0^2}\left(\frac{\partial b_{i\ell}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}}\frac{\Gamma^{ih}}{\sigma_0^2}\mathrm{cum}_4(y_j, y_t, y_\ell, y_h).
\end{aligned}
$$

The first term in (24) simplifies as

$$
\sum_{i,j,t,\ell,h=1}^{T} A_{j\ell th} = \sum_{i,j,t,\ell,h=1}^{T} \frac{\partial b_{ij}}{\partial \boldsymbol{\theta}_0}\Gamma^{ii}\left(\frac{\partial b_{i\ell}}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}}\Gamma_{j\ell}.
$$

The second term simplifies as

$$
\sum_{i,j,t,\ell,h=1}^{T} C_{jht\ell} = T\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(0)}{\sigma_0^2}\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(0)}{\sigma_0^2}\right)^{\mathrm{T}}.
$$

The third term in (24), which involves the fourth cumulant, vanishes as for Gaussian linear processes all the cumulant functions cum_k for $k > 3$ are identically null Brockwell and Davis (1991). Hence convergence of $\mathrm{var}(J_T(\boldsymbol{\theta}_0))$ is evaluated by considering only the first non-constant term (25).

Eq. (25) can be rewritten as

$$
\sum_{i,j,t,\ell,h=1}^{T} A_{j\ell th} = \sum_{i,j,\ell=1}^{T} \frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(i-j)}{\sigma_0^2}\gamma^{-1}(0)\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(i-\ell)}{\sigma_0^2}\right)^{\mathrm{T}}\gamma(j-\ell),
$$

applying the same substitutions and conditions used in (22) we obtain

$$
\sum_{k,r=-T}^{T} (T - \max\{|k|, |k+r|, |r|\})\,\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k)}{\sigma_0^2}\gamma^{-1}(0)\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k+r)}{\sigma_0^2}\right)^{\mathrm{T}}\gamma(r).
$$

Taking limits we have then

$$
\begin{aligned}
&\lim_{T\to\infty}\sum_{r,k=-T}^{T}\frac{(T - \max\{|k|, |k+r|, |r|\})}{T}\\
&\quad\times \frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k)}{\sigma_0^2}\gamma^{-1}(0)\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k+r)}{\sigma_0^2}\right)^{\mathrm{T}}\gamma(r)\\
&= \sum_{r,k=-\infty}^{\infty}\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k)}{\sigma_0^2}\gamma^{-1}(0)\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k+r)}{\sigma_0^2}\right)^{\mathrm{T}}\gamma(r).
\end{aligned}
$$

Combining Eqs. (26) and (27) we obtain

$$
V(\boldsymbol{\theta}_0) = \left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(0)}{\sigma_0^2}\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(0)}{\sigma_0^2}\right)^{\mathrm{T}} + \sum_{r,k=-\infty}^{\infty}\frac{\partial}{\partial \boldsymbol{\theta}_0}\frac{\gamma^{-1}(k)}{\sigma_0^2}\gamma^{-1}(0)\left(\frac{\partial \gamma^{-1}(k+r)}{\partial \boldsymbol{\theta}_0}\right)^{\mathrm{T}}\gamma(r),
$$

and then

$$
\mathrm{var}(J_T(\boldsymbol{\theta}_0)) \longrightarrow V(\boldsymbol{\theta}_0).
$$

Since $J(\boldsymbol{\theta}_0)$ depends on the B_i's, which involve the sample auto-covariance, it follows from the asymptotic normality of the sample auto-covariance of ARMA processes that $J_T(\boldsymbol{\theta}_0)$ is also asymptotically normal with zero mean and variance V. From (20) and (28) we obtain the asymptotic normality of $\widehat{\boldsymbol{\theta}}_T$:

$$
\sqrt{T}(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N_{m-1}\left(0, M(\boldsymbol{\theta}_0)^{-1}V(\boldsymbol{\theta}_0)M^{\top}(\boldsymbol{\theta}_0)^{-1}\right).
$$

References

Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. Monographs on Statistics and Applied Probability, vol. 52, Chapman and Hall, London.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., 1976. Time Series Analysis, Forecasting and Control, third ed. Holden-Day.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods. Springer, New York.

