ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2017)
79, *Part* 2, *pp.* 507–524

Oracally efficient estimation and consistent model selection for auto-regressive moving average time series with trend

Qin Shao

Soochow University, Suzhou, People's Republic of China, and University of Toledo, USA

and Lijian Yang

Tsinghua University, Beijing, People's Republic of China

[Received August 2014. Final revision December 2015]

Summary. Most time series that are encountered in practice contain non-zero trend, yet textbook approaches to time series analysis are typically focused on zero-mean stationary auto-regressive moving average (ARMA) processes. Trend is often estimated by *ad hoc* methods and subtracted from time series, and the residuals are used as the true ARMA noise for data analysis and inference, including parameter estimation, lag selection and prediction. We propose a theoretically justified two-step method to analyse time series consisting of a smooth trend function and ARMA error term, which is computationally efficient and easy for practitioners to implement. The trend is estimated by *B*-spline regression, and the maximum likelihood estimator based on residuals is shown to be oracally efficient in the sense that it is asymptotically as efficient as if the true trend function were known and then removed to obtain the ARMA errors. In addition, consistency of the Bayesian information criterion for model selection is established for the detrended residual sequence. Finite sample performance of the procedure is illustrated by simulation studies and real data analysis.

Keywords: Auto-regressive moving average; Bayesian information criterion; *B*-splines; Maximum likelihood estimator; Mixing; Oracle efficiency

1. Introduction

Inference for stationary auto-regressive moving average (ARMA) processes $\{x_t\}_{t=-\infty}^{\infty}$ is critical to understanding the dynamical process, and hence it has received extensive attention; see for example Brockwell and Davis (1991) and Fan and Yao (2003). Theoretical results on inference of model coefficients and identification of significant lags have been obtained assuming that an ARMA time series $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ is actually observed. However, in practice time series are often observed with trends that need to be removed before elaborating the analysis. Our method is motivated by the demand of analysing real data and the approach recommended by for example Brockwell and Davis (1991) and Fan and Yao (2003). It is intuitive and straightforward, and can be applied to time series that contain a smooth trend and ARMA error term. Essentially, it accomplishes the analysis in two steps: in the first step, the trend is estimated and the residual sequence $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n)^{\mathrm{T}}$ is calculated; in the second step, the inference about ARMA time

Address for correspondence: Lijian Yang, Center for Statistical Science and Department of Industrial Engineering, Tsinghua University, Beijing 10084, People's Republic of China.
E-mail: yanglijian@tsinghua.edu.cn

 1369–7412/17/79507

series, which includes lag selection and parameter estimation, is carried out by using the residuals $\hat{\mathbf{x}}$ in place of $\mathbf{x}$. This two-step model fitting procedure is appealing, as all well-known results for ARMA time series can be applied to the residual sequence if the substitution is appropriate.

The time series $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ that is considered here can be described by the model

$$\mathbf{y} = \mathbf{g} + \mathbf{x},$$

where $\mathbf{g} = (g(\omega_1), \ldots, g(\omega_n))^{\mathrm{T}}$ with $\omega_t = t/n$ and $g(\cdot)$ being a smooth trend function, and the error term $\mathbf{x}$ is an ARMA time series with AR order p and MA order q (ARMA(p, q)) satisfying

$$x_t - \sum_{k=1}^{p} \phi_k x_{t-k} = \epsilon_t + \sum_{k=1}^{q} \theta_k \epsilon_{t-k}, \tag{1}$$

where the white noise $\{\epsilon_t\}_{t=-\infty}^{\infty}$ is independent and identically distributed (IID) with mean 0 and variance σ^2 ($\epsilon_t \sim \mathrm{IID}(0, \sigma^2)$). The most important parameter in model (1) consists of ARMA coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{p+q})^{\mathrm{T}}$ with $\alpha_k = \phi_k$ for $1 \leqslant k \leqslant p$ and $\alpha_k = \theta_{k-p}$ for $p+1 \leqslant k \leqslant p+q$.

Denote by $\hat{g}$ an estimator for the trend function g, $\hat{\mathbf{x}}$ the residuals that are obtained by subtracting the trend estimate $\hat{\mathbf{g}} = (\hat{g}(\omega_1), \ldots, \hat{g}(\omega_n))^{\mathrm{T}}$ from the observed $\mathbf{y}$,

$$\hat{x}_t = y_t - \hat{g}(\omega_t), \tag{2}$$

and let $\tilde{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\alpha}}$ be two maximum likelihood estimators (MLEs) of $\boldsymbol{\alpha}$ based on $\mathbf{x}$ and $\hat{\mathbf{x}}$ respectively. The primary goal of the present paper is to establish oracle efficiency of the MLE $\hat{\boldsymbol{\alpha}}$ obtained from detrended residuals. If the substitution of $\hat{\mathbf{x}}$ for $\mathbf{x}$ is appropriate so that its effect on the estimation of the model parameters is negligible, then $\hat{\boldsymbol{\alpha}}$ should have the same consistency and asymptotically normal distribution as $\tilde{\boldsymbol{\alpha}}$. This property of the estimator $\hat{\boldsymbol{\alpha}}$ is called *oracle efficiency*, i.e. $\hat{\boldsymbol{\alpha}}$ has the same asymptotic behaviour as $\tilde{\boldsymbol{\alpha}}$.

Oracle efficiency is of utmost importance, as it ensures the legitimacy of substituting the residuals for the unobserved noise variables. Although some results have been obtained when the trend is estimated parametrically (see for example Tsay and Tiao (1984) and Garel and Hallin (1995)), little has been done for non-parametric estimation of the trend. In contrast, non-parametric trend estimation does not need to specify the analytical function form, and thus it is much more flexible and reduces the error from model misspecification. Consider, for instance, the ARMA(1,1) examples in Section 4. Misspecifying a non-linear trend as a linear function produces extremely large bias in the estimation of some ARMA(1,1) model coefficients, whereas B-spline and kernel trend estimates result in consistent estimates, as shown in Table 1 in Section 4. Truong (1991) was a pioneering work on oracle efficiency of Yule–Walker estimators for AR time series based on residuals, with trend estimated by a kernel method. Shao and Yang (2011) and Qiu *et al.* (2013) respectively extended Truong's work to non-Gaussian time series, using B-spline and local linear detrending methods. These previous oracle efficiency results are extended to general ARMA time series, by using MLEs instead of Yule–Walker estimators, with significantly more challenging proofs. Although spline and kernel smoothing have been our main trending tools, we might expect that other approaches, such as the unbalanced Haar wavelet as in Schroeder and Fryzlewicz (2013), would work equally well or possibly better.

Model selection is an important aspect of model fitting. Box *et al.* (1994) suggested that one starts with candidate models and finds the most parsimonious and accurate model by using model selection criteria for the given data. On the basis of the oracle efficiency of the proposed MLEs we can extend the Bayesian information criterion (BIC), which is one of the commonly used model selection criteria, to the residual sequence and choose the true model with probability 1 as the sample size increases to ∞.

A second objective of this paper is to provide practitioners with a data-driven and easily applicable detrending procedure which is also an asymptotically efficient estimation of ARMA coefficients. We recommend *B*-spline smoothing especially for large data because of its computational advantage, which is to say that only a single optimization is needed for the unknown function over an entire range, rather than optimization at every point in a kernel smoothing method. Indeed, *B*-spline smoothing is computationally much more efficient than kernel smoothing; see Xue and Yang (2006) and Wang and Yang (2007) for detailed discussions. It is worth emphasizing that the method proposed is computationally accessible to practitioners. One can utilize any software package that has a built-in function for ARMA models, although we carry out the calculations by using the `arima` package in R which is an open-access environment for statistical computing and graphics developed by the R Core Team (2013).

The paper is organized as follows. Section 2 introduces *B*-splines and describes the actual steps to compute maximum likelihood estimates for ARMA coefficients from observations. Section 3 presents oracle efficiency and other asymptotic properties of $\hat{\boldsymbol{\alpha}}$, as well as consistency of the BIC for model selection. Section 4 illustrates, by means of simulation studies, the performance of $\hat{\boldsymbol{\alpha}}$ and the BIC with several MA(1) and ARMA(1,1) time series, and it applies the proposed procedure to the analysis of annual normalized tree ring widths between 1131 and 1975. Concluding remarks are given in Section 5, and all technical proofs are in Appendix A and a separate on-line supplement.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Construction of estimators

2.1. Estimator of auto-regressive moving average coefficients

To be precise, denote the true parameters of the ARMA process (1) by σ_0^2 and $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0,p+q})^{\mathrm{T}}$ with $\alpha_{0k} = \phi_{0k}$ for $1 \leqslant k \leqslant p$ and $\alpha_{0k} = \theta_{0,k-p}$ for $p+1 \leqslant k \leqslant p+q$. We rewrite model (1) as

$$\Phi(\boldsymbol{\alpha}_0, B) x_t = \Theta(\boldsymbol{\alpha}_0, B)\epsilon_t, \tag{3}$$

where $\Phi(\boldsymbol{\alpha}_0, B) = 1 - \Sigma_{k=1}^{p} \phi_{0k} B^k$ and $\Theta(\boldsymbol{\alpha}_0, B) = 1 + \Sigma_{k=1}^{q} \theta_{0k} B^k$ with B being the backward shift operator defined by $B^k x_t = x_{t-k}$. Throughout the paper, we assume that the time series $\{x_t\}_{t=-\infty}^{\infty}$ is causal and invertible; hence there are sequences of constants $\{\psi_{0j}\}_{j=0}^{\infty}$ and $\{\pi_{0j}\}_{j=0}^{\infty}$ such that $\Sigma_{j=0}^{\infty} |\psi_{0j}| < \infty$, $\Sigma_{j=0}^{\infty} |\pi_{0j}| < \infty$ and

$$x_t = \sum_{j=0}^{\infty} \psi_{0j} \epsilon_{t-j}, \qquad \epsilon_t = \sum_{j=0}^{\infty} \pi_{0j} x_{t-j}.$$

According to equation (3.1.19) in theorem 3.1.2 of Brockwell and Davis (1991), the constants $\{\pi_{0j}\}_{j=0}^{\infty}$ satisfy

$$\sum_{j=0}^{\infty} \pi_{0j} z^j = \Phi(\boldsymbol{\alpha}_0, z)/\Theta(\boldsymbol{\alpha}_0, z), \qquad |z| \leqslant 1. \tag{4}$$

