Statistical Science
2011, Vol. 26, No. 1, 21–46
DOI: 10.1214/10-STS345

Feature Matching in Time Series Modeling

Yingcun Xia and Howell Tong

Abstract. Using a time series model to mimic an observed time series has a long history. However, with regard to this objective, conventional estimation methods for discrete-time dynamical models are frequently found to be wanting. In fact, they are characteristically misguided in at least two respects: (i) assuming that there is a true model; (ii) evaluating the efficacy of the estimation as if the postulated model is true. There are numerous examples of models, when fitted by conventional methods, that fail to capture some of the most basic global features of the data, such as cycles with good matching periods, singularities of spectral density functions (especially at the origin) and others. We argue that the shortcomings need not always be due to the model formulation but the inadequacy of the conventional fitting methods. After all, all models are wrong, but some are useful *if they are fitted properly*. The practical issue becomes one of how to best fit the model to data.

Thus, in the absence of a true model, we prefer an alternative approach to conventional model fitting that typically involves one-step-ahead prediction errors. Our primary aim is to match the joint probability distribution of the observable time series, including long-term features of the dynamics that underpin the data, such as cycles, long memory and others, rather than short-term prediction. For want of a better name, we call this specific aim *feature matching*.

The challenges of model misspecification, measurement errors and the scarcity of data are forever present in real time series modeling. In this paper, by synthesizing earlier attempts into an extended-likelihood, we develop a systematic approach to empirical time series analysis to address these challenges and to aim at achieving better feature matching. Rigorous proofs are included but relegated to the Appendix. Numerical results, based on both simulations and real data, suggest that the proposed catch-all approach has several advantages over the conventional methods, especially when the time series is short or with strong cyclical fluctuations. We conclude with listing directions that require further development.

Key words and phrases: ACF, Bayesian statistics, black-box models, blowflies, Box's dictum, calibration, catch-all approach, ecological populations, data mining, epidemiology, feature consistency, feature matching, least squares estimation, maximum likelihood, measles, measurement errors, misspecified models, model averaging, multi-step-ahead prediction, nonlinear time series, observation errors, optimal parameter, periodicity, population models, sea levels, short time series, SIR epidemiological model, skeleton, substantive models, sunspots, threshold autoregressive models, Whittle's likelihood, XT-likelihood.

Yingcun Xia is Professor of Statistics, Department of Statistics and Applied Probability, Risk Management Institute, National University of Singapore, Singapore (e-mail: staxyc@nus.edu.sg). Howell Tong is Emeritus Chair Professor of Statistics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom (e-mail: howell.tong@gmail.com).

[1]Discussed in 1012.14/11-STS345A, 1012.14/11-STS345C, 1012.14/11-STS345B and 1012.14/11-STS345D; rejoinder at 1012.14/11-STS345REJ.

1. INTRODUCTION

Dynamical models, either in continuous time or in discrete time, have been widely used to describe the changing world. Interestingly, salient features of many seemingly complex observations can sometimes be captured by simple dynamical models, as demonstrated most eloquently by Sir Isaac Newton in the seventeenth century when he used his model, Newton's law of universal gravitation, to explain Kepler's observations concerning planetary motion. In statistics, dynamical models are the *raison d'être* of time series analysis. For a time series, the dynamics transmits information about its future from observations made in the past and the present. Of particular interest are the long-term future, the periodicity and so on. To capture salient features, there are essentially two approaches: substantive and black-box. Examples of both approaches abound. The former is often preferred if available in the context in which we find ourselves. If not available, then a black-box approach might be the only choice. We shall include examples of both approaches. Let us first mention two substantive examples as they are relevant to our later discussion.

1.1 Two Substantive Models and Related Features

1. *Animal populations*. There are numerous ecological models describing the time evolution of animal populations. The single-species model of Oster and Ipaktchi (1978) can be written as

$$\frac{dx_t}{dt} = b(x_{t-\tau})x_{t-\tau} - \mu x_t, \tag{1.1}$$

where x_t is the number of adults at time t; τ is the delayed regulation duration due to the time taken for the young to develop into adults or discrete breeding seasons; $b(\cdot)$ is the birth rate; and μ is the death rate. There are different specifications for $b(\cdot)$. Gurney, Blythe and Nisbet (1980) suggested $b(u) = c\exp(-u/N_0)$, where N_0 is the reciprocal of the exponential decay rate and c is a parameter related to the reproductive rate of adults. Ellner, Seifu and Smith (2002) investigated the estimation of model (1.1) by replacing $b(x_{t-\tau})x_{t-\tau}$ and μx_t with unknown functions $B(x_{t-\tau})$ and $D(x_t)$, respectively, which they then used a nonparametric method to estimate. Wood (2001) considered a similar approach. There are several discrete-time versions of (1.1) in biology. See, for example, Varley, Gradwell and Hassell (1973). If we approximate dx_t/dt by $x_t - x_{t-1}$, then we obtain a nonlinear time series model in discrete time

$$x_t = b(x_{t-\tau})x_{t-\tau} + \nu x_{t-1}, \tag{1.2}$$

where $\nu = 1 - \mu$.

In ecology, population cycles are often observed and are an issue of paramount importance. For example, the blowfly data show a cycle of 39 days and the Canadian lynx shows a cycle of about 9.7 years. Some ecologists have even suggested chaotic patterns, although we are skeptical about this possibility. Most ecologists consider the dynamics underlying population cycles as one of the major challenges in their discipline.

2. *Transmission of infectious diseases*. The conventional compartmental SIR model partitions a community with population N into three compartments S_t (for susceptible), I_t (for infectious) and R_t (for recovered): $N = S_t + I_t + R_t$ at any time instant t. The SIR model is simple but very useful in investigating many infectious diseases including measles, mumps, rubella and SARS. Each member of the population typically progresses from susceptible to infectious to recovered or death.

Infectious diseases tend to occur in cycles of outbreaks due to the variation in the number of susceptible individuals over time. During an epidemic, the number of susceptible individuals falls rapidly as more of them are infected and thus enter the infectious and recovered compartments. The disease cannot break out again until the number of susceptible has built back up as a result of babies being born into the susceptible compartment.

Consider a population characterized by a death rate μ and a birth rate equal to the death rate, in which an infectious disease is spreading. The differential equations of the SIR model are

$$\begin{aligned}
\frac{dS_t}{dt} &= \mu(N - S_t) - \beta\frac{I_t}{N}S_t,\\
\frac{dI_t}{dt} &= \beta\frac{I_t}{N}S_t - (\nu + \mu)I_t,\\
\frac{dR_t}{dt} &= \nu I_t - \mu R_t,
\end{aligned}$$

where β is the contact rate and ν is the recovery rate of the disease. See, for example, Anderson and May (1991) and Isham and Medley (2008) for details. This model has been extensively investigated and very successfully used in the control of infectious diseases. Discrete-time versions of the model have been proposed. An example is

$$I_{t+1} = r_0 S_t I_t/N, \quad S_{t+1} = S_t - I_{t+1} + \mu N,$$

where μ is the birth rate and r_0 is the basic reproductive rate of transmission. See, for example, Bartlett (1957, 1960), Anderson and May (1991) and the discussion in Section 6.

Again, an important feature for the transmission of infectious disease is the periodicity, to understand which it is essential to understand the effect of such factors as the birth rate, the seasonal force, the transmission rate and the incubation time on the dynamics, the phase difference that is related to the transmission in different areas, and the interaction between different diseases; see, for example, Earn et al. (2000) and Rohani et al. (2003). The model can also be used to guide the policy maker in controlling the spread of the disease. See, for example, Bartlett (1957), Hethcote (1976), Keeling and Grenfell (1997) and Dye and Gay (2003).

1.2 The Objectives

Our primary concern is parametric time series modeling with the objective of achieving good *matching* of the joint probabilistic distribution of the observable time series, including, in particular, salient features, such as cycles and others. Short-term prediction is secondary in this paper. Accepting G. E. P. Box's (1976) dictum: *All models are wrong, but some are useful*, we use parametric time series models only as means to an end. We are typically less interested in the consistency of estimators of unknown parameters in the conventional sense, which is predicated on the assumed truth of the postulated model. In fact, we are more interested in improving the matching capability of the postulated model.

Suppose we postulate the following model:

$$x_t = g_\theta(x_{t-1}, \ldots, x_{t-p}) + \varepsilon_t, \tag{1.3}$$

where ε_t is the innovation and the function $g_\theta(\cdot)$ is known up to parameters θ. To indicate the dependence of x_t on θ, we also write it as $x_t(\theta)$. Following Tong (1990), we call (1.3) with $\mathrm{Var}(\varepsilon_t) = 0$ the *skeleton* of the model. In postulating the above model, we recognize that it is generally just an approximation of the true underlying dynamics no matter how the function $g_\theta(\cdot)$ is specified. Of particular note is the fact that conventional methods of estimation of θ in the present setup are usually not different from those used for a cross-sectional model: with observations $\{y_1, y_2, \ldots, y_T\}$ and postulated model g_θ, typically a loss function is based on the errors and takes the following form:

$$L(\theta) = (T-p)^{-1} \sum_{t=p+1}^{T} \{y_t - g_\theta(y_{t-1}, \ldots, y_{t-p})\}^2,$$

where, here and elsewhere, T denotes the sample size. The errors above happen to coincide with the one-step-ahead prediction errors. Under general conditions, minimizing this loss function is known mathematically to lead to efficient estimation *if the postulated model is true*. However, the postulated model is, by the Box dictum, almost invariably wrong, in which case the above loss function is not necessarily fit for purpose. To illustrate, let observations $\{y_1, y_2, \ldots, y_T\}$ be given and, of the postulated model (1.3), let the function g_θ be linear and ε_t be Gaussian with zero mean and finite variance. Let $\mathcal{T} = \{C(j), j = 0, 1, 2, \ldots, T-1\}$ denote a set of sample autocovariances of the y-data. Then minimizing $L(\theta)$ yields well-known estimates of θ that are functions of $\mathcal{S} = \{C(0), C(1), \ldots, C(p)\}$. If the postulated model is "right," then $\mathcal{S}$ is a minimal set of sufficient statistics (ignoring boundary effects) and all is well. However, if it is wrong, then it is unlikely that $\mathcal{S}$ will remain so. Since the model is typically wrong, then restricting to $\mathcal{S}$ is unfit for the purpose of estimating θ; $\mathcal{T}$ may be preferable.