Cattelan, M., Sartori, N., 2016. Empirical and simulated adjustments of composite likelihood ratio statistics. J. Stat. Comput. Simul. 86, 1056–1067.

Chatfield, C., 1979. Inverse autocorrelations. J. R. Stat. Soc. A 142, 363–377.

Cleveland, W.S., 1972. The inverse autocorrelations of a time series and their applications. Technometrics 14, 277–293.

Davis, R.A., Yau, C.Y., 2011. Comments on pairwise likelihood in time series models. Statist. Sinica 21, 255–277.

Dawid, A.P., 2007. The geometry of proper scoring rules. Ann. Inst. Statist. Math. 59, 77–93.

Dawid, A.P., Lauritzen, S.L., 2005. The geometry of decision theory. In: Proceedings of the Second International Symposium on Information Geometry and Its Applications. University of Tokyo, pp. 22–28.

Dawid, A.P., Lauritzen, S., Parry, M., 2012. Proper local scoring rules on discrete sample spaces. Ann. Statist. 40 (1), 593–608.

Dawid, A.P., Musio, M., 2014. Theory and applications of proper scoring rules. Metron 72, 169–183.

Dawid, A.P., Musio, M., 2015. Bayesian model selection based on proper scoring rules. Bayesian Anal. 10 (2), 479–499.

Dawid, A., Musio, M., Columbu, S., 2017. A note on Bayesian model selection for discrete data using proper scoring rules. Statist. Probab. Lett. 129, 101–106.

Dawid, A.P., Musio, M., Ventura, L., 2016. Minimum scoring rule inference. Scand. J. Stat. 43, 123–138.

Ehm, W., Gneiting, T., 2012. Local proper scoring rules of order two. Ann. Statist. 40 (1), 609–637.

Gilbert, P., Varadhan, R., 2012. Numderiv: Accurate numerical derivatives. URL: http://CRAN.R-project.org/package=numDeriv.

Good, I.J., 1952. Rational decisions. J. R. Stat. Soc. Ser. B Stat. Methodol. 14 (1), 107–114.

Hosking, J.R.M., 1980. The asymptotic distribution of the sample inverse autocorrelations of an autoregressive-moving average process. Biometrika 67, 223–226.

Hosking, J.R.M., 1981. Fractional differencing. Biometrika 68 (1), 165–172.

Hyvärinen, A., 2005. Estimation of non-normalized statistical models by score matching. J. Mach. Learn. Res. 6, 695–709.

Joe, H., Lee, Y., 2009. On weighting of bivariate margins in pairwise likelihood. J. Multivariate Anal. 100 (4), 670–685.

Mardia, K., Kent, J., Laha, A., 2016. Score matching estimators for directional distributions. arXiv:1604.08470.

Pace, L., Salvan, A., 1997. Principles of Statistical Inference from a Neo-Fisherian Perspective. World Scientific, Singapore.

Parry, M., Dawid, A.P., Lauritzen, S.L., 2012. Proper local scoring rules. Ann. Statist. 40 (1), 561–592.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Shao, S., Jacob, P.E., Ding, J., Tarokh, V., 2019. Bayesian model comparison with the Hyvärinen score: Computation and consistency. J. Amer. Statist. Assoc. 114 (528), 1826–1837.

Takasu, Y., Yano, K., Komaki, F., 2018. Scoring rules for statistical models on spheres. Statist. Probab. Lett. 138, 111–115.

Varin, C., 2008. On composite marginal likelihoods. AStA Adv. Stat. Anal. 92 (1), 1–28.

Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. Statist. Sinica 21, 5–42.

Von Rosen, D., 1997. On moments of the inverted wishart distribution. Statistics 30 (3), 259–278.