If $\{\epsilon_t\}_{t=1}^{n}$ were available and the white noise normally distributed, the MLE would be calculated by minimizing $n \log(\sigma^2) + \Sigma_{t=1}^{n} \epsilon_t^2/\sigma^2$, which is proportional to the log-likelihood function $l(\boldsymbol{\alpha}, \sigma^2; \mathbf{x})$. Therefore we define the objective function $Q_n(\boldsymbol{\alpha}; \mathbf{x})$ as

$$Q_n(\boldsymbol{\alpha};\mathbf{x}) = n^{-1}\sum_{t=p+1}^{n}\left(\sum_{j=0}^{t-1}\pi_j x_{t-j}\right)^2$$

in which the $\{\pi_j\}_{j=0}^{\infty}$ are rational functions of candidate value $\boldsymbol{\alpha}$ defined by mimicking condition (4):

$$\sum_{j=0}^{\infty}\pi_j z^j = \sum_{j=0}^{\infty}\pi_j(\boldsymbol{\alpha})z^j = \Phi(\boldsymbol{\alpha},z)/\Theta(\boldsymbol{\alpha},z). \tag{5}$$

Minimizing $Q_n(\boldsymbol{\alpha};\mathbf{x})$ produces the following estimators $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\sigma}^2$:

$$\tilde{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} Q_n(\boldsymbol{\alpha};\mathbf{x});$$
$$\tilde{\sigma}^2 = Q_n(\tilde{\boldsymbol{\alpha}};\mathbf{x}).$$

Under certain conditions, these estimators are equivalent to the MLE that is obtained from $l(\boldsymbol{\alpha},\sigma^2;\mathbf{x})$ in the sense that they have the same asymptotic properties according to Pierce (1971). Hereafter, we call $(\tilde{\boldsymbol{\alpha}}^{\mathrm{T}},\tilde{\sigma}^2)$ the 'infeasible' MLE of $(\boldsymbol{\alpha}_0^{\mathrm{T}},\sigma_0^2)$, as it relies on unobservable sequence $\{x_t\}_{t=1}^n$, when only $\{y_t\}_{t=1}^n$ is observed. We propose to estimate $\boldsymbol{\alpha}_0$ by replacing $\{x_t\}_{t=1}^n$ with the residuals $\{\hat{x}_t\}_{t=1}^n$ calculated by equation (2), i.e.

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}}),$$
$$\hat{\sigma}^2 = Q_n(\hat{\boldsymbol{\alpha}};\hat{\mathbf{x}}), \tag{6}$$

where $Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}}) = n^{-1}\Sigma_{t=p+1}^{n}(\Sigma_{j=0}^{t-1}\pi_j\hat{x}_{t-j})^2$. We call $(\hat{\boldsymbol{\alpha}}^{\mathrm{T}},\hat{\sigma}^2)$ the MLE of $(\boldsymbol{\alpha}_0^{\mathrm{T}},\sigma_0^2)$ hereafter. In the next section, we shall show that $\hat{\boldsymbol{\alpha}}$ is oracally efficient, i.e. it is asymptotically indistinguishable from $\tilde{\boldsymbol{\alpha}}$, whereas $\hat{\sigma}^2$ is as consistent as $\tilde{\sigma}^2$.

This residual replacement approach can be applied to other objective functions which provide asymptotically equivalent estimators to the infeasible MLEs, e.g. the prediction-based algorithms discussed in Brockwell and Davis (1991). We summarize the estimation procedure for the ARMA coefficient $\boldsymbol{\alpha}_0$ as follows.

Step 1: estimate the trend function $g(\omega)$ and calculate the residuals $\{\hat{x}_t\}_{t=1}^n$ by equation (2).
Step 2: find the MLEs $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$ by expression (6).

2.2. Trend estimator

We now describe one convenient estimator of the trend. Denote a sequence of interior knots $\tau_1,\ldots,\tau_N, 0<\tau_1<\ldots<\tau_N<1$, which divide the interval $[0,1]$ into subintervals of equal length $h=(N+1)^{-1}$, $J_j=[jh,(j+1)h)$, $j=0,\ldots,N-1$, and $J_N=[Nh,1]$. For any given $\omega\in[0,1]$, define $j(\omega)=[\omega/h]$ where $[a]$ denotes the integer part of a real number a, so that $\omega\in J_{j(\omega)}$. For an integer $m>0$, let $G_N^{(m-2)}=G_N^{(m-2)}[0,1]$ be the space of functions that are polynomial of degree $m-1$ on each J_j and have continuous $(m-2)$th derivative, and denote its B-spline basis as $\mathbf{b}_m(\omega)=(b_{-m+1,m}(\omega),\ldots,b_{N,m}(\omega))^{\mathrm{T}}$; see chapter IX of de Boor (2001). In particular, for $m=1$, the B-spline basis for the constant spline space $G_N^{(-1)}$ is $\mathbf{b}_1(\omega)=(b_{0,1}(\omega),\ldots,b_{N,1}(\omega))^{\mathrm{T}}$, where $b_{j,1}(\omega)$ is the indicator function of J_j, i.e.

$$b_{j,1}(\omega)=\begin{cases}1, & j=j(\omega),\\ 0, & \text{otherwise},\end{cases}$$

whereas, for $m=2$, the B-spline basis for the piecewise linear spline space $G_N^{(0)}$ is $\mathbf{b}_2(\omega)=(b_{-1,2}(\omega),\ldots,b_{N,2}(\omega))^{\mathrm{T}}$, where $b_{j,2}(\omega)$ is defined as

$$b_{j,2}(\omega)=\begin{cases} j(\omega)+1-\omega/h, & j=j(\omega)-1,\\ \omega/h-j(\omega), & j=j(\omega),\\ 0, & \text{otherwise;}\end{cases}$$

see Shao and Yang (2011).

Denote for any function $\varphi(\cdot)$ in $L^2[0,1]$ the norm as $\|\varphi\|_2=\{\int_0^1\varphi^2(x)\,\mathrm{d}x\}^{1/2}$. For any $\omega\in[0,1]$, the standardized B-spline basis $\mathbf{c}_m(\omega)$ is defined as

$$\mathbf{c}_m(\omega)=(c_{-m+1,m}(\omega),\ldots,c_{N,m}(\omega))^{\mathrm{T}},\qquad c_{j,m}(\omega)=\frac{b_{j,m}(\omega)}{\|b_{j,m}\|_2}=\frac{b_{j,m}(\omega)}{\left\{\int_0^1 b_{j,m}^2(\omega)\,\mathrm{d}\omega\right\}^{1/2}}. \tag{7}$$

The design matrix $\mathbf{C}_m$ is defined as

$$\mathbf{C}_m=(\mathbf{c}_{-m+1,m},\ldots,\mathbf{c}_{N,m}),\qquad \mathbf{c}_{j,m}=(c_{j,m}(\omega_1),\ldots,c_{j,m}(\omega_n))^{\mathrm{T}},\quad -m+1\leqslant j\leqslant N. \tag{8}$$

The polynomial spline estimator of $g(\omega)$ based on a realization of time series $\mathbf{y}$ is

$$\hat{g}_m(\omega)=\sum_{j=1-m}^{N}c_{j,m}(\omega)\hat{\beta}_{j,m},\qquad \omega\in[0,1], \tag{9}$$

where $c_{j,m}(\omega)$ is defined in equation (7) and $\hat{\boldsymbol{\beta}}_m=(\hat{\beta}_{-m+1,m},\ldots,\hat{\beta}_{N,m})^{\mathrm{T}}$ is obtained by

$$\hat{\boldsymbol{\beta}}_m=\arg\min_{\boldsymbol{\beta}}\{(\mathbf{y}-\mathbf{C}_m\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{C}_m\boldsymbol{\beta})\}. \tag{10}$$

According to linear model theory, equations (9) and (10) imply

$$\hat{g}_m(\omega)=\mathbf{c}_m^{\mathrm{T}}(\omega)\left(\frac{1}{n}\mathbf{C}_m^{\mathrm{T}}\mathbf{C}_m\right)^{-1}\frac{1}{n}\mathbf{C}_m^{\mathrm{T}}\mathbf{y}. \tag{11}$$

3. Main results

We begin by listing the assumptions that are needed for the main theorems 1, 2 and 3. The only requirement on the trend estimator $\hat{g}(\omega)$ is the rather generic assumption (c).

(a) The parameter space Ξ is compact and consists of $\boldsymbol{\alpha}$ such that all roots of $\Phi(\boldsymbol{\alpha},z)=0$ and $\Theta(\boldsymbol{\alpha},z)=0$ are larger than 1 in absolute value, and they have no common roots. The true parameter value $\boldsymbol{\alpha}_0$ is in the interior of the parameter space Ξ.

(b) The innovations $\{\epsilon_t\}_{t=-\infty}^{\infty}$ are IID with $E(\epsilon_1^4)<\infty$.

(c) The trend estimator $\hat{g}(\omega)$ satisfies the constraints

$$\max_{1\leqslant t\leqslant n}E\{g(\omega_t)-\hat{g}(\omega_t)\}^2=o(n^{-1/2}), \tag{12}$$

$$\max_{1\leqslant k\leqslant p+q}n^{-1}\left|\sum_{t=p+1}^{n}\left(\sum_{j=0}^{t-1}\frac{\partial\pi_{0j}}{\partial\alpha_k}x_{t-j}\right)\sum_{j=0}^{t-1}\pi_{0j}\{\hat{g}(\omega_{t-j})-g(\omega_{t-j})\}\right|=o_p(n^{-1/2}), \tag{13}$$

$$\max_{1\leqslant k\leqslant p+q}n^{-1}\left|\sum_{t=p+1}^{n}\left(\sum_{j=0}^{t-1}\pi_{0j}x_{t-j}\right)\sum_{j=0}^{t-1}\frac{\partial\pi_{0j}}{\partial\alpha_k}\{\hat{g}(\omega_{t-j})-g(\omega_{t-j})\}\right|=o_p(n^{-1/2}), \tag{14}$$

in which $\partial\pi_{0j}/\partial\alpha_k=\partial\pi_j/\partial\alpha_k|_{\boldsymbol{\alpha}=\boldsymbol{\alpha}_0}$.