To reconcile with the Box spirit, diagnostic checks, goodness-of-fit tests and other post-modeling devices are recommended. Indeed Box and Jenkins (1970) have stressed these post-modeling devices. See also Tsay (1992) for some later developments. These are undoubtedly very important developments. However, the challenge remains as to whether we can adopt the Box spirit more seriously right at the modeling stage rather than at the post-modeling stage.

It is worth recalling the fact that the classic autoregressive (AR) model of Yule (1927) and the moving average (MA) model of Slutsky (1927) were originally proposed to capture the sunspot cycle and the business cycle, respectively, rather than for the purpose of short-term prediction.

2. THE MATCHING APPROACH

We shall use the letters y and x to signify respectively the real time series under study and the time series generated by the postulated model. The adjective observable is reserved for a stochastic process. An observed time series consisting of observations constitutes (possibly part of) a realization of a stochastic process. In order for model (1.3) to be able to approximate an observable $\{y_t : t = 1, 2, \ldots\}$ well, it is natural to require throughout this paper that the state space of $\{x_t(\theta) : t = 1, 2, \ldots\}$ covers that of the observable $\{y_t : t = 1, 2, \ldots\}$. For expositional simplicity, let $p = 1$. Starting from $x_0(\theta) = y_0$, the postulated model

is said to match an observable time series under study perfectly if their conditional distributions are the same, namely,

$$P\{x_1(\theta_0) < u_1, \ldots, x_n(\theta_0) < u_n | x_0(\theta_0) = y_0\} \equiv P\{y_1 < u_1, \ldots, y_n < u_n | y_0\} \tag{2.1}$$

almost surely for some θ_0 and any n and any real values $u_1, \ldots, u_n$. We call the approach based on the above model, including all its weaker versions, some of which will be described in the next two subsections, collectively by the name *catch-all approach*.

However, formulation (2.1) is usually quite difficult to implement in practice. In the next two subsections, we suggest two weaker forms, although other forms are obviously possible.

In the econometric literature, the notion of calibration has been introduced (e.g., Kydland and Prescott, 1996). It has many alternative definitions. Broadly speaking, calibration consists of a series of steps intended to provide quantitative answers to a particular economic question. A crucial step involves some so-called "computational experiments" with a substantive model of relevance to economic theory; it is acknowledged that the model is unlikely to be the true model for the observed economic data. At the philosophical level, calibration and our feature matching share almost the same aim. However, there are some fundamental differences in methodology. Our methodology provides a statistical and *coherent* framework (in a non-Bayesian sense) to estimate *all* the parameters of a postulated (and usually wrong) model. As far as we know, calibration seems to be in need of such a framework. See, for example, Canova (2007), esp. page 239. The hope is that our methodology will be useful to substantive modelers in all fields, including ecology, economics, epidemiology and others. At the other end of the scale, it has been suggested that our methodology has potential in data mining (K.S. Chan, private communication).

2.1 Matching Up-to-m-Step-Ahead Point Predictions

If we are interested in the mean conditional on some initial observation, say y_0, we can weaken the matching requirement (2.1) to

$$E[(x_1(\theta_0), \ldots, x_m(\theta_0))|x_0(\theta_0) = y_0] \equiv E[(y_1, \ldots, y_m)|y_0],$$

where the length m of the random vector is, in practice, bounded above by the sample size under consideration. The expectation is taken with respect to the relevant joint distribution of the random vector conditional on the initial value being y_0. Since a postulated model is just an approximation of the underlying dynamics, we set θ_0 to minimize the difference of the prediction vectors, that is,

$$E\{\|E[(x_1(\theta), \ldots, x_m(\theta))|x_0(\theta) = y_0] - E[(y_1, \ldots, y_m)|y_0]\|^2\}. \tag{2.2}$$

Here, $\|\cdot\|$ denotes the Euclidean norm of a vector. In other words, we choose θ by minimizing up-to-m-step-ahead prediction errors. It is basically based on a *catch-all* idea. It is easy to see that the best θ based on minimizing (2.2) depends on m. Generally speaking, we set $m = 1$, when and only when we have complete faith in the model, which is what the conventional methods do. Denote the m-step-ahead prediction of y_{t+m} based on model (1.3) by

$$g_\theta^{[m]}(y_t) = \mathbf{E}(x_{t+m}|x_t = y_t).$$

If model (1.3) is deterministic [i.e., $\mathrm{Var}(\varepsilon_t) = 0$] or linear, $g_\theta^{[m]}(y_t)$ is simply a composite function,

$$g_\theta^{[m]}(y_t) = \underbrace{g_\theta(g_\theta(\cdots g_\theta}_{m \text{ folds}}(y_t)\cdots)).$$

Let

$$Q(y_t, x_t(\theta)) = \sup_{w_m} \sum_{m=1}^{\infty} w_m \big[\mathbf{E}\big\{y_{t+m} - g_\theta^{[m]}(y_t)\big\}^2\big], \tag{2.3}$$

where $w_m \geq 0$ and $\sum w_m = 1$. Since

$$\mathbf{E}\big[\{y_{t+m} - \mathbf{E}(y_{t+m}|y_t)\}\{\mathbf{E}(y_{t+m}|y_t) - g_\theta^{[m]}(y_t)\}\big] = 0,$$

we have

$$\mathbf{E}\big\{y_{t+m} - g_\theta^{[m]}(y_t)\big\}^2 = \mathbf{E}\{y_{t+m} - \mathbf{E}(y_{t+m}|y_t)\}^2 + \mathbf{E}\big\{\mathbf{E}(y_{t+m}|y_t) - g_\theta^{[m]}(y_t)\big\}^2.$$

Let

$$\tilde{Q}(y_t, x_t(\theta)) = \sup_{w_m} \sum_{m=1}^{\infty} w_m \big[\mathbf{E}\big\{\mathbf{E}(y_{t+m}|y_t) - g_\theta^{[m]}(y_t)\big\}^2\big].$$

If the observable y_t indeed follows the model of x_t, then $\min_\theta \tilde{Q}(y_t, x_t(\theta)) = 0$. Otherwise we generally expect $\min_\theta \tilde{Q}(y_t, x_t(\theta)) > 0$. Minimizing $\tilde{Q}(y_t,$

$x_t(\theta)$) is for $x_t(\theta)$ to arrive at a choice within the postulated model that gives all (suitably weighted) multiple-step-ahead predictions of y_t as accurately as possible in the mean squared sense.

Note that the above measure of the difference between two time series is based on a (weighted) least squares loss function. Clearly there exist many other possible measures. For example, if the distribution of the innovation is known, a likelihood type measure of the difference can be used instead. A Bayesian may perhaps then endow $\{w_m\}$ with some prior distribution. This line of development may be worth further exploration as suggested by an anonymous referee. Intuitively speaking, a J-shaped $\{w_m\}$ tends to emphasize low-pass filtering, because $\mathbf{E}(y_{t+m}|y_t)$ is a slowly varying function for sufficiently large m. Similarly, an inverted-J-shaped $\{w_m\}$ tends to emphasize high-pass filtering. An optimal choice of $\{w_m\}$ strikes a good balance between high-pass filtering and low-pass filtering.

The most commonly used estimation method in time series modeling is probably that based on minimizing the sum of squares of the errors of one-step-ahead prediction. This has been extended to the sum of squares of errors of other single-step-ahead prediction. See, for example, Cox (1961), Tiao and Xu (1993), Bhansali and Kokoszka (2002) and Chen, Yang and Hafner (2004). Clearly, the former method is predicated on the model being *true*. The latter extension recognizes that this is an unrealistic assumption for multi-step-ahead prediction. Instead, a *panel of models* is constructed so that a different model is used for the prediction at each different horizon. The focus of the extension is prediction.

The approach that we develop here essentially builds on the above extension. First, we shift the focus away from prediction. Second, we transform the prediction based on a *panel of models* into the fitting of a *single time series model*. We effectively synthesize the panel into a catch-all methodology. Specifically, we propose to minimize the sum of squares of errors of prediction *over all* (*allowable*) *steps ahead*, as given in (2.3). We stress again that our primary objective is feature matching rather than prediction. Of course, it is conceivable that good feature matching may sometimes lead to better prediction, especially for the medium and long term. Clearly each member of the panel can be recovered, at least formally, from the catch-all setup by setting, in turn, the weight, w_j, to unity, leaving the rest to zero.