Define $\Pi_j=\sup_{\boldsymbol{\alpha}\in\Xi}|\pi_j(\boldsymbol{\alpha})|$, $0\leqslant j<\infty$. Under assumption (a), π_j defined in equation (5) satisfies equation (3.3.6) of Brockwell and Davis (1991), which entails that there are constants

C_π and $0<\rho_\pi<1$ such that

$$\Pi_j \leqslant C_\pi \rho_\pi^j, \qquad 0 \leqslant j < \infty. \tag{15}$$

For instance, for an ARMA(1,1) model $|\pi_j(\phi_1,\theta_1)| \leqslant |\theta_1^{j-1}|(|\phi_1|+|\theta_1|)$, which satisfies condition (15).

The following result indicates that the MLEs $\hat{\boldsymbol{\alpha}}$ and $\hat{\sigma}^2$ from equation (6) are consistent, i.e. they converge to their true values in probability.

Theorem 1. Under assumptions (a)–(c), as $n\to\infty$, $\hat{\boldsymbol{\alpha}}\to^{\mathrm{P}}\boldsymbol{\alpha}_0$ and $\hat{\sigma}^2\to^{\mathrm{P}}\sigma_0^2$.

Denote two AR processes $\{u_t\}_{t=-\infty}^{\infty}$ and $\{v_t\}_{t=-\infty}^{\infty}$ by $\Phi(\boldsymbol{\alpha}_0,B)u_t=\epsilon_t$ and $\Theta(\boldsymbol{\alpha}_0,B)v_t=\epsilon_t$ respectively. Under assumptions (a) and (b), the infeasible MLE $\tilde{\boldsymbol{\alpha}}$ is asymptotically distributed as a $(p+q)$-dimensional multivariate normal distribution i.e., as $n\to\infty$,

$$\sqrt{n}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0) \overset{\mathrm{D}}{\Rightarrow} N_{p+q}(\mathbf{0},\mathbf{V}),$$

where $\mathbf{V}$ is the $(p+q)\times(p+q)$ covariance matrix defined as

$$\mathbf{V}=\sigma_0^2\begin{pmatrix} E(\mathbf{u}_t\mathbf{u}_t^{\mathrm{T}}) & E(\mathbf{u}_t\mathbf{v}_t^{\mathrm{T}}) \\ E(\mathbf{v}_t\mathbf{u}_t^{\mathrm{T}}) & E(\mathbf{v}_t\mathbf{v}_t^{\mathrm{T}}) \end{pmatrix}^{-1}$$

with $\mathbf{u}_t=(u_t,\ldots,u_{t+1-p})^{\mathrm{T}}$ and $\mathbf{v}_t=(v_t,\ldots,v_{t+1-q})^{\mathrm{T}}$. Details about the asymptotic normality of $\tilde{\boldsymbol{\alpha}}$ can be found, for example, in chapter 8 of Brockwell and Davis (1991). The proposed MLE $\hat{\boldsymbol{\alpha}}$ has the same asymptotic distribution as $\tilde{\boldsymbol{\alpha}}$, which is summarized in theorem 2.

Theorem 2. Under assumptions (a)–(c), $\hat{\boldsymbol{\alpha}}$ is oracally efficient, i.e. as $n\to\infty$

$$\begin{aligned} \hat{\boldsymbol{\alpha}}-\tilde{\boldsymbol{\alpha}} &= o_p(n^{-1/2}), \\ \sqrt{n}\,(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0) &\overset{\mathrm{D}}{\Rightarrow} N_{p+q}(\mathbf{0},\mathbf{V}). \end{aligned} \tag{16}$$

Theorems 1 and 2 provide a formal justification for the proposed two-step procedure in which the residual sequence is used to replace the unobserved stationary time series. According to theorems 1 and 2, the modelling procedure for $\mathbf{x}$ can be adopted for $\hat{\mathbf{x}}$, such as the BIC, which is the commonly used criterion for selecting AR order p and MA order q. Ignoring the constant terms, the BIC in Shumway and Stoffer (2011), page 53, for an ARMA model is defined as

$$\mathrm{BIC}(p',q',\tilde{\boldsymbol{\alpha}}) = \log(Q_n)\,(\tilde{\boldsymbol{\alpha}};\mathbf{x}) + \frac{p'+q'}{n}\log(n),$$

and the selected orders $(\tilde{p},\tilde{q})$ minimize $\mathrm{BIC}(p',q',\tilde{\boldsymbol{\alpha}})$, i.e.

$$(\tilde{p},\tilde{q}) = \arg\min_{(p',q')} \mathrm{BIC}(p',q',\tilde{\boldsymbol{\alpha}}). \tag{17}$$

It is known (Hannan, 1980) that the BIC is consistent in the sense that, for the true orders (p,q),

$$\lim_{n\to\infty} P(\tilde{p}=p;\tilde{q}=q)=1.$$

Suppose that $(\hat{p},\hat{q})$ are the selected orders from equation (17) with $\mathbf{x}$ replaced by $\hat{\mathbf{x}}$ and $\tilde{\boldsymbol{\alpha}}$ replaced by $\hat{\boldsymbol{\alpha}}$; we would expect that $(\hat{p},\hat{q})$ are also consistent, which is the next theorem.

Theorem 3. If $(\hat{p},\hat{q})=\arg\min_{(p',q')}\mathrm{BIC}(p',q',\hat{\boldsymbol{\alpha}})$, then, under assumptions (a)–(c),

$$\lim_{n\to\infty} P(\hat{p}=p;\hat{q}=q)=1. \tag{18}$$

Whereas assumptions (a) and (b) are standard for time series, we shall provide elementary conditions under which the B-spline estimator $\hat{g}_m(\omega)$ satisfies the high level assumption (c) so that the general theorems 1–3 hold at least for a B-spline estimator. For an integer $m' \geqslant 0$ and $\nu \in [0, 1]$, denote by $C^{(m',\nu)}[0, 1]$ the space of functions whose m'th derivatives satisfy Hölder conditions of order ν, i.e.

$$C^{(m',\nu)}[0,1] = \left\{ \phi : [0,1] \to \mathbb{R} \,\middle|\, \|\phi\|_{m',\nu} = \sup_{0 \leqslant x < y \leqslant 1} \frac{|\phi^{(m')}(x) - \phi^{(m')}(y)|}{|x-y|^{\nu}} < \infty \right\}. \tag{19}$$

The following assumptions are straightforward.

Condition 1. The trend function $g(\cdot) \in C^{(m',\nu)}[0, 1], m' < m$.

Condition 2. The number of interior knots $N = N_n$ satisfies $n^{1/4(m'+\nu)} \ll N \ll n^{1/2}$.

Note that assumption 2 necessitates that $4(m' + \nu) > 2$, or $m' + \nu > \frac{1}{2}$; in other words, g either has Hölder continuous derivatives of order $m' > 0$ or it must be Hölder continuous of order greater than $\frac{1}{2}$.

Theorem 4. Under assumptions (a) and (b) and 1 and 2, assumption (c) is fulfilled for the B-spline estimator $\hat{g}_m(\omega)$ and hence theorems 1–3 hold under these assumptions.

Proofs of the general theorems 1–3 are in Appendix A. They indicate that no asymptotic properties of the MLE and BIC are affected by the replacement of $\mathbf{x}$ with $\hat{\mathbf{x}}$. The on-line supplement contains a detailed proof of theorem 4 for any sequence $N = N_n$ under assumption 2 but, since the last term $N^{-2(m'+\nu)} + n^{-1}N$ of expression (S.6) in the supplement, of order $o(n^{-1/2})$, is not uniform for all such sequences, the proof does not hold uniformly either over all possible N_ns that satisfy assumption 2. For instance, if $m' + \nu > \frac{5}{4}$, then, for any $c > 0$, $N_c = [cn^{1/5}]$ satisfies $N_c^{-2(m'+\nu)} + n^{-1}N_c = o(n^{-1/2})$; yet, for any integer $n > 0$, $\sup_{c>0}(N_c^{-2(m'+\nu)} + n^{-1}N_c) = \infty$, so $\sup_{c>0}(N_c^{-2(m'+\nu)} + n^{-1}N_c) \neq o(n^{-1/2})$.

4. Data examples

All computing in this section is carried out using R which is an open-access environment for statistical computing and graphics, developed by the R Core Team (2013).

To implement the default B-spline trend estimator in equation (11), it is crucial to choose the number N of knots on the basis of data. When regression errors are IID an optimal N was provided by Ma (2014) which minimizes the mean integrated squared error of the regression function. A modified version is adapted as follows which incorporates the correlated structure of data:

$$N_{\text{opt}} = \max\left\{ \left[n^{1/5} \left(\frac{\hat{\mu}_2}{180\hat{\gamma}_0} \right)^{1/5} \right] - 1, b \right\}, \tag{20}$$

where $\hat{\gamma}_0$ is the estimate of $\gamma_0 = \text{var}(x_t)$ calculated from equation (2.6) of Hall and Van Keilegom (2003), $\hat{\mu}_2 = \int_0^1 \hat{g}^{(2)}(u)^2 \mathrm{d}u$ with second-order derivative estimate $\hat{g}^{(2)}(u)$ using cubic splines with number of knots $N = [n^{1/9}]$, and $[a]$ is the integer part of a. The positive integer b is the only tuning parameter, and our simulation experiments suggest that $b = 3$ is a robust choice. The above specifications are used in all our numerical work.

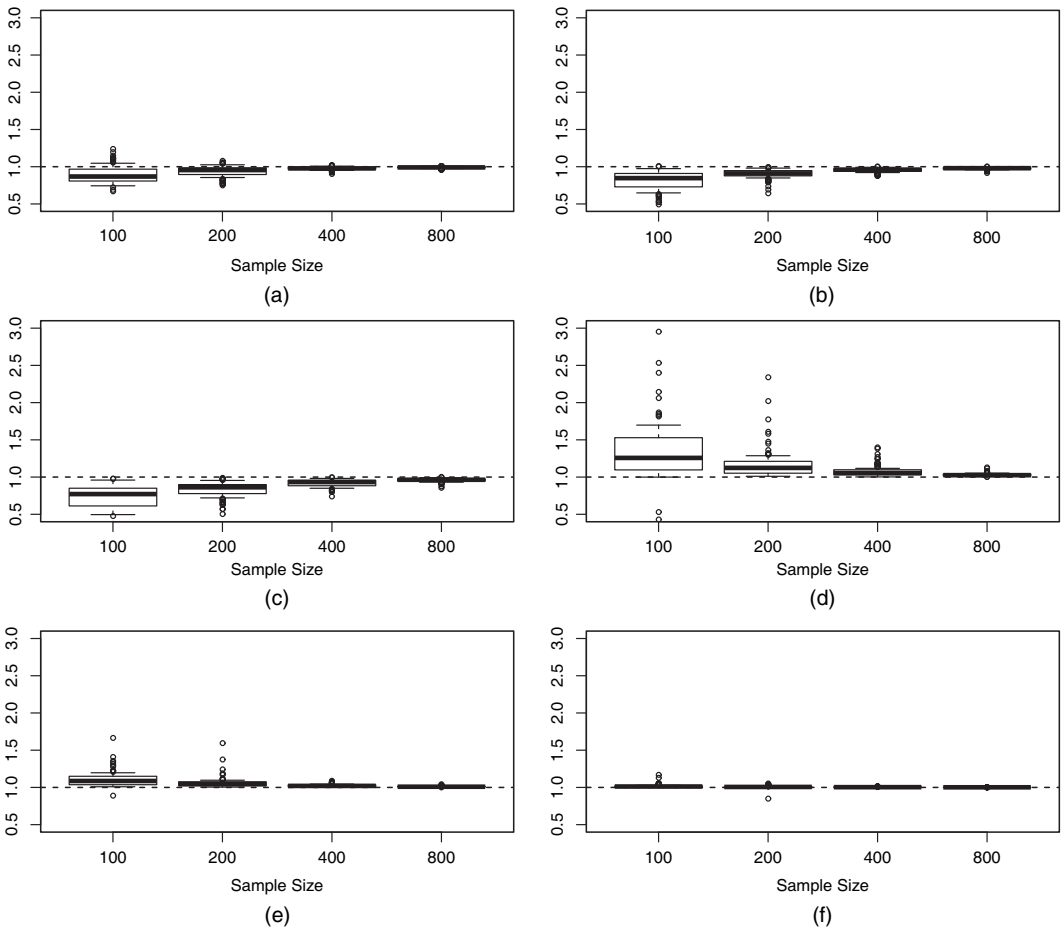

Fig. 1. Boxplots of $\tilde{\theta}/\hat{\theta}$ for the MA(1) process with trend g in equation (21): (a) $\theta_1 = -0.8$; (b) $\theta_1 = -0.4$; (c) $\theta_1 = -0.2$; (d) $\theta_1 = 0.2$; (e) $\theta_1 = 0.4$; (f) $\theta_1 = 0.8$

4.1. Simulation studies

Sample paths of MA(1) and ARMA(1,1) time series errors are generated 100 times, with sample size $n = 400, 800$. The variance $\gamma_0 = 2$ and the trend functions are

$$g(\omega) = \tfrac{1}{2}\sin(\pi\omega), \qquad \omega \in [0, 1]. \tag{21}$$

The parameters are $\theta_{01} = -0.8, -0.4, -0.2, 0.2, 0.4, 0.8$ for the MA(1) and $(\phi_{01}, \theta_{01}) = (-0.8, -0.8), (-0.6, -0.6), (-0.2, -0.2), (-0.8, -0.4), (0.4, -0.8), (-0.8, 0.2), (0.2, 0.8), (0.6, -0.4), (0.1, -0.2), (0.8, 0.8), (0.8, -0.6), (0.6, 0.6)$ for the ARMA(1,1) process. The absolute values of these parameters scatter from close to 0 through close to 1 so the simulated sequences are representative of a wide range of causal and invertible stationary time series. For example, an MA(1) time series is 1 dependent, i.e. the autocovariance function is 0 at any lag larger than 1, and in particular $\mathrm{cov}(x_t, x_{t-1}) = \theta_{01}\sigma_0^2$. Therefore, the simulated MA(1) sequences have correlations ranging from close to -1 to close to 1 at lag 1.

For each sample path of the MA(1) errors, $\hat{\theta}_1$ is computed with residuals from a B-spline trend, and boxplots of the 100 ratios $\tilde{\theta}_1/\hat{\theta}_1$ are in Fig. 1, with the horizontal broken line $y = 1$.

Table 1. Sample means and standard deviations of ARMA(1,1) coefficients' estimates

True coefficients	*B-spline trend estimates*		*Kernel trend estimates*		*Linear trend estimates*	
	$n=400$	$n=800$	$n=400$	$n=800$	$n=400$	$n=800$
$\phi_{01}=-0.8$	-0.798 ± 0.028	-0.797 ± 0.023	-0.812 ± 0.026	-0.808 ± 0.023	-0.846 ± 0.023	-0.846 ± 0.020
$\theta_{01}=-0.8$	-0.757 ± 0.035	-0.767 ± 0.023	-0.578 ± 0.022	-0.630 ± 0.016	-0.007 ± 0.032	-0.006 ± 0.021
$\phi_{01}=-0.6$	-0.598 ± 0.049	-0.597 ± 0.032	-0.612 ± 0.048	-0.606 ± 0.031	-0.683 ± 0.045	-0.683 ± 0.029
$\theta_{01}=-0.6$	-0.606 ± 0.054	-0.605 ± 0.032	-0.546 ± 0.045	-0.568 ± 0.026	-0.135 ± 0.044	-0.133 ± 0.031
$\phi_{01}=-0.2$	-0.150 ± 0.135	-0.191 ± 0.094	-0.164 ± 0.128	-0.197 ± 0.094	-0.274 ± 0.113	-0.301 ± 0.083
$\theta_{01}=-0.2$	-0.266 ± 0.135	-0.212 ± 0.094	-0.246 ± 0.128	-0.204 ± 0.094	-0.073 ± 0.117	-0.043 ± 0.086
$\phi_{01}=-0.8$	-0.795 ± 0.033	-0.800 ± 0.025	-0.802 ± 0.032	-0.804 ± 0.024	-0.843 ± 0.026	-0.846 ± 0.020
$\theta_{01}=-0.4$	-0.418 ± 0.050	-0.408 ± 0.032	-0.369 ± 0.042	-0.379 ± 0.030	0.033 ± 0.035	0.032 ± 0.026
$\phi_{01}=0.4$	0.439 ± 0.103	0.417 ± 0.061	0.392 ± 0.084	0.400 ± 0.058	0.210 ± 0.101	0.209 ± 0.059
$\theta_{01}=-0.8$	-0.856 ± 0.076	-0.817 ± 0.040	-0.801 ± 0.054	-0.797 ± 0.035	-0.514 ± 0.079	-0.505 ± 0.042
$\phi_{01}=-0.8$	-0.793 ± 0.038	-0.793 ± 0.030	-0.795 ± 0.038	-0.794 ± 0.030	-0.818 ± 0.034	-0.818 ± 0.026
$\theta_{01}=0.2$	0.188 ± 0.069	0.186 ± 0.046	0.195 ± 0.068	0.189 ± 0.046	0.293 ± 0.058	0.288 ± 0.041
$\phi_{01}=0.2$	0.176 ± 0.058	0.192 ± 0.040	0.184 ± 0.059	0.194 ± 0.040	0.208 ± 0.058	0.214 ± 0.042
$\theta_{01}=0.8$	0.798 ± 0.035	0.801 ± 0.024	0.796 ± 0.036	0.801 ± 0.024	0.792 ± 0.037	0.797 ± 0.024
$\phi_{01}=0.6$	0.490 ± 0.189	0.529 ± 0.125	0.527 ± 0.174	0.548 ± 0.121	0.651 ± 0.182	0.656 ± 0.101
$\theta_{01}=-0.4$	-0.303 ± 0.200	-0.333 ± 0.136	-0.335 ± 0.190	-0.350 ± 0.132	-0.449 ± 0.212	-0.448 ± 0.123
$\phi_{01}=0.1$	0.276 ± 0.478	0.136 ± 0.389	0.205 ± 0.460	0.111 ± 0.394	-0.111 ± 0.389	-0.199 ± 0.315
$\theta_{01}=-0.2$	-0.394 ± 0.471	-0.241 ± 0.379	-0.314 ± 0.450	-0.213 ± 0.385	0.038 ± 0.382	0.126 ± 0.315
$\phi_{01}=0.8$	0.751 ± 0.039	0.782 ± 0.023	0.772 ± 0.036	0.789 ± 0.023	0.797 ± 0.033	0.801 ± 0.021
$\theta_{01}=0.8$	0.807 ± 0.029	0.803 ± 0.024	0.805 ± 0.029	0.802 ± 0.024	0.802 ± 0.030	0.801 ± 0.024
$\phi_{01}=0.8$	0.670 ± 0.133	0.750 ± 0.078	0.722 ± 0.103	0.765 ± 0.076	0.801 ± 0.078	0.818 ± 0.064
$\theta_{01}=-0.6$	-0.483 ± 0.148	-0.557 ± 0.092	-0.528 ± 0.123	-0.570 ± 0.090	-0.599 ± 0.104	-0.619 ± 0.087
$\phi_{01}=0.6$	0.570 ± 0.048	0.582 ± 0.030	0.582 ± 0.047	0.587 ± 0.029	0.607 ± 0.045	0.604 ± 0.030
$\theta_{01}=0.6$	0.605 ± 0.041	0.605 ± 0.030	0.602 ± 0.041	0.604 ± 0.030	0.595 ± 0.042	0.600 ± 0.030

Table 2. Sample percentages of correct ARMA(1,1) model selection by the BIC

True coefficients (ϕ_{01},θ_{01})	*Results (%) for time series with trend*		*Results (%) for time series without trend*	
	$n=400$	$n=800$	$n=400$	$n=800$
$(-0.8,-0.8)$	0.92	1.00	0.97	0.99
$(-0.8,-0.6)$	0.93	0.98	0.95	0.99
$(-0.8,0.6)$	0.55	0.86	0.71	0.98
$(-0.6,-0.6)$	0.93	0.97	0.95	0.98
$(0.8,0.8)$	0.77	0.84	0.95	0.99
$(0.8,0.6)$	0.70	0.82	0.98	1.00
$(0.8,-0.6)$	0.45	0.84	0.76	0.95
$(0.6,0.6)$	0.79	0.87	0.98	0.98

These boxplots agree with theorem 2: for each θ_{01} the boxes not only become narrower but also closer to 1, as the sample size increases from 100 to 800. Our estimator works better for moderate and highly correlated data, such as $|\theta_{01}|=0.4,0.8$. Compared with $\tilde{\theta}_1$, $\hat{\theta}_1$ tends to overestimate the true value θ_{01} for $-1<\theta_{01}<0$ and to underestimate it otherwise.