2.2 Matching ACFs

Suppose that the observable $\{y_t\}$ and $\{x_t(\theta)\}$ are both second-order stationary. If we are interested in second-order moments, then a weaker form of (2.1) is the following difference or distance function:

$$D_C(y_t, x_t(\theta)) = \sup_{\{w_m\}} \sum_{m=0}^{\infty} w_m \{\gamma_{x(\theta)}(m) - \gamma_y(m)\}^2.$$

Here, the suffixes of y and $x(\theta)$ are self-explanatory. We assume that the spectral density function (SDF) of the observable y_t exists; it is given by

$$f_y(\omega) = \frac{1}{2\pi}\gamma(0) + \frac{1}{\pi}\sum_{k=1}^{\infty}\gamma_y(k)\cos(k\omega).$$

The SDF of $x_t(\theta)$, which we also assume to exist, can be defined similarly. We can also measure the difference between two time series by reference to the difference between their SDFs, for example,

$$D_F(y_t, x_t(\theta)) = \int_{-\pi}^{\pi}\left\{\frac{f_y(\omega)}{f_x(\omega)} + \log\left(\frac{f_x(\omega)}{f_y(\omega)}\right) - 1\right\} d\omega,$$

which is called the Itakura–Saito distortion measure; see also Whittle (1962). Further discussion on measuring the difference between two SDFs can be found in Georgiou (2007).

Suppose that $\{x_t(\theta)\}$ and the observable $\{y_t\}$ have the same marginal distribution and they each have second-order moments. Then we can prove that

$$D_C(y_t, x_t(\theta)) \le C_1 \tilde{Q}(y_t, x_t(\theta)),$$
$$D_F(y_t, x_t(\theta)) \le C_2 \tilde{Q}(y_t, x_t(\theta))$$

for some positive constants C_1 and C_2. Moreover, if $\{x_t(\theta)\}$ and the observable $\{y_t\}$ are linear AR models, then there are some positive constants C_3 and C_4 such that

$$\tilde{Q}(y_t, x_t(\theta)) \le C_3 D_C(y_t, x_t(\theta)),$$
$$\tilde{Q}(y_t, x_t(\theta)) \le C_4 D_F(y_t, x_t(\theta)).$$

For further details, see Theorem A in the Appendix.

For linear AR models under the above setup, $\tilde{Q}(\cdot,\cdot)$, $D_C(\cdot,\cdot)$ and $D_F(\cdot,\cdot)$ are equivalent. However, the equivalence is not generally true. A counterexample can be constructed easily by reference to the classic random telegraph signal process. [See, e.g., Parzen (1962), page 115].

Let us close this section by describing one way of implementing the ACF criterion for an ARMA model

with normal innovation. Suppose $y_1, \ldots, y_T$ are observations from the observable $\{y_t\}$. Whittle (1962) considered a "likelihood function" for ARMA models in terms of the SDF. Let

$$I(w) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} y_t \exp(-\iota\omega t)\right|^2$$

be the periodogram of the sample, where ι is the imaginary unit. Let $f_\theta(\omega)$ be the theoretical SDF of an ARMA model with parameters θ. Whittle (1962) proposed to estimate θ by

$$\hat{\theta} = \min_{\theta} \sum_{j=1}^{T}\left\{\frac{I(\omega_j)}{f_\theta(\omega_j)} + \log(f_\theta(\omega_j))\right\},$$

where $\omega_j = 2\pi j/T$. From the perspective of feature matching, the celebrated Whittle's likelihood is not a conventional likelihood but a precursor of the extended-likelihood approach. It matches the second-order moments, by using a natural sample version of $D_F(y_t, x_t(\theta))$ up to a constant. For this reason, it is expected that for misspecified models, Whittle's estimator can lead to better matching of the ACFs of the observed time series than the innovation driven methods [e.g., the least squares estimation (LSE) or the maximum likelihood estimation (MLE)]. We shall give some numerical comparison between Whittle's estimator and the others in Sections 5 and 6 below.

3. TIME SERIES WITH MEASUREMENT ERRORS

To illustrate the advantages of the catch-all approach, which involves minimal assumptions on the observed time series, we give detailed analyses of two cases involving measurement errors, one of which is related to a linear AR(p) model and the other a nonlinear skeleton model. They can be considered special cases of model misspecification in that the observable y-time series is a measured version of the x-time series subject to measurement errors. For the linear case, measurement error is an old problem in time series analysis that was studied at least as early as Walker (1960). Some new lights will be shed.

3.1 Linear AR(p) Models

Consider the following AR(p) model:

$$x_t = \theta_1 x_{t-1} + \cdots + \theta_p x_{t-p} + \varepsilon_t. \tag{3.1}$$

Stationarity is assumed. By the Yule–Walker equations, we have the recursive formula for the ACF, $\{\gamma(j)\}$, of the x-time series, namely,

$$\begin{aligned}\gamma(k) &= \gamma(k-1)\theta_1 + \gamma(k-2)\theta_2 + \cdots \\ &\quad + \gamma(k-p)\theta_p, \quad k = 1, 2, \ldots.\end{aligned} \tag{3.2}$$

Let $m \geq p$ and $\Upsilon_m = (\gamma(1), \gamma(2), \ldots, \gamma(m))^\top$, $\theta = (\theta_1, \ldots, \theta_p)^\top$ and

$$\Gamma_m = \begin{pmatrix} \gamma(0) & \gamma(-1) & \cdots & \gamma(-p+1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(-p+2) \\ \vdots & & & \\ \gamma(m-1) & \gamma(m-2) & \cdots & \gamma(m-p) \end{pmatrix}.$$

The Yule–Walker equations can be written as

$$\Gamma_m \theta = \Upsilon_m.$$

Suppose that the observable y-time series is given by $y_t = x_t + \eta_t$, for $t = 0, 1, 2, \ldots$, where $\{\eta_t\}$ is independent of $\{x_t\}$ and is a sequence of independent and identically distributed random variables each with zero mean and finite variance. Clearly, $\{y_t\}$ is no longer given by an AR(p) model of the form (3.1).

Let $\{\tilde{\gamma}(j)\}$ denote the ACF of the observable y-time series. Let $\tilde{\Gamma}_m$ and $\tilde{\Upsilon}_m$ denote the analogously defined matrix and vector of ACFs for the observable y-time series.

Suppose we are now given the observations $\{y_1, y_2, \ldots, y_T\}$, and we wish to fit the wrong model of the form (3.1) to them. We may estimate $\tilde{\gamma}(j)$ by $\hat{\gamma}(j) = \hat{\gamma}(-j) = T^{-1}\sum_{t=1}^{T-j}(y_t - \bar{y})(y_{t+j} - \bar{y})$, $\bar{y}$ being the sample mean. Let $\hat{\Gamma}_m$ and $\hat{\Upsilon}_m$ denote the obvious sample version of $\tilde{\Gamma}_m$ and sample version of $\tilde{\Upsilon}_m$, respectively.

Since any p equations can be used to determine the parameters, the Yule–Walker estimators typically use the first p equations, that is,

$$\hat{\theta} = \hat{\Gamma}_p^{-1}\hat{\Upsilon}_p \quad \text{or} \quad \hat{\theta} = (\hat{\Gamma}_p^\top \hat{\Gamma}_p)^{-1}\hat{\Gamma}_p^\top \hat{\Upsilon}_p,$$

which is also the minimizer of $\sum_{k=1}^{p}\{\hat{\gamma}(k) - \hat{\gamma}(k-1)\theta_1 - \hat{\gamma}(k-2)\theta_2 - \cdots - \hat{\gamma}(k-p)\theta_p\}^2$, involving the ACF only up to lag p. We can achieve closer matching of the ACF by incorporating lags beyond p as well. For example, we may consider estimating θ by minimizing

$$\begin{aligned}&\sum_{k=1}^{m}\{\hat{\gamma}(k) - \hat{\gamma}(k-1)\theta_1 \\ &\qquad - \hat{\gamma}(k-2)\theta_2 - \cdots - \hat{\gamma}(k-p)\theta_p\}^2, \\ &\qquad\qquad m \geq p.\end{aligned}$$

Denoting the minimizer by $\hat{\theta}_{\{m\}}$, we have

$$\begin{aligned}\hat{\theta}_{\{m\}} &= (\hat{\Gamma}_m^\top \hat{\Gamma}_m)^{-1} \hat{\Gamma}_m^\top \hat{\Upsilon}_m \\ &= \left\{ \sum_{k=0}^{m} \breve{\Upsilon}_k \breve{\Upsilon}_k^\top \right\}^{-1} \sum_{k=0}^{m} \breve{\Upsilon}_k \hat{\gamma}(k+1),\end{aligned} \tag{3.3}$$

where $\breve{\Upsilon}_k = (\hat{\gamma}(k), \hat{\gamma}(k+1), \ldots, \hat{\gamma}(k+p-1))^\top$. Let us call the estimator $\hat{\theta}_{\{m\}}$ the up-to-lag-m Yule–Walker estimator (or AYW($\leq m$)). For the error-free case, that is, $\eta_t = 0$ with probability 1, it is easy to see that $\hat{\theta}_{\{p\}}$ is the most efficient amongst all $\hat{\theta}_{\{m\}}$, $m = p, p+1, \ldots$. Otherwise, under some regularity conditions, we have in distribution

$$\sqrt{n}\{\hat{\theta}_{\{m\}} - \vartheta\} \to N(0, \tilde{\Sigma}_m),$$

where $\vartheta = (\tilde{\Gamma}_m^\top \tilde{\Gamma}_m)^{-1} \tilde{\Gamma}_m^\top \tilde{\Upsilon}_m$ and $\tilde{\Sigma}_m$ is a positive definite matrix. For $\mathrm{Var}(\varepsilon_t) > 0$ and $\mathrm{Var}(\eta_t) = \sigma_\eta^2 > 0$, the above asymptotic result holds with $\vartheta = \theta + \sigma_\eta^2(\Gamma_m^\top \Gamma_m + 2\sigma_\eta^2 \Gamma_p + \sigma_\eta^4 I)^{-1}(\Gamma_p + \sigma_\eta^2 I)\theta$. For further details, see Theorem B in the Appendix.

Clearly the bias $\sigma_\eta^2(\Gamma_m^\top \Gamma_m + 2\sigma_\eta^2 \Gamma_p + \sigma_\eta^4 I)^{-1}(\Gamma_p + \sigma_\eta^2 I)\theta$ in the estimator will be smaller when m is larger. For sufficiently large sample size, the smaller bias can lead to higher efficiency in the sense of mean squared errors (MSE). Let $\bar{\Upsilon}_k = (\gamma(k), \gamma(k+1), \ldots, \gamma(k+p-1))^\top$. Then

$$\Gamma_m^\top \Gamma_m = \Gamma_p^\top \Gamma_p + \sum_{k=p}^{m} \bar{\Upsilon}_k \bar{\Upsilon}_k^\top.$$

Thus, the bias can be reduced more substantially if the ACF decays very slowly and a larger m is used. For example, a highly cyclical time series usually has slowly decaying ACF, in which case the AYW will provide a substantial improvement over the Yule–Walker estimators. However, even with the ACF slowly decaying, a large m may cause larger variability of the estimator. Therefore, a good choice of m is also important in practice. We shall return to this issue later.

In fact, Walker (1960) suggested using exactly p equations to estimate the coefficients giving

$$\hat{\theta}_{W.\ell} = \arg\min_\theta \left\{ \sum_{k=p+\ell}^{2p-1+\ell} \breve{\Upsilon}_k \breve{\Upsilon}_k^\top \right\}^{-1} \sum_{k=p+\ell}^{2p-1+\ell} \breve{\Upsilon}_k \hat{\gamma}(k+1).$$

Note the difference between AYW and $\hat{\theta}_{W.\ell}$. Walker (1960) showed that in the presence of measurement error, then $\ell = p$ is the optimal choice amongst all candidates with $\ell \geq p$, by reference to MSE. However, Walker's method seems counterintuitive because it relies on the sample ACF at higher lags to a greater extent than those at the lower lags. Further discussion on Walker's method can be found in Sakai, Soeda and Tokumaru (1979) and Staudenmayer and Buonaccorsi (2005). It is well known that an autoregressive model plus independent additive white noise results in an ARMA model. Walker's approach essentially treats the resulting ARMA model as a true model. This approach has attracted attention in the engineering literature. See, for example, Friedlander and Sharman (1985) and Stoica, Moses and Li (1991). The essential difference between this approach and the catch-all approach is that the latter postulates an autoregressive model to match the observations. And we know that it is a wrong model, as we consistently do with all postulated models. Note that the use of sample ACFs at all possible lags has points of contact with the so-called generalized method of moments, used extensively in econometrics. See, for example, Hall (2005).

Next, we consider estimation based on $Q(\cdot, \cdot)$. Given a finite sample size, we may stop at, say, the m-step-ahead prediction. Let $e_1 = (1, 0, \ldots, 0)^\top$ and

$$\Phi = \begin{pmatrix} \theta_1 & \theta_2 & \cdots & \theta_{p-1} & \theta_p \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

We estimate θ by

$$\begin{aligned}\tilde{\theta}_{\{m\}} = \arg\min_\theta \sum_{t=p+1}^{T} \sum_{k=1}^{m} w_k \{ y_{t-1+k} \\ - e_1^\top \Phi^k (y_{t-1}, \ldots, y_{t-p})^\top \}^2,\end{aligned} \tag{3.4}$$

where w_k is a weight function, typically positive definite. A reasonable choice of w_k is the absolute value of the autocorrelation function of the observed time series, that is, $w_k = |r_y(k)|$. We call $\tilde{\theta}_{\{m\}}$ in (3.4) the up-to-m-step-ahead prediction estimator [APE or APE($\leq m$)].

The asymptotic properties of $\tilde{\theta}_{\{m\}}$ will be discussed later.

3.2 Nonlinear Skeletons

A deterministic nonlinear dynamic model with measurement error is commonly used in many applied areas, for example, ecology, dynamical systems and others. See, for example, May (1976), Gurney, Blythe and Nisbet (1980), Tong (1990), Anderson and May (1991), Alligood, Sauer and Yorke (1997), Grenfell, Bjørnstad and Finkenstädt (2002), Chan and Tong

(2001) and the examples in Section 6. Consider using the following nonlinear skeleton:

$$x_t = g_\theta(x_{t-1}, \ldots, x_{t-p}) \tag{3.5}$$

to match the observable time series $\{y_t\}$.

Employing the $Q(\cdot,\cdot)$ criterion, the estimator is given by

$$\tilde{\theta}_{\{m\}} = \arg\min_\theta \sum_{t=p+1}^{T} \sum_{k=1}^{m} w_k \{y_{t-1+k} - g_\theta^{[m]}(y_{t-1}, \ldots, y_{t-p})\}^2, \tag{3.6}$$

which we again call the up-to-m-step-ahead prediction estimator [APE or APE($\le m$)]. Here the weight function $\{w_k\}$ is as defined in (2.3).

For ease of explanation, we consider again $y_t = x_t + \eta_t$ and $p = 1$. Starting from any state $\tilde{x}_0 = x_0$, let $\tilde{x}_t = g_\theta^{[m]}(x_0)$. Suppose the dynamical system has a negative Lyapunov exponent

$$\lambda_\theta(x_0) = \lim_{n\to\infty} n^{-1} \sum_{t=0}^{n-1} \log(|g'_\theta(\tilde{x}_t)|) < 0,$$

for all states x_0. Similarly starting from x_t let $x_{t+m} = g_{\theta_0}^{[m]}(x_t)$. We predict x_{t+m} by $\hat{y}_{t+m} = g_\theta^{[m]}(y_t)$. By the definition of the Lyapunov exponent, we have

$$|g_\theta^{[m]}(x_t + \eta_t) - g_\theta^{[m]}(x_t)| \approx \exp\{m\lambda_\theta(x_t)\}|\eta_t|.$$

More generally, suppose the system $x_t = g_{\theta_0}(x_{t-1}, \ldots, x_{t-p})$ has a finite-dimensional state space and admits only limit cycles, but x_t is observed as $y_t = x_t + \eta_t$, where $\{\eta_t\}$ are independent with mean 0. Suppose that the function $g_\theta(v_1, \ldots, v_p)$ has bounded derivatives in both θ in the parameter space Θ and $v_1, \ldots, v_p$ in a neighborhood of the state space. Suppose that the system $z_t = g_\theta(z_{t-1}, \ldots, z_{t-p})$ has only negative Lyapunov exponents in a small neighborhood of $\{x_t\}$ and in $\theta \in \Theta$. Let $X_t = (x_t, x_{t-1}, \ldots, x_{t-p})$ and $Y_t = (y_t, y_{t-1}, \ldots, y_{t-p})$. If the observed $Y_0 = X_0 + (\eta_0, \eta_{-1}, \ldots, \eta_{-p})$ is taken as the initial values of $\{x_t\}$, then for any n,

$$\begin{aligned} &f(y_{m+1}, \ldots, y_{m+n}|X_0) \\ &\quad - f(y_{m+1}|X_0 = Y_0) \\ &\quad \cdots f(y_{m+n}|X_0 = Y_0) \to 0 \end{aligned} \tag{3.7}$$

as $m \to \infty$. Suppose the equation $\sum_{X_{t-1}} \{g_\theta(X_{t-1}) - x_t\}^2 = 0$ has a unique solution in θ, where the summation is taken over all limiting states. Let $\theta_{\{m\}} = \arg\min_\theta m^{-1} \sum_{k=1}^m \mathbf{E}\{y_{t-1+k} - g_\theta^{[k]}(Y_{t-1})\}^2$. If the noise takes value in a small neighborhood of the origin, then

$$\theta_{\{m\}} \to \theta_0 \quad \text{as } m \to \infty.$$

Note that $|f(y_1|X_0) - f(y_1|X_0 = Y_0)| \neq 0$ implies that

$$\begin{aligned} &f(y_1, \ldots, y_n|X_0 = Y_0) \\ &\quad \neq f(y_1|X_0 = Y_0) f(y_2|X_1 = Y_1) \\ &\quad \cdots f(y_n|X_{n-1} = Y_{n-1}), \end{aligned}$$

which challenges the commonly used (conditional) MLE. Equation (3.7) indicates that using high step-ahead prediction can reduce the effect of noisy data (e.g., due to measurement errors), and provide a better approximation of the conditional distribution. The second part suggests that using high step-ahead prediction errors in a criterion can reduce the bias caused by the presence of η_t. It also implies that any set of past values, for example, $(y_{t-1}, \ldots, y_{t-p})$ for $t > p$, can offer us an estimator with the first summation in (3.6) removed. However, the summation over all past values is more efficient statistically. For further details, see Theorem C in the Appendix.