Our method has the advantage that the user does not need to know the form of the trend

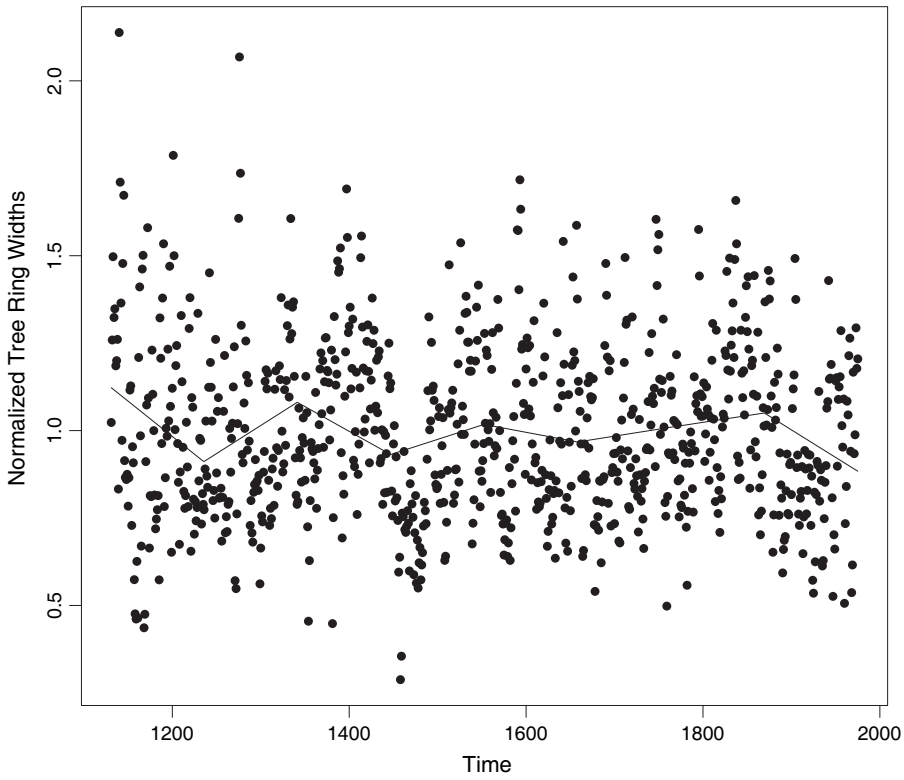

Fig. 2. Annual normalized tree rings with *B*-spline trend estimates

function. To investigate the effects of model misspecification, we estimated the ARMA(1,1) coefficients for the above simulated data. Table 1 includes the sample means and standard deviations of (ϕ_{01}, θ_{01}) estimates from *B*-spline, kernel and linear trend estimates, with the kernel trend estimate computed according to Qiu *et al*. (2013). The non-parametric trends by *B*-spline and kernel exhibit similar accuracy, and both outperform the linear trend. This points to the possibility that a kernel trend could be a viable alternative to *B*-splines, if a version of theorem 4 can be proved for kernel trend.

We also summarize the percentages of correctly selecting the ARMA(1,1) models according to the BICs from the 100 samples. For the sequence with trend, we first obtain residuals based on a *B*-spline trend, and then compute the BICs to choose the model according to equation (17). To ease the comparison, Table 2 includes both the percentages calculated from the simulated ARMA(1,1) sequences ('time series without trend') and the *B*-spline residuals ('time series with trend'). As expected, the correct selection percentages in both cases increase with n. Although the simulated ARMA(1,1) sequences without trend always have higher percentages than those calculated from the *B*-spline residuals for both sample sizes, the percentages become relatively close at $n = 800$. In other words, for the ARMA(1,1) coefficients specification the difference diminishes with increasing n and is quite insignificant for certain specifications, all confirming theorem 3.

These findings suggest that, in place of the latent ARMA series, one could use residuals obtained by removing *B*-spline trend estimates for coefficients' estimation and model selection. In the next subsection, a real data example is studied to illustrate such an application.

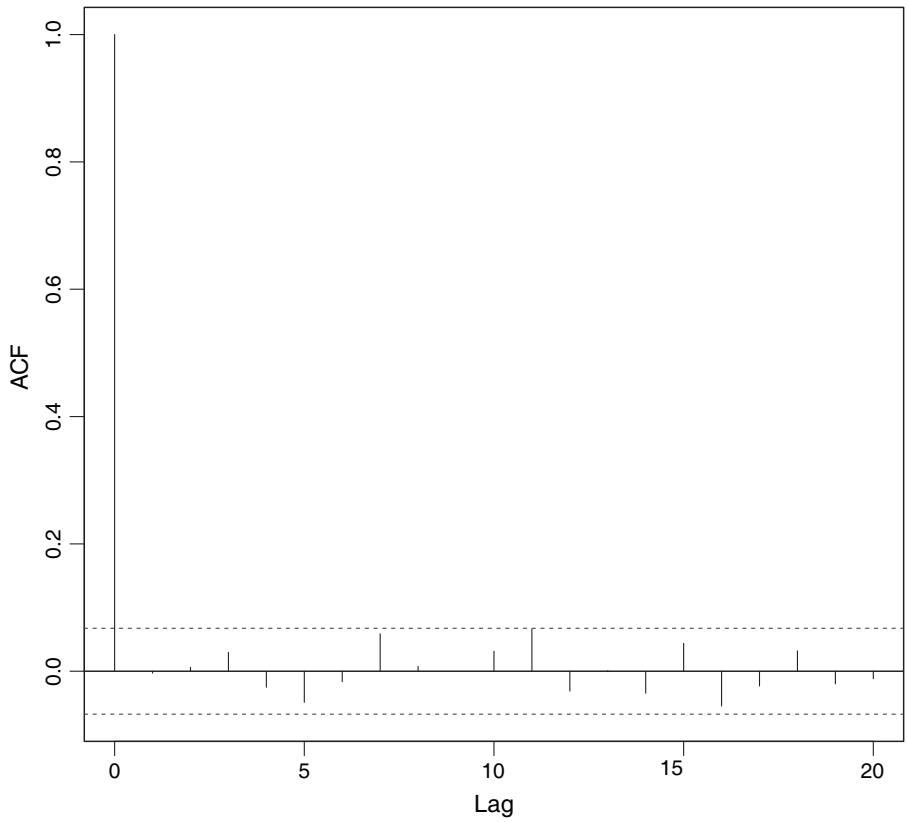

Fig. 3. ARMA(1,2) residuals auto-correlations of normalized tree rings

4.2. Application

In this subsection, we apply the tested model selection and coefficients' estimation tools from the previous subsection to real time series data, depicted in Fig. 2. The data were downloaded from datamarket.com and consist of 845 observations of annual normalized tree ring widths at San Gabriel, Auch, 1042A in Chile, Cipres from 1131 to 1975. Each observation corresponds to a normalized tree ring width in dimensionless units. The sample estimate of $\sqrt{\gamma_0}$ is 0.249 based on the formula of Hall and Van Keilegom (2003), and the range of trend estimates is $\max_{1\leqslant t\leqslant n}\hat{g}(u_t) - \min_{1\leqslant t\leqslant n}\hat{g}(u_t) = 0.072$. The ratio between the estimated $\sqrt{\gamma_0}$ and the range of the estimated trend is very similar to that in the simulation studies. The number of knots calculated by using equation (20) is $N_{\text{opt}} = 3$, and the fitted trend function is given by the full curve in Fig. 2.

The best model selected by the BIC rule is an ARMA(1,2) process, with parameter estimates as well as their standard errors $\hat{\phi}_1 = 0.716 \pm 0.082$, $\hat{\theta}_1 = -0.066 \pm 0.093$, $\hat{\theta}_2 = -0.212 \pm 0.070$ and $\hat{\sigma} = 0.199$. Model diagnostic checking was conducted by using the auto-correlation functions of the ARMA(1,2) residuals. All auto-correlations of the residuals at lags 0–20 are within a 95% confidence interval in Fig. 3; therefore an appropriate ARMA(1,2) model for the annual normalized tree ring x_t is

$$x_t - 0.716x_{t-1} = \epsilon_t - 0.066\epsilon_{t-1} - 0.212\epsilon_{t-2}, \qquad \epsilon_t \sim \text{IID}(0, 0.199^2).$$

5. Concluding remarks

Oracle efficiency was proved for MLEs of ARMA(p,q) coefficients by using residuals from trend estimates that meet certain general assumptions, together with consistency of the well-known BIC for model selection. The B-spline trend estimate is shown in the on-line supplement to satisfy the assumptions, and two-step estimation and model selection procedures can be easily implemented by using any software package such as R, which has a built-in function or subroutine for estimation of the ARMA parameters $\boldsymbol{\alpha}$. These procedures are not only theoretically well justified, but also computationally simple, and we reasonably expect that similar results hold for time series errors that are much more complicated than ARMA.

Acknowledgements

The work was partially supported by funding from US National Science Foundation award DMS 1007594, Jiangsu Specially-Appointed Professor Program SR10700111, Jiangsu Key Discipline Program (Statistics) ZY107992, National Natural Science Foundation of China award NSFC 11371272 and Research Fund for the Doctoral Program of Higher Education of China award 20133201110002. The authors are grateful to three reviewers, the Associate Editor and the Joint Editor for thoughtful comments leading to significant improvements of the paper.

Appendix A

A.1. Preliminaries

Before we provide the detailed proofs for theorems 1–3, we introduce some notation that will be used frequently. For any vector $\mathbf{d}=(d_1,\ldots,d_n)^{\mathrm{T}}$, we denote $\|\mathbf{d}\|_\infty=\sup_{1\leqslant t\leqslant n}|d_t|$ is the supremum norm and $\|\mathbf{d}\|=\sqrt{\Sigma_{j=1}^n d_j^2}$ is the Euclidean norm. Unless otherwise indicated, throughout the appendix, C and ρ denote constants satisfying $C>0$ and $0<\rho<1$.

Definition 1. A deterministic vector $\mathbf{a}=(a_0,\ldots,a_{n-1})^{\mathrm{T}}$ is called (C,ρ) exponentially bounded if $|a_j|\leqslant C\rho^j$ for any $0\leqslant j\leqslant n-1$; a random vector $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^{\mathrm{T}}$ is called (C,ρ) exponentially correlated if $E(\xi_k)=0$ and $|E(\xi_k\xi_l)|\leqslant C\rho^{|k-l|}$, $1\leqslant k,l\leqslant n$.