There are other interesting special cases. For example, when the postulated model has a chaotic skeleton, the initial values play a crucial role. One approach is to treat the initial values as unknown parameters. See, for example, Chan and Tong (2001) for more details. Another example is when the postulated model is nonlinear, and is driven by nonadditive white noise with an unknown distribution. Here, the exact least squares multi-step-ahead prediction is quite difficult to obtain theoretically and time consuming to calculate numerically; see, for example, Guo, Bai and An (1999). In this case, the up-to-m-step-ahead prediction method is difficult to implement directly. However, our simulations suggest that approximating the multi-step-ahead prediction by its skeleton is sometimes helpful in feature matching, especially when the observed time series is quite cyclical (Chan, Tong and Stenseth, 2009).

4. ISSUES OF THE ESTIMATION METHOD

We now turn to some theoretical issues and calculation problems. In conventional statistical theory for parameter estimation, by consistency is generally meant that the estimated parameter vector converges to the true parameter vector in some sense as the sample size tends to infinity. The postulated model is assumed to be the true model in the above conventional approach.

In the absence of a true model and *ipso facto* true parameter vector, we propose an alternative definition

of consistency. Specifically, by consistency we mean that the estimated parameter vector will, in some sense, tend to the optimal parameter vector that represents the best achievable feature matching of the postulated model to the observable time series. To be more precise, for some positive integer m (which may be infinite), we define the optimal parameter by

$$\vartheta_{m,\mathbf{w}} = \arg\min_{\theta} \sum_{k=1}^{m} w_k \mathbf{E}[y_{t+k} - \mathbf{E}\{x_{t+k}(\theta)|X_t(\theta) = Y_t\}]^2,$$

where $X_t(\theta) = (x_t(\theta), \ldots, x_{t-p+1}(\theta))$ and $\{w_k\}$ defines the weight function, typically positive and summing to unity. For ease of exposition, we assume that the solution to the above minimization is unique. Now, we say that an estimator is *feature-consistent* if it converges to $\vartheta_{m,\mathbf{w}}$ in probability as the sample size tends to infinity. It is easy to prove that under some regularity conditions, $\tilde{\theta}_{\{m\}}$ is asymptotically normal, that is,

$$T^{-1/2}\big(\tilde{\theta}_{\{m\}} - \vartheta_{m,\mathbf{w}}\big) \xrightarrow{D} N(0, \Omega)$$

for some positive definite matrix Ω. For further details, see Theorem D in the Appendix.

The optimal parameter depends on m and the weight function w_k. As discussed in Section 3.1, when the autocorrelation decays less slowly, we should consider using a larger m. Alternatively, we can consider assigning heavier weights for larger k. Our experience suggests that, for a postulated linear time series model, w_k can be selected as the absolute value of the sample ACF function. For a postulated nonlinear time series model aiming to match possibly high degrees of periodicity, w_k can be chosen as constant lasting for approximately one, two or three periods. Note that by setting $w_1 = 1$ and all other w_j's zero, the estimation is equivalent to the LSE, and the MLE in the case of exponential family of distributions.

The above feature suggests that we may regard $\tilde{\theta}_{\{m\}}$ as a *maximum extended-likelihood estimator* and functions such as $\sum_{t=p+1}^{T}\sum_{k=1}^{m} w_k\{y_{t-1+k} - e_1^\top \Phi^k(y_{t-1}, \ldots, y_{t-p})^\top\}^2$ or their equivalents as *extended-likelihoods* (or *XT-likelihoods* for short), with Whittle's likelihood as a precursor. An XT-likelihood carries with it the interpretation as a weighted average of likelihoods of a cluster of models around the postulated model. In this sense, it is related to Akaike's notion of the likelihood of a model (Akaike, 1978).

For the numerical calculation involved in (3.4) and (3.6), the gradient and the Hessian matrix of the loss function can be obtained recursively for different steps of prediction. Consider (3.6) as an example. Let $g_\theta^{[m]}$ stand for $g_\theta^{[m]}(y_{t-1}, \ldots, y_{t-p})$ and write $g_\theta(v_1, \ldots, v_p)$ as $g(v_1, \ldots, v_p, \theta_1, \ldots, \theta_q)$. Let $g_\theta^{[0]} = y_{t-1}, \ldots, g_\theta^{[-p+1]} = y_{t-p}$, $\partial g_\theta^{[m]}/\partial\theta_k = 0$ and $\partial^2 g_\theta^{[m]}/(\partial\theta_k\partial\theta_\ell) = 0, k, \ell = 1, \ldots, q$ if $m \le 0$. Then for $m \ge 1$,

$$g_\theta^{[m]} = g\big(g_\theta^{[m-1]}, \ldots, g_\theta^{[m-p]}, \theta_1, \ldots, \theta_q\big)$$

and

$$\frac{\partial g_\theta^{[m]}}{\partial\theta_k} = \sum_{i=1}^{p} \dot{g}_i \frac{\partial g_\theta^{[m-i]}}{\partial\theta_k} + \dot{g}_{p+k}\big(g_\theta^{[m-1]}, \ldots, g_\theta^{[m-p]}, \theta_1, \ldots, \theta_q\big), \quad k = 1, \ldots, q,$$

where $\dot{g}_k(v_1, \ldots, v_p, \ldots, v_{p+q}) = \partial g(v_1, \ldots, v_p, \ldots, v_{p+q})/\partial v_k, k = 1, \ldots, p+q$, and

$$\begin{aligned}\frac{\partial^2 g_\theta^{[m]}}{\partial\theta_k\,\partial\theta_\ell} &= \sum_{i=1}^{p}\sum_{j=1}^{p} \ddot{g}_{i,j} \frac{\partial g_\theta^{[m-i]}}{\partial\theta_k}\frac{\partial g_\theta^{[m-j]}}{\partial\theta_\ell} + \sum_{i=1}^{p} \dot{g}_i \frac{\partial^2 g_\theta^{[m-i]}}{\partial\theta_k\,\partial\theta_\ell} \\ &\quad + \sum_{i=1}^{p} \ddot{g}_{p+k,i}\big(g_\theta^{[m-1]}, \ldots, g_\theta^{[m-p]}, \theta_1, \ldots, \theta_q\big)\frac{\partial g_\theta^{[m-i]}}{\partial\theta_\ell} \\ &\quad + \ddot{g}_{p+k,p+\ell}\big(g_\theta^{[m-1]}, \ldots, g_\theta^{[m-p]}, \theta_1, \ldots, \theta_q\big),\end{aligned}$$

where $\ddot{g}_{k,\ell}(v_1, \ldots, v_p, \ldots, v_{p+q}) = \partial^2 g(v_1, \ldots, v_p, \ldots, v_{p+q})/(\partial v_k \partial v_\ell)$ for $k, \ell = 1, \ldots, p+q$. The Newton–Raphson method can then be used for the minimization.

5. SIMULATION STUDY

There are many different ways to measure the goodness of matching the observed by the postulated model, depending on the features of interest. We suggest two here. (1) The ACFs are clearly important features in the context of linear time series, and relevant even for nonlinear time series analysis. Therefore, a natural measure can be based on the differences of the ACFs, for example,

$$\left[\sum_{k=0}^{N}\{r_y(k) - r_x(k)\}^2/N\right]^{1/2} \tag{5.1}$$

for some N, sufficiently large or even infinite, where $r_y(k)$ and $r_x(k)$ are the theoretical ACFs (if available) or sample ACFs. Clearly, we can use other distances to measure the differences of the ACFs. (2) For highly cyclical $\{y_t\}$, we can measure the differences between the observed and the attractor (i.e., the limiting state) generated by the skeleton of postulated model, after allowing for possible phase shifts. Thus, we can use the following quasi-sample-path measure:

$$\min_k \sum_{t=1}^{T} |y_t - x_{t+k}|/T, \tag{5.2}$$

where T is the sample size as before.

To check the efficacy of estimation of parameters, especially in a simulation study, we can use an obvious measure: $\{(\hat{\theta}-\theta)^\top(\hat{\theta}-\theta)/p\}^{1/2}$ for any estimator $\hat{\theta}$ of $\theta = (\theta_1, \ldots, \theta_p)^\top$. Obviously, it is a function of the number of steps m in APE($\leq m$) or AYW($\leq m$). Note $m = 1$ corresponds to the commonly used estimation method based on the least squares, or the maximum likelihood when normality is assumed. Note that the MLE is also based on the one-step-ahead prediction for dynamical models that are driven by Gaussian white noise. In our plotting below, results for APE(≤ 1) and AYW(≤ 1) are not marked separately from those for APE($\leq m$) and AYW($\leq m$) with $m > 1$.

EXAMPLE 5.1 (Model misspecification). We postulate an AR(p) model to match data generated by fractionally integrated noise $(1-B)^d y_t = \varepsilon_t$, where $0.5 > d > -0.5$ and B is the back-shift operator and $\{\varepsilon_t\}$ are i.i.d. $N(0,1)$. The process is stationary, but has long-memory when $0.5 > d > 0$. The closer is d to 0.5, the longer is the memory. For the use of low-order ARMA models for short-term prediction of this type of long-memory model, see, for example, Man (2002). Any AR(p) model with finite p is a "wrong" model for the process. In the following analysis, the order p is assumed unknown and determined by AIC.

The simulation results shown in Figure 1 are based on 2,000 replications. We have the following observations. (1) With a misspecified model, the APE($\leq m$) and the AYW($\leq m$) with $m > 1$ show better matching of the ACFs than the APE(≤ 1) and AYW(≤ 1). When d is closer to 0.5, the AR model is less likely to fit the data well, thus necessitating a larger m. (2) When the autocorrelation is not strong, which is the case with d being close to zero, the AYW with large m shows better matching of the ACF than the APE; otherwise APE shows better matching. It is interesting to note that although APE does not target the ACF directly, it

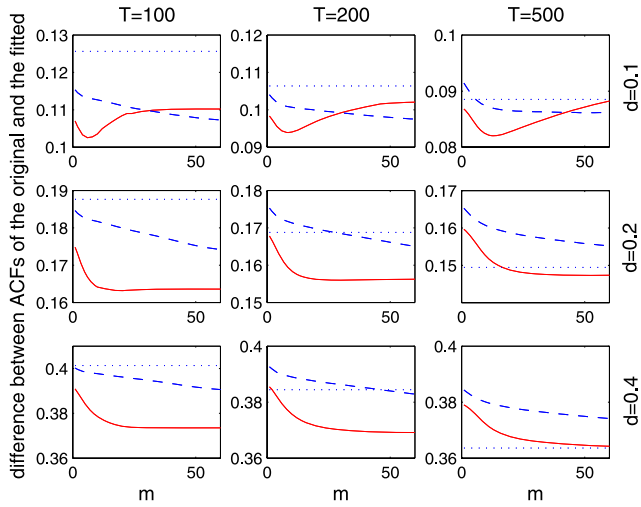

FIG. 1. *Simulation results for Example 5.1 with different sample size T, index d and the number of steps m in* AYW($\leq m$) *or* APE($\leq m$). *In each panel, the dotted line, the solid line and the dashed line correspond to the Whittle estimator, the APE and the AYW, respectively.*

can match the ACF well in comparison with the AYW. (3) For small sample size or when d is not so close to 0.5, the APE($\leq m$) with $m > 1$ show better matching than the Whittle estimator; otherwise the Whittle estimator shows better matching.

EXAMPLE 5.2 (State–space model). Consider the AR(4) model with observation errors

$$x_t = \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \beta_4 x_{t-4} + \varepsilon_t,$$
$$y_t = x_t + \eta_t.$$

This is also a special case of a state–space model. The estimation of the state model is of interest and has attracted considerable attention. See, for example, Durbin and Koopman (2001) and Staudenmayer and Buonaccorsi (2005).

To cover as widely as possible all admissible values on the parameter space, we choose $\beta_1, \beta_2, \beta_3$ and β_4 uniformly distributed in the stationary region. In the model, $\{\varepsilon_t\}$ is a sequence of independently and identically distributed random variables, each with a unit normal distribution, or i.i.d. $N(0,1)$ for short; $\{\eta_t\}$ is i.i.d. $N(0, \sigma_\eta^2)$, such that the signal-noise ratio $\sigma_\eta^2/\operatorname{Var}(y_t) = sn$ is fixed. Again, we run the simulation 2,000 times. The results are summarized in Figures 2 and 3. When p is known, Figure 2 suggests that APE($\leq m$) and AYW($\leq m$) with $m > 1$ can usually produce models that better match the dynamics of the hidden state time series $\{x_t\}$ than APE(≤ 1)

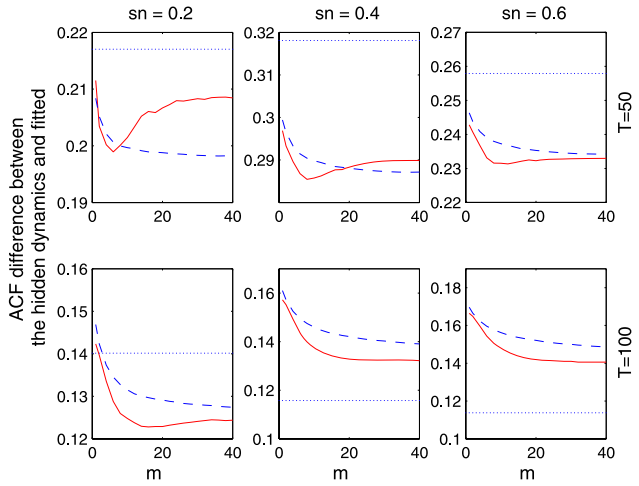

FIG. 2. *Results for Example 5.2 when the order $p = 4$ is known. In each panel, the dotted line, the solid line and the dashed line correspond respectively to the Kalman filter, the* APE($\leq m$) *and the* AYW($\leq m$) *over different m.*

and AYW(≤ 1). When p is selected by AIC, Figure 3 suggests that APE($\leq m$) and AYW($\leq m$) with $m > 1$ can still lead to better matching than APE(≤ 1) and AYW(≤ 1).

To compare with the Kalman filter approach which utilizes the maximum likelihood method or other methods such as the EM algorithm, we apply the R package "dlm" kindly provided by Professor Giovanni Petris. The results are shown by dotted lines in Figure 2. When the order is known, the Kalman filter shows good performance in estimating the coefficients and in matching the ACF, but it shows very unstable performance

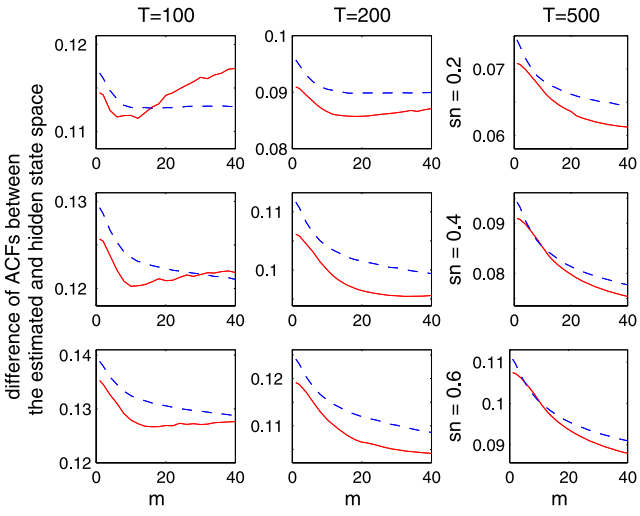

FIG. 3. *Results for Example 5.2 when the order p is selected by AIC. In each panel, the solid line is for* APE($\leq m$) *and the dashed line is for* AYW($\leq m$).

when the sample size is small. Even worse, if the order is selected by the AIC, the Kalman filter appears to be incapable of producing reasonable matching, so much so that the results are outside the range in Figure 3 in the wrong direction.

EXAMPLE 5.3 (Nonlinear time series model 1: smooth model). Consider the simple nonlinear model

$$x_t = b_1 x_{t-1} + b_2 x_{t-1}^2 + \sigma_0 \varepsilon_t;$$
$$y_t = x_t + \sigma_1 \eta_t$$

with parameters $b_1 = 3.2$ and $b_2 = -0.2$; both ε_t and η_t are i.i.d. $N(0, 1)$ but ε_t is truncated to lie in $[-4, 4]$. We replicate our simulation 1,000 times for each set of variances σ_0^2 and σ_1^2. The matching results are shown in Figure 4.

By coping well with noisy data due to $\sigma_1 \eta$, APE($m > 1$) demonstrates substantial improvement on the parameter estimation (in panel 1 of Figure 4), the ACF-matching of the hidden time series x_t (panel 2 of Figure 4) and the ACF-matching of the observed time series (in panel 3 of Figure 4). It is not surprising that when the model is perfectly specified (i.e., $\sigma_1 = 0$), the APE(≤ 1) can provide better performance than APE($\leq m$) with $m > 1$ in terms of the parameter estimation and the ACF-matching; see panels 4–5 of Figure 4. However, APE($\leq m$) with $m > 1$ is still useful in matching features of the observed time series as shown

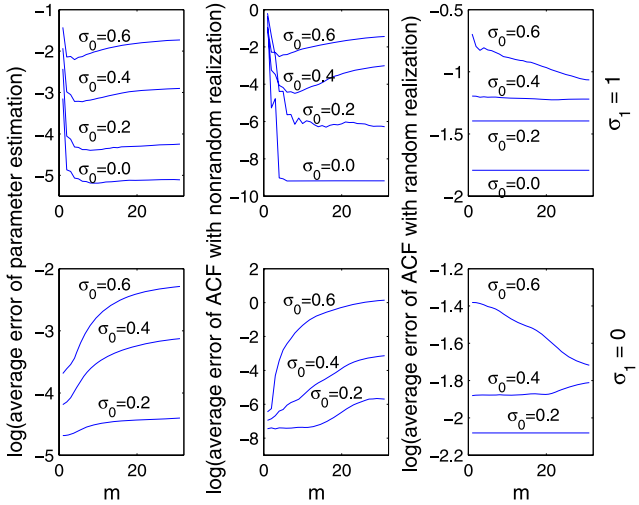

FIG. 4. *Results for Example 5.3 with $T = 50$ and different σ_0 and σ_1. The first panel is the estimation error of (b_1, b_2) with $\sigma_1 = 1$; the second panel is the difference of ACFs between the matching skeleton and the true ACF with $\sigma_1 = 1$; the third panel is the difference of ACFs between the matching skeleton and the estimated ACFs based on random realizations with $\sigma_1 = 1$. Panels 4–6 are respectively the corresponding results of panels 1–3 but with $\sigma_0 = 0$.*