Lemma 1. For any deterministic vector $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^{\mathrm{T}}$ and (C,ρ) exponentially bounded sequence $\mathbf{a}=(a_0,\ldots,a_{n-1})^{\mathrm{T}}$, define $\boldsymbol{\nu}=(\nu_1,\ldots,\nu_n)^{\mathrm{T}}$ as $\nu_t=\Sigma_{j=0}^{t-1}a_j\mu_{t-j}$, $\|\boldsymbol{\nu}\|_\infty\leqslant C(1-\rho)^{-1}\|\boldsymbol{\mu}\|_\infty$.

Proof. By definition 1 of (C,ρ) exponential boundedness, $|a_j|\leqslant C\rho^j$ for any $0\leqslant j\leqslant n-1$; hence

$$\|\boldsymbol{\nu}\|_\infty=\max_{1\leqslant t\leqslant n}\left|\sum_{j=0}^{t-1}a_j\mu_{t-j}\right|\leqslant\|\boldsymbol{\mu}\|_\infty\max_{1\leqslant t\leqslant n}C\sum_{j=0}^{t-1}\rho^j\leqslant C\,(1-\rho)^{-1}\,\|\boldsymbol{\mu}\|_\infty.$$

The proof is complete. □

Lemma 2. For any deterministic vector $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_n)^{\mathrm{T}}$ and a (C,ρ) exponentially correlated random vector $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_n)^{\mathrm{T}}$, $E(\Sigma_{j=1}^n\mu_j\xi_j)^2\leqslant C(1-\rho)^{-1}n\|\boldsymbol{\mu}\|_\infty^2$; hence $\Sigma_{j=1}^n\mu_j\xi_j=O_p(n^{1/2}\|\boldsymbol{\mu}\|_\infty)$ as $n\to\infty$.

Proof. By definition 1, $|E(\xi_k\xi_l)|\leqslant C\rho^{|k-l|}$ for any $1\leqslant k,l\leqslant n$. Note that

$$\sum_{j_1=1}^n\sum_{j_2=1}^n\rho^{|j_1-j_2|}\leqslant n(1-\rho)^{-1},\tag{22}$$

which implies that

$$E\left(\sum_{j=1}^n\mu_j\xi_j\right)^2\leqslant\sum_{j_1=1}^n\sum_{j_2=1}^n|\mu_{j_1}\|\mu_{j_2}\|E(\xi_{j_1}\xi_{j_2})|\leqslant\|\boldsymbol{\mu}\|_\infty^2\sum_{j_1=1}^n\sum_{j_2=1}^nC\rho^{|j_1-j_2|}\leqslant C(1-\rho)^{-1}\,n\|\boldsymbol{\mu}\|_\infty^2.$$

The proof is complete. □

Lemma 3. For a (C_ξ, ρ_ξ) exponentially correlated random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^{\mathrm{T}}$ and a (C_a, ρ_a) exponentially bounded vector $\mathbf{a} = (a_0, \ldots, a_{n-1})^{\mathrm{T}}$, define $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$ as $\eta_t = \Sigma_{j=0}^{t-1} a_j \xi_{t-j}$, $1 \leqslant t \leqslant n$. Then $\boldsymbol{\eta}$ is (C_η, ρ_η) exponentially correlated for some $C_\eta > 0$ and $0 < \rho_\eta < 1$.

Proof. By definition 1, $|E(\xi_k \xi_l)| \leqslant C_\xi \rho_\xi^{|k-l|}$ for any $1 \leqslant k, l \leqslant n$, and $|a_j| \leqslant C_a \rho_a^j$ for any $0 \leqslant j \leqslant n-1$. Define $C_\eta = \max[C_a^2 C_\xi (1-\rho_\eta^2)^{-2}, C_a^2 C_\xi \{1 - \min(\rho_\xi, \rho_a)/\rho_\eta\}^{-1}(1-\rho_\eta^2)^{-1}]$ and $\rho_\eta = \max(\rho_\xi, \rho_a)$. It is obvious that $E(\eta_j) = 0$ for any $1 \leqslant j \leqslant n$. We shall show that $\boldsymbol{\eta}$ is (C_η, ρ_η) exponentially correlated.

Without loss of generality, we assume that $l \geqslant k$. Define $S_1 = \Sigma_{j_1=1}^{k} \Sigma_{j_2=1}^{l-k+j_1} \rho_a^{j_1} \rho_\xi^{l-k+j_1} (\rho_a/\rho_\xi)^{j_2}$ and $S_2 = \Sigma_{j_1=1}^{k} \Sigma_{j_2=l-k+j_1+1}^{l} \rho_a^{j_1+j_2} \rho_\xi^{k-j_1-l+j_2}$. It is obvious that $S_1 \leqslant \rho_\eta^{l-k} \{1 - \min(\rho_\xi, \rho_a)/\rho_\eta\}^{-1}(1-\rho_\eta^2)^{-1}$ and $S_2 \leqslant \rho_\eta^{l-k}(1-\rho_\eta^2)^{-2}$. Thus

$$
\begin{aligned}
|E(\eta_k \eta_l)| &= \left| \sum_{j_1=1}^{k} \sum_{j_2=1}^{l} a_{j_1} a_{j_2} E(\xi_{k-j_1+1} \xi_{l-j_2+1}) \right| \leqslant \sum_{j_1=1}^{k} \sum_{j_2=1}^{l} C_a^2 \rho_a^{j_1+j_2} C_\xi \rho_\xi^{|k-j_1-l+j_2|} \\
&= C_a^2 C_\xi \sum_{j_1=1}^{k} \sum_{j_2=l-k+j_1+1}^{l} \rho_a^{j_1+j_2} \rho_\xi^{k-j_1-l+j_2} + C_a^2 C_\xi \sum_{j_1=1}^{k} \sum_{j_2=1}^{l-k+j_1} \rho_a^{j_1+j_2} \rho_\xi^{l-k+j_1-j_2} \\
&= C_a^2 C_\xi (S_1 + S_2) \leqslant C_\eta \rho_\eta^{|k-l|}.
\end{aligned}
$$

Therefore, $\boldsymbol{\eta}$ is (C_η, ρ_η) exponentially correlated. □

Lemma 4. Under assumptions (a)–(c), as $n \to \infty$

$$
R_{0n} = n^{-1} \sup_{\alpha \in \Xi} \sum_{t=p+1}^{n} \left(\sum_{j=0}^{t-1} \pi_j x_{t-j} \right)^2 = O_p(1), \tag{23}
$$

$$
R_{1n} = \sup_{\alpha \in \Xi} n^{-1} \sum_{t=p+1}^{n} \left[\sum_{j=0}^{t-1} \pi_j \{ g(\omega_{t-j}) - \hat{g}(\omega_{t-j}) \} \right]^2 = o_p(n^{-1/2}), \tag{24}
$$

$$
R_{2n} = \sup_{\alpha \in \Xi} n^{-1} \left| \sum_{t=p+1}^{n} \left(\sum_{j=0}^{t-1} \pi_j x_{t-j} \right) \sum_{j=0}^{t-1} \pi_j \{ g(\omega_{t-j}) - \hat{g}(\omega_{t-j}) \} \right| = o_p(n^{-1/4}). \tag{25}
$$

Proof. By Cauchy–Schwarz inequality, lemma 1 and equation (15), we obtain

$$
\begin{aligned}
E(R_{0n}) &= n^{-1} E \left(\sup_{\alpha \in \Xi} \sum_{t=p+1}^{n} \sum_{j_1=0}^{t-1} \sum_{j_2=0}^{t-1} \pi_{j_1} \pi_{j_2} x_{t-j_1} x_{t-j_2} \right) \\
&\leqslant n^{-1} \sum_{t=p+1}^{n} \sum_{j_1=0}^{t-1} \sum_{j_2=0}^{t-1} \Pi_{j_1} \Pi_{j_2} E|x_{t-j_1} x_{t-j_2}| \\
&\leqslant n^{-1} C_\pi \sum_{t=p+1}^{n} \left\{ \sum_{j=0}^{t-1} \rho_\pi^j E(x_{t-j}^2)^{1/2} \right\}^2 = O\left\{ n^{-1} C \sum_{t=p+1}^{n} \left(\sum_{j=0}^{t-1} \rho_\pi^j \right)^2 \right\} = O(1),
\end{aligned}
$$

so equation (23) follows because $R_{0n} \geqslant 0$. Similarly

$$
\begin{aligned}
E(R_{1n}) &\leqslant n^{-1} C_\pi \sum_{t=p+1}^{n} \left(\sum_{j=0}^{t-1} \rho_\pi^j [E\{ g(\omega_{t-j}) - \hat{g}(\omega_{t-j}) \}^2]^{1/2} \right)^2 \\
&= O\left[n^{-1} C \sum_{t=p+1}^{n} \left(\sum_{j=0}^{t-1} \rho_\pi^j \right)^2 \max_{1 \leqslant t \leqslant n} E\{ \hat{g}(\omega_t) - g(\omega_t) \}^2 \right] = o(n^{-1/2})
\end{aligned}
$$

by applying equation (12) in assumption (c), and equation (24) follows because $R_{1n} \geqslant 0$. Finally

$$
R_{2n} = \sup_{\alpha \in \Xi} n^{-1} \left| \sum_{t=p+1}^{n} \left(\sum_{j=0}^{t-1} \pi_j x_{t-j} \right) \sum_{j=0}^{t-1} \pi_j \{ g(\omega_{t-j}) - \hat{g}(\omega_{t-j}) \} \right| \leqslant R_{0n}^{1/2} R_{1n}^{1/2} = o_p(n^{-1/4})
$$

by equations (23) and (24).