TABLE 1
The simulation results for Example 5.4

Model setting	Method	Matching error	Cycle periods	Frequency of correct periods (%)
$T = 50$, $d = 2$,	APE(≤ 1)	2.1352 (1.0334)	5.3806 (0.6301)	31
period $= 6$	APE(≤ 50)	0.8523 (0.6591)	5.8629 (0.5141)	92
$T = 50$, $d = 3$,	APE(≤ 1)	2.5301 (1.6729)	9.4839 (0.5824)	34
period $= 10$	APE(≤ 50)	1.3987 (0.8180)	9.9340 (0.1472)	66
$T = 100$, $d = 2$	APE(≤ 1)	1.5260 (1.0643)	5.5884 (0.6912)	57
period $= 6$	APE(≤ 50)	0.6471 (0.5301)	5.9180 (0.3940)	95
$T = 100$, $d = 3$	APE(≤ 1)	2.7196 (1.6411)	9.4005 (0.6224)	34
period $= 10$	APE(≤ 50)	1.1502 (0.5133)	9.9705 (0.0770)	78

in the last panel. Our results suggest that APE($\leq m$) with $m > 1$ leads to less improvement over APE(≤ 1) when σ_0 (for the dynamic noise) is larger but greater improvement when σ_1 (for the observation noise) is larger.

EXAMPLE 5.4 (Nonlinear time series model 2: SETAR model). Now, we consider a self-exciting threshold autoregressive model (SETAR model) with skeleton

$$x_t = \begin{cases} a_0 + b_0 x_{t-1}, & \text{if } x_{t-d} \leq c, \\ a_1 + b_1 x_{t-1}, & \text{if } x_{t-d} > c, \end{cases}$$

where parameters $a_0 = 3, b_0 = 1, a_1 = -3, b_1 = 1$ and $c = 0$. A realization is shown in the first panel of Figure 5. It reveals a period of 6 when $d = 2$, and 10 (not shown) when $d = 3$. Suppose that we observe $y_t = x_t + \eta_t$, where $\{\eta_t\}$ are i.i.d. $N(0, 1)$. A typical realization is also shown in the second panel of Figure 5.

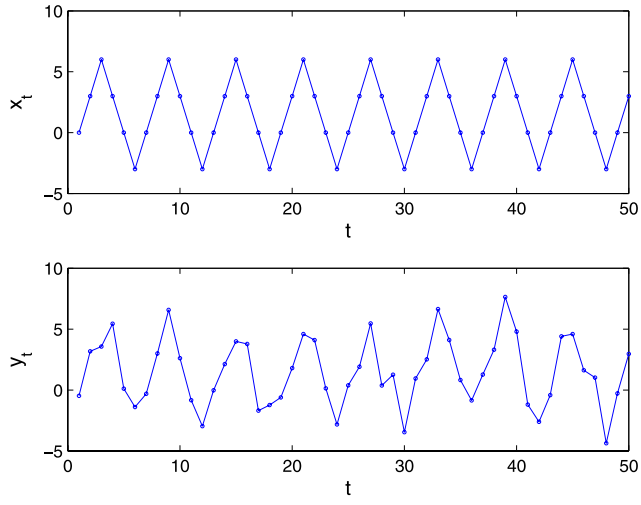

FIG. 5. *The upper panel is a realization of the hidden skeleton in Example 5.4. The lower panel is an observed time series subject to additive measurement error from* $N(0, 1)$.

Using the APE approach to the simulated data, we denote the matching skeleton by x_t and measure the matching error defined in (5.2) with $T = 100$. Based on 100 replications, we summarize the results in Table 1. The matching errors have means and standard deviations in the parentheses in column 3; the average and standard error (in the parentheses) of the periods in all the matching models are listed in column 4. Our results suggest that the APE($\leq m$) with $m > 1$ performs much better than the APE(≤ 1), both in terms of matching the dynamic range and the periodicity.

6. APPLICATION TO REAL DATA SETS

In this section, we study four real time series, some of which are very well known but others less so. They are the sea levels data, the annual sunspot numbers, Nicholson's blowflies data, and the measles infection data in London after the massive vaccination in the late 1960s.

6.1 Sea Levels Data

Long-term mean sea level change is of considerable interest in the study of global climate change. Measurements of the change can provide an important corroboration of predictions by climate models of global warming. Starting from 1992, in each year 34 equally spaced observations were recorded. The data with the linear trend and seasonality removed are available at http://sealevel.colorado.edu/current/sl_noib_ns_global.txt. The time series is depicted in the first panel of Figure 6. Note that the data are subject to measurement errors of 3–4 mm.

As an experiment with using a much less than ideal model to match this data set, let us postulate an AR

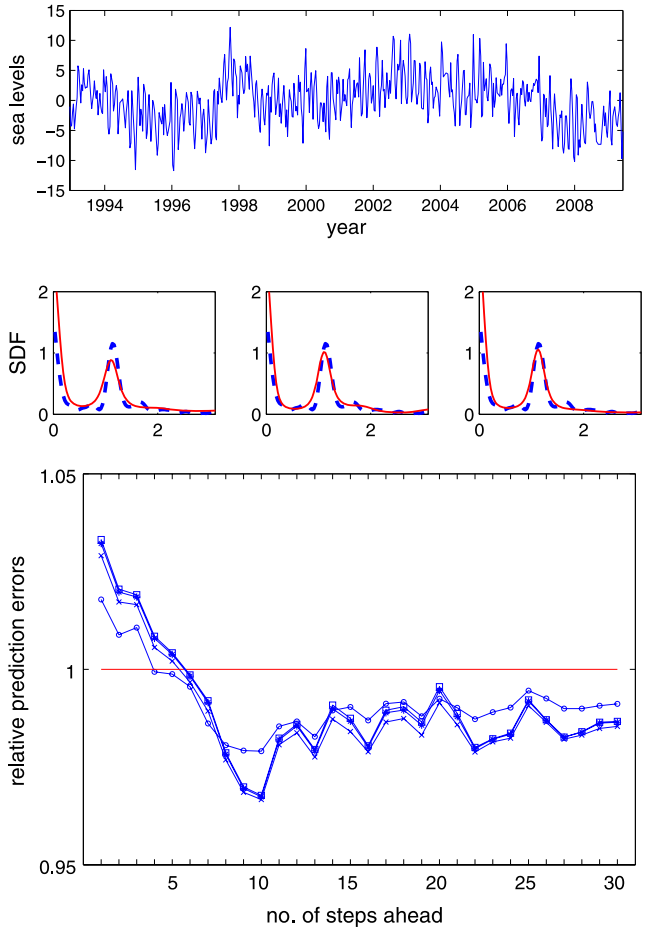

FIG. 6. *Results for the sea level data. The data with linear trend and seasonality removed are shown in the first panel. Panels* 2–4 *are the smoothed sample SDF and those of the fitted models by MLE, the Whittle method and* APE(≤ 20), *respectively. Panel* 5 *is the relative averaged multi-step-ahead prediction errors by taking those of the one-step method as one unit. The curves marked by "$\circ$," "$\times$," "$\star$" and "$\diamond$" are for* APE(≤ 10), APE(≤ 20), APE(≤ 30) *and* APE(≤ 50), *respectively.*

model. By AIC, the order of the AR model is selected as 6. Next, we apply the MLE (equivalently the one-step-ahead prediction estimation method), the Whittle method and the up-to-m-step-ahead prediction estimation method to the data. The results are shown in Figure 6. The sample spectral density function (SDF) is estimated by the method of Fan and Zhang (2004). The results show clear evidence of long-memory with the singularity at the origin, which is well captured by all three methods. However, for the peak away from the origin, the Whittle estimation and APE($\leq m$) show very similar matching capability and both show much better match than the MLE.

To investigate further, we build an AR(6) model for every span of observations of length $T = 100$ and make predictions from 1 step ahead to 30 steps ahead. For the different estimation methods, their averaged prediction errors based on all periods are displayed in the bottom panels of Figure 6. The MLE method shows clear superior performance for short-term prediction, while the reverse is true from 5 steps onward.