A.2. Proofs of theorems 1 and 2

A.2.1. Proof of theorem 1

Note that $\sup_{\boldsymbol{\alpha}\in\Xi}|Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}})-Q_n(\boldsymbol{\alpha};\mathbf{x})|\leqslant R_{1n}+2R_{2n}$, where R_{1n} and R_{2n} are defined in equations (24) and (25). Lemma 4 implies that

$$\sup_{\boldsymbol{\alpha}\in\Xi}|Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}})-Q_n(\boldsymbol{\alpha};\mathbf{x})|=o_p(1). \tag{26}$$

According to assumptions (a) and (b), the roots of $\Theta(\boldsymbol{\alpha},z)=0$ are larger than or equal to $1+\epsilon$ for some $\epsilon>0$. According to the proof of theorem 8.4.1 of Fuller (1996), page 432, $Q_n(\boldsymbol{\alpha};\mathbf{x})\to^{\mathrm{P}}\mathrm{var}\{\epsilon_t(\boldsymbol{\alpha})\}$ uniformly in $\boldsymbol{\alpha}\in\Xi$, where $\epsilon_t(\boldsymbol{\alpha})=\Phi(\boldsymbol{\alpha},B)\Theta^{-1}(\boldsymbol{\alpha},B)x_t$ with the backward shift operator B defined in equation (3). Moreover, the function $\mathrm{var}\{\epsilon_t(\boldsymbol{\alpha})\}$ has a unique minimizer at $\boldsymbol{\alpha}_0$. Therefore, theorem 4.1.1 of Amemiya (1985), page 106, ensures that $\hat{\boldsymbol{\alpha}}\to^{\mathrm{P}}\boldsymbol{\alpha}_0$, and

$$\begin{aligned}|\hat{\sigma}^2-\sigma_0^2|&=|Q_n(\hat{\boldsymbol{\alpha}};\hat{\mathbf{x}})-\mathrm{var}\{\epsilon_t(\boldsymbol{\alpha}_0)\}|\\&\leqslant|Q_n(\hat{\boldsymbol{\alpha}};\hat{\mathbf{x}})-Q_n(\hat{\boldsymbol{\alpha}};\mathbf{x})|+|Q_n(\hat{\boldsymbol{\alpha}};\mathbf{x})-\mathrm{var}\{\epsilon_t(\hat{\boldsymbol{\alpha}})\}|+|\mathrm{var}\{\epsilon_t(\hat{\boldsymbol{\alpha}})\}-\mathrm{var}\{\epsilon_t(\boldsymbol{\alpha}_0)\}|\xrightarrow{\mathrm{P}}0;\end{aligned}$$

hence $\hat{\sigma}^2\to^{\mathrm{P}}\sigma_0^2$. The proof is complete. □

Lemma 5. Under assumptions (a)–(c), as $n\to\infty$, for $1\leqslant k\leqslant p+q$,

$$\frac{\partial Q_n(\boldsymbol{\alpha}_0;\hat{\mathbf{x}})}{\partial\alpha_k}-\frac{\partial Q_n(\boldsymbol{\alpha}_0;\mathbf{x})}{\partial\alpha_k}=o_p(n^{-1/2}). \tag{27}$$

Proof. For any $k=1,\ldots,p+q$, we obtain that

$$\frac{\partial Q_n(\boldsymbol{\alpha}_0;\hat{\mathbf{x}})}{\partial\alpha_k}-\frac{\partial Q_n(\boldsymbol{\alpha}_0;\mathbf{x})}{\partial\alpha_k}=S_{1n}+S_{2n}+S_{3n},$$

where

$$\begin{aligned}S_{1n}&=\frac{1}{n}\sum_{t=p+1}^{n}\left[\sum_{j=0}^{t-1}\pi_{0j}\{g(\omega_{t-j})-\hat{g}(\omega_{t-j})\}\right]\sum_{j=0}^{t-1}\frac{\partial\pi_{0j}}{\partial\alpha_k}\{g(\omega_{t-j})-\hat{g}(\omega_{t-j})\},\\S_{2n}&=\frac{1}{n}\sum_{t=p+1}^{n}\left(\sum_{j=0}^{t-1}\frac{\partial\pi_{0j}}{\partial\alpha_k}x_{t-j}\right)\sum_{j=0}^{t-1}\pi_{0j}\{\hat{g}(\omega_{t-j})-g(\omega_{t-j})\},\\S_{3n}&=\frac{1}{n}\sum_{t=p+1}^{n}\left(\sum_{j=0}^{t-1}\pi_{0j}x_{t-j}\right)\sum_{j=0}^{t-1}\frac{\partial\pi_{0j}}{\partial\alpha_k}\{\hat{g}(\omega_{t-j})-g(\omega_{t-j})\}.\end{aligned}$$

From inequality (27) of Yao and Brockwell (2006), $\{\partial\pi_{0j}/\partial\alpha_k\}_{j=0}^{n-1}$ is (C,s) exponentially bounded for some $C>0, s\in(0,1)$, so

$$\begin{aligned}&E\left(n^{-1}\sum_{t=p+1}^{n}\left[\sum_{j=0}^{t-1}\frac{\partial\pi_{0j}}{\partial\alpha_k}\{g(\omega_{t-j})-\hat{g}(\omega_{t-j})\}\right]^2\right)\\&\quad=n^{-1}\sum_{t=p+1}^{n}\sum_{j_1=0}^{t-1}\sum_{j_2=0}^{t-1}\frac{\partial\pi_{0j_1}}{\partial\alpha_k}\frac{\partial\pi_{0j_2}}{\partial\alpha_k}E\{g(\omega_{t-j_1})-\hat{g}(\omega_{t-j_1})\}\{g(\omega_{t-j_2})-\hat{g}(\omega_{t-j_2})\}\\&\quad\leqslant n^{-1}\sum_{t=p+1}^{n}\sum_{j_1=0}^{t-1}\sum_{j_2=0}^{t-1}Cs^{j_1}Cs^{j_2}\max_{1\leqslant t'\leqslant n}E\{\hat{g}(\omega_{t'})-g(\omega_{t'})\}^2\\&\quad\leqslant\max_{1\leqslant t'\leqslant n}E\{\hat{g}(\omega_{t'})-g(\omega_{t'})\}^2n^{-1}C^2\sum_{t=p+1}^{n}\sum_{j_1=0}^{t-1}\sum_{j_2=0}^{t-1}s^{j_1+j_2}\\&\quad=\max_{1\leqslant t\leqslant n}E\{\hat{g}(\omega_t)-g(\omega_t)\}^2O(1)=o(n^{-1/2}).\end{aligned}$$

Thus

$$n^{-1}\sum_{t=p+1}^{n}\left[\sum_{j=0}^{t-1}\frac{\partial\pi_{0j}}{\partial\alpha_k}\{g(\omega_{t-j})-\hat{g}(\omega_{t-j})\}\right]^2=o_p(n^{-1/2}). \tag{28}$$

Meanwhile, equation (24) entails that

$$n^{-1}\sum_{t=p+1}^{n}\left[\sum_{j=0}^{t-1}\pi_{0j}\{g(\omega_{t-j})-\hat{g}(\omega_{t-j})\}\right]^2 \leqslant R_{1n}=o_p(n^{-1/2}). \tag{29}$$

Therefore, expressions (28) and (29) and the Cauchy–Schwarz inequality lead to $S_{1n}=o_p(n^{-1/2})$.

Noting next that equations (13) and (14) in assumption (c) state exactly that $S_{2n}=o_p(n^{-1/2})$ and $S_{3n}=o_p(n^{-1/2})$, the proof is complete. □

Lemma 6. Under assumptions (a)–(c), as $n\to\infty$, for $1\leqslant j,k\leqslant p+q$,

$$\sup_{\boldsymbol{\alpha}\in\Xi}\left|\frac{\partial^2 Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}})}{\partial\alpha_j\partial\alpha_k}-\frac{\partial^2 Q_n(\boldsymbol{\alpha};\mathbf{x})}{\partial\alpha_j\partial\alpha_k}\right|=o_p(1). \tag{30}$$

Proof. For brevity we give the details of derivation for only

$$\sup_{\boldsymbol{\alpha}\in\Xi}\left|\frac{\partial^2 Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}})}{\partial\phi_j\partial\phi_k}-\frac{\partial^2 Q_n(\boldsymbol{\alpha};\mathbf{x})}{\partial\phi_j\partial\phi_k}\right|=o_p(1). \tag{31}$$

Note that, from the definition of $\{\pi_l\}_{l=0}^{\infty}$, $\partial^2\pi_l/\partial\phi_j\partial\phi_k=0$. Thus, for $1\leqslant j,k\leqslant p$,

$$\frac{\partial^2 Q_n(\boldsymbol{\alpha};\mathbf{x})}{\partial\phi_j\partial\phi_k}=\frac{2}{n}\sum_{t=p+1}^{n}\left(\sum_{l=0}^{t-1}\frac{\partial\pi_l}{\partial\phi_j}x_{t-l}\right)\sum_{l=0}^{t-1}\frac{\partial\pi_l}{\partial\phi_k}x_{t-l},$$
$$\frac{\partial^2 Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}})}{\partial\phi_j\partial\phi_k}=\frac{2}{n}\sum_{t=p+1}^{n}\left(\sum_{l=0}^{t-1}\frac{\partial\pi_l}{\partial\phi_j}\hat{x}_{t-l}\right)\sum_{l=0}^{t-1}\frac{\partial\pi_l}{\partial\phi_k}\hat{x}_{t-l}.$$

Following a similar procedure to the proof of lemma 4, it is obvious that result (31) is true.

A.2.2. Proof of theorem 2

Consider the Taylor series expansions

$$\frac{\partial Q_n(\tilde{\boldsymbol{\alpha}},\mathbf{x})}{\partial\boldsymbol{\alpha}}=\frac{\partial Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}}+\frac{\partial^2 Q_n(\boldsymbol{\alpha}_1,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}(\tilde{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0),$$
$$\frac{\partial Q_n(\hat{\boldsymbol{\alpha}},\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}}=\frac{\partial Q_n(\boldsymbol{\alpha}_0,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}}+\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}_0),$$

where $\boldsymbol{\alpha}_1$ is a point on the linear segment between $\tilde{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_0$, and $\boldsymbol{\alpha}_2$ between $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{\alpha}_0$. First note that, because of consistency of $\hat{\boldsymbol{\alpha}}$ given by theorem 1, $\boldsymbol{\alpha}_2\to^{\mathrm{P}}\boldsymbol{\alpha}_0$ and likewise $\boldsymbol{\alpha}_1\to^{\mathrm{P}}\boldsymbol{\alpha}_0$ as $n\to\infty$. Hence

$$\frac{\partial^2 Q_n(\boldsymbol{\alpha}_1,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}-\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\xrightarrow{\mathrm{P}}0.$$

Thus applying equation (30)

$$\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}-\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}=\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}-\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}+\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}-\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\xrightarrow{\mathrm{P}}0.$$

The boundedness in probability of $\{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}\}^{-1}$ implies that, as $n\to\infty$, we have

$$\left|\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_1,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}\right|+\left|\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}\right|=O_p(1), \tag{32}$$

$$\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_1,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}=\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}+o_p(1), \tag{33}$$

$$\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}=\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}+o_p(1). \tag{34}$$

Next write $\tilde{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}$ as

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}&=\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}\frac{\partial Q_n(\boldsymbol{\alpha}_0,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}}-\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_1,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}\frac{\partial Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}}\\&=\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}\left\{\frac{\partial Q_n(\boldsymbol{\alpha}_0,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}}-\frac{\partial Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}}\right\}\\&\quad+\left[\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_2,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}-\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}\right]\frac{\partial Q_n(\boldsymbol{\alpha}_0,\hat{\mathbf{x}})}{\partial\boldsymbol{\alpha}}\\&\quad+\left[\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}-\left\{\frac{\partial^2 Q_n(\boldsymbol{\alpha}_1,\mathbf{x})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}}\right\}^{-1}\right]\frac{\partial Q_n(\boldsymbol{\alpha}_0,\mathbf{x})}{\partial\boldsymbol{\alpha}}.\end{aligned}$$

It is known that $\partial Q_n(\boldsymbol{\alpha}_0,\mathbf{x})/\partial\boldsymbol{\alpha}=O_p(n^{-1/2})$, which, together with equation (27), implies that $\partial Q_n(\boldsymbol{\alpha}_0;\hat{\mathbf{x}})/\partial\boldsymbol{\alpha}=O_p(n^{-1/2})$. These facts on the gradient of Q_n and equations (32), (33) and (34) on the Hessian matrix of Q_n, entail that $\tilde{\boldsymbol{\alpha}}-\hat{\boldsymbol{\alpha}}=o_p(n^{-1/2})$. The second part of expression (16) then follows by Slutsky's theorem.

A.3. Proof of theorem 3

To show consistency of the BIC based on $\hat{\mathbf{x}}$, we introduce some notation. Let (p',q') be candidate orders of ARMA processes, (p,q) the unique true orders and $Q_n(\tilde{\boldsymbol{\alpha}}_{p',q'};\mathbf{x})$ and $Q_n(\tilde{\boldsymbol{\alpha}}_{p,q};\mathbf{x})$ correspond to orders (p',q') and (p,q) respectively. In particular, it is obvious that $Q_n(\tilde{\boldsymbol{\alpha}}_{p',q'};\mathbf{x})$ is the mean-squared error with the MLE $\tilde{\boldsymbol{\alpha}}_{p',q'}$ for orders (p',q') based on $\mathbf{x}$. Likewise, $Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})$ and $Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})$ denote the mean-squared errors based on $\hat{\mathbf{x}}$. The following definition is from Hannan (1980).

Definition 2. If the candidate orders (p',q') satisfy $p'\geqslant p$, $q'\geqslant q$ and $p'+q'>p+q$, they overfit; if $p'<p$ or $q'<q$, they underfit.

A.3.1. Proof of theorem 3

It is clear that result (18) follows if we show that

$$\lim_{n\to\infty}P\{\mathrm{BIC}(p',q',\hat{\boldsymbol{\alpha}})-\mathrm{BIC}(p,q,\hat{\boldsymbol{\alpha}})>0\}=1 \tag{35}$$

for any orders $(p',q')\neq(p,q)$. We shall show result (35) in the cases of overfitting and underfitting.

A.3.1.1. Case of overfitting. For the correct orders (p,q), $\hat{\boldsymbol{\alpha}}_{p,q}=\hat{\boldsymbol{\alpha}}$. Hence, according to theorem 2, $\hat{\boldsymbol{\alpha}}_{p,q}-\boldsymbol{\alpha}_0=O_p(n^{-1/2})$. Meanwhile, $\partial Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})/\partial\boldsymbol{\alpha}\equiv 0$, as $\hat{\boldsymbol{\alpha}}_{p,q}$ minimizes $Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}})$, and lemma 6 ensures that $\partial^2 Q_n(\boldsymbol{\alpha};\hat{\mathbf{x}})/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathrm{T}}$ is uniformly bounded for $\boldsymbol{\alpha}\in\Xi$. Therefore, applying the second-order Taylor series expansion it follows that

$$Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})-Q_n(\boldsymbol{\alpha}_0;\hat{\mathbf{x}})=O_p(n^{-1}).$$

For overfitting, $p'-p\geqslant 0$ and $q'-q\geqslant 0$. Note that the true ARMA(p,q) model is also an ARMA(p',q') model with coefficients vector $\boldsymbol{\alpha}_{p',q',0}=(\phi_{01},\ldots,\phi_{0p},\mathbf{0}_{p'-p}^{\mathrm{T}},\theta_{01},\ldots,\theta_{0q},\mathbf{0}_{q'-q}^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{0}_r$ denotes an r-dimensional vector of 0s. Hence we conclude that

$$Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})-Q_n(\boldsymbol{\alpha}_{p',q',0};\hat{\mathbf{x}})=O_p(n^{-1}).$$

By definition, $Q_n(\boldsymbol{\alpha}_0;\hat{\mathbf{x}})\equiv Q_n(\boldsymbol{\alpha}_{p',q',0};\hat{\mathbf{x}})$, so the above two equations imply that

$$Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})-Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})=O_p(n^{-1}).$$

Applying this to the BICs, we have

$$\begin{aligned}\mathrm{BIC}(p',q',\hat{\boldsymbol{\alpha}})-\mathrm{BIC}(p,q,\hat{\boldsymbol{\alpha}}) &= \log\left\{1+\frac{Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})-Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})}{Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})}\right\}+\frac{p'+q'-p-q}{n}\log(n)\\ &= \frac{Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})-Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})}{Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})}\{1+o_p(1)\}+\frac{p'+q'-p-q}{n}\log(n),\\ &= O_p(n^{-1})+\frac{p'+q'-p-q}{n}\log(n)=\frac{p'+q'-p-q}{n}\log(n)\{1+o_p(1)\}.\end{aligned}$$

Since $p'+q'-p-q>0$, equation (35) follows.

A.3.1.2. Case of underfitting. In the case of underfitting, either $p'<p$ or $q'<q$. By equation (3) of Hannan (1980)

$$Q_n(\tilde{\boldsymbol{\alpha}}_{p',q'};\mathbf{x})-Q_n(\tilde{\boldsymbol{\alpha}}_{p,q};\mathbf{x})\to c_{p',q'} \qquad \text{almost surely}$$

for some constant $c_{p',q'}>0$. Next, equation (26) implies that $|Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})-Q_n(\tilde{\boldsymbol{\alpha}}_{p,q};\mathbf{x})|=o_p(1)$, and likewise one can show that $|Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})-Q_n(\tilde{\boldsymbol{\alpha}}_{p',q'};\mathbf{x})|=o_p(1)$. Hence we conclude that

$$Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})-Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})\xrightarrow{\mathrm{P}}c_{p',q'}.$$

Applying this equation to the BICs, we have

$$\begin{aligned}\mathrm{BIC}(p',q',\hat{\boldsymbol{\alpha}})-\mathrm{BIC}(p,q,\hat{\boldsymbol{\alpha}}) &= \log\left\{1+\frac{Q_n(\hat{\boldsymbol{\alpha}}_{p',q'};\hat{\mathbf{x}})-Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})}{Q_n(\hat{\boldsymbol{\alpha}}_{p,q};\hat{\mathbf{x}})}\right\}+\frac{p'+q'-p-q}{n}\log(n)\\ &= \log\left(1+\frac{c_{p',q'}}{\sigma_0^2}\right)\{1+o_p(1)\}+\frac{p'+q'-p-q}{n}\log(n).\end{aligned}$$

Therefore, result (35) holds.

References

Amemiya, T. (1985) *Advanced Econometrics*. Cambridge: Harvard University Press.

de Boor, C. (2001) *A Practical Guide to Splines*. New York: Springer.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994) *Time Series Analysis: Forecasting and Control*, 3rd edn. Englewood Cliffs: Prentice Hall.

Brockwell, P. J. and Davis, R. A. (1991) *The Analysis of Time Series: Theory and Methods*, 2nd edn. New York: Springer.

Fan, J. and Yao, Q. W. (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer.

Fuller, W. A. (1996) *Introduction to Statistical Time Series*, 2nd edn. New York: Wiley.

Garel, B. and Hallin, M. (1995) Local asymptotic normality of multivariate ARMA processes with a linear trend. *Ann. Inst. Statist. Math.*, **47**, 551–579.

Hall, P. and Van Keilegom, I. (2003) Using difference-based methods for inference in nonparametric regression with time series errors. *J. R. Statist. Soc.* B, **65**, 443–456.

Hannan, E. J. (1980) The estimation of the order of an ARMA process. *Ann. Statist.*, **8**, 1071–1081.

Ma, S. (2014) A plug-in the number of knots selector for polynomial spline regression. *J. Nonparam. Statist.*, **26**, 489–507.

Pierce, D. A. (1971) Least squares estimation in regression model with autoregressive-moving average errors. *Biometrika*, **58**, 299–312.

Qiu, D., Shao, Q. and Yang, L. (2013) Efficient inference for autoregressive coefficients in the presence of trend. *J. Multiv. Anal.*, **114**, 40–53.

R Core Team (2013) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Schroeder, A. L. and Fryzlewicz, P. (2013) Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statist. Interfc.*, **6**, 449–461.

Shao, Q. and Yang, L. (2011) Autoregressive coefficient estimation in nonparametric analysis. *J. Time Ser. Anal.*, **32**, 587–597.

Shumway, R. H. and Stoffer, D. S. (2011) *Time Series Analysis and Its Applications with R Examples*, 3rd edn. New York: Springer.

Truong, Y. K. (1991) Nonparametric curve estimation with time series errors. *J. Statist. Planng Inf.*, **28**, 167–183.

Tsay, R. S. and Tiao, G. C. (1984) Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *J. Am. Statist. Ass.*, **79**, 84–96.
Wang, L. and Yang, L. (2007) Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *Ann. Statist.*, **35**, 2474–2503.
Xue, L. and Yang, L. (2006) Additive coefficient modelling via polynomial spline. *Statist. Sin.*, **16**, 1423–1446.
Yao, Q. W. and Brockwell, P. J. (2006) Gaussian maximum likelihood estimation for ARMA models: I, time series. *J. Time Ser. Anal.*, **27**, 857–875.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplement to "Oracally efficient estimation and consistent model selection for ARMA time series with trend"'.