6.2 Annual Sunspot Numbers

Sunspots, as an index of solar activity, are relatively cooler and darker areas on the sun's surface resulting from magnetic storms. Sunspots have a cycle of length varying from about 9 to 13 years. Statisticians have fitted several models to predict sunspot numbers. They have also noticed that the cycles are asymmetric and that the time from the initial minimum of a cycle to its next maximum, called the rise time, and the time from a cycle maximum to its next minimum, called the fall time, are fairly regular. Due to their link to other kinds of solar activity, sunspots are helpful in predicting space weather and the state of the ionosphere. Thus, sunspots can help predict conditions of short-wave radio propagation as well as satellite communications. Historical data of the sunspots have been recorded in different parts of the world. The data we use are the annual sunspot numbers for the period 1700–2008 which are obtainable from http://www.ngdc.noaa.gov/stp/SOLAR/. Yule (1927) was the first statistician to model the sunspot number using a model, now known as the autoregressive model, with lag 2. Later refinements of stationary linear models can be found in, for example, Brockwell and Davis (1991) and others; higher-order AR models or ARMA models are used. Akaike (1978) suggested that the data are better modeled as nonstationary over a long period. Tong and Lim (1980) noticed nonlinearity in the data dynamics and proposed the use of a self-exciting threshold autoregressive model (or a SETAR model for short). In the following, we postulate a two-regime SETAR model of order 3 with delay parameter equal to 2 for the annual sunspot numbers (1700–2008). Specifically,

$$x_t = \begin{cases} a_0 + b_0 x_{t-1} + c_0 x_{t-2} + d_0 x_{t-3}, & \text{if } x_{t-2} \leq \tau_0, \\ a_1 + b_1 x_{t-1} + c_1 x_{t-2} + d_1 x_{t-3}, & \text{if } x_{t-2} > \tau_0, \end{cases}$$

where $x_t = \log(\text{no. of sunspots} + 1)$. Note that Cheng and Tong (1992) recommended a nonparametric AR(4) model. We also tried SETAR model of order 4 with delay parameter equal to 2. The performances of both models are very similar.

We use each fixed span of T observations to fit the postulated model and then use it to do a post-sample prediction based on the skeleton of the fitted model.

TABLE 2
The averaged difference (and its standard error) of cycle periods in the data and matching models and the number of unstable matching models [in squared brackets]

m in	Length of time series			
APE($\le m$)	20	35	50	100
1	2.5448 (3.0084) [42]	1.7115 (1.8162) [2]	1.3355 (1.5718) [0]	1.5934 (1.4051) [0]
10	1.3454 (1.7082) [13]	0.9576 (0.8499) [0]	0.8459 (0.9584) [0]	0.4487 (0.5427) [0]
20	1.2972 (1.7143) [10]	0.8975 (1.1257) [0]	0.7580 (0.6074) [0]	0.4134 (0.9715) [0]
30		0.8802 (1.1415) [1]	0.8449 (0.5807) [0]	0.3640 (0.5894) [0]
50			0.8548 (0.5813) [0]	0.3538 (0.4267) [0]

We measure the following: (1) the difference of cycle periods between the data and the fitted model; (2) the frequency of stable fitted models; (3) the out-of-sample prediction errors based on the skeletons of models fitted by the APE($\le m$) for different m; (4) the difference between the observed time series and the time series generated by the best fitting skeleton by reference to (5.2).

The results are shown in Figure 7 and Table 2. We may draw the following conclusions. (1) When the observed time series is short (e.g., $T = 20, 35$), APE($\le m$) with $m > 1$ show better matching than APE(≤ 1) in both one-step-ahead prediction and multi-step-ahead prediction; see panels 1 and 2 in Figure 7. When the length of the time series is longer (e.g., $T = 50, 100$), APE(≤ 1) can lead to fitted models with better short-term (less than 4 steps ahead) prediction than APE($\le m$) with $m > 1$, but for prediction beyond 4 steps ahead, the reverse appears to be the case, in line with our understanding of the APE method. (2) When the observed time series is short, APE($\le m$) with $m > 1$ shows its ability in avoiding unstable models; see the numbers in the square brackets of Table 2. (3) For both short time series and long time series, models fitted by APE($\le m$) with $m > 1$ show better matching of the observed time series in terms of their cycles; see Table 2 and the horizontal lines in Figure 7.

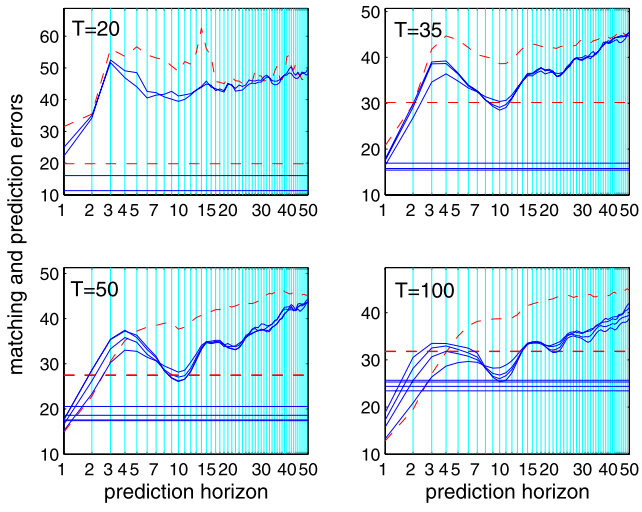

FIG. 7. *The dashed curves are the averaged prediction errors based on* APE(≤ 1), *the solid curves are those based on* APE($\le m$) *with* $m = 10, 20, 30, 50$, *respectively. The horizontal dashed lines are the matching errors for the* APE(≤ 1), *the solid lines are those for* APE($\le m$) *with* $m = 10, 20, 30, 50$, *respectively.*

6.3 Nicholson's Blowflies

The data consist of the total number of blowflies (Lucilia cuprina) in a population under controlled laboratory conditions. The data represent counts for every second day. The developmental delay (from egg to adult) is between 14 and 15 days for the blowflies under the conditions employed (Gurney, Blythe and Nisbet, 1980). Nicholson obtained 361 bi-daily recordings over a 2-year period (722 days). However, due to biological evolution (Stokes et al., 1988), the whole series cannot be considered to represent the same system; a major transition appears to have occurred around day 400. Following Tong (1990), we consider the first part of the time series (to day 400, thus $T = 200$), for which the population has a 19 bi-days cycle; see Figure 8.

Next, we postulate the single species animal population discrete model (1.2) with $b(x_{t-\tau}) = cx_{t-\tau}^{\alpha-1} \cdot \exp(-x_{t-\tau}/N_0)$, and thus

$$x_t = cx_{t-\tau}^{\alpha} \exp(-x_{t-\tau}/N_0) + \nu x_{t-1},$$

where we take $\tau = 8$ (bi-days) corresponding to the time taken for an egg to develop into an adult. Note that we specify $b(x_{t-\tau})$ slightly differently from Gurney, Blythe and Nisbet (1980) by adding an exponent $\alpha - 1$ to $x_{t-\tau}$, which is usually necessary when a differential equation model is discretized and approximated by a time series model; see Glass, Xia and Grenfell (2003). In the model, there are four parameters:

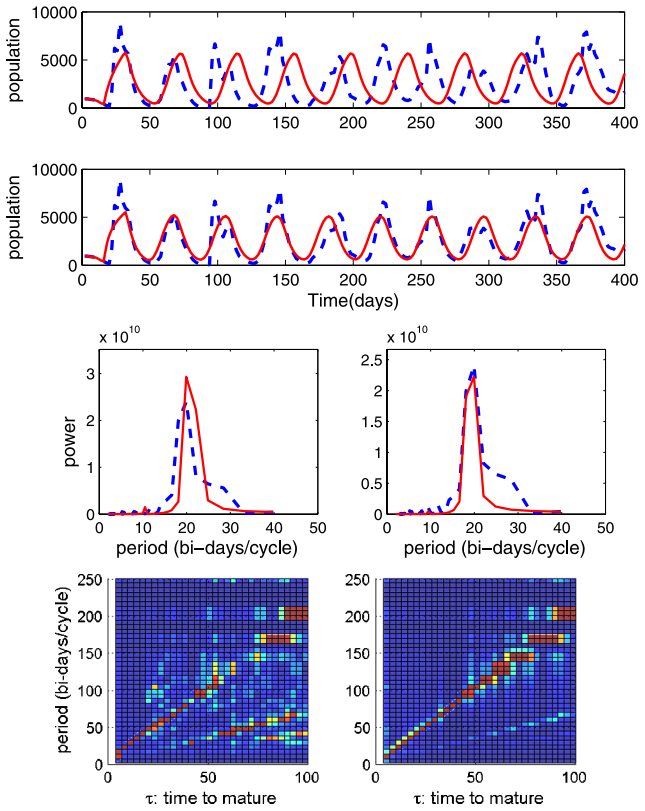

FIG. 8. *Results for the Nicholson's blowflies data. In the first two panels, the dashed lines are for the observed population; the solid lines are for realizations from models fitted by* APE(≤ 1) *and* APE($\leq T$), *respectively. The dashed lines in panels* 3 *and* 4 *are the periodograms of the observed data, and the solid lines are those of the models fitted by* APE(≤ 1) *and* APE($\leq T$), *respectively. In panels* 5 *and* 6, *for each* τ *marked in the* x*-axis, the vertical column is the periodogram with the values color-coded, brighter color (blue being dull) corresponding to higher power value. Thus the brightest point indicates the cycle-period of the dynamics at* τ.

c, α, N_0 and ν. The (one-step-ahead prediction) MLE estimates for the parameters are

$$\hat{c} = 20.1192, \quad \hat{N}_0 = 589.5553,$$
$$\hat{\nu} = 0.7598, \quad \hat{\alpha} = 0.8461.$$

The APE method gives

$$\hat{c} = 591.5801, \quad \hat{N}_0 = 1307.0,$$
$$\hat{\nu} = 0.6469, \quad \hat{\alpha} = 0.2633.$$

The skeletons based on the postulated model with parameters estimated by above methods are shown in panels 1 and 2 in Figure 8, respectively. They show that APE($\leq T$) results in a model whose skeleton matches the observed cycles to a much greater extent than APE(≤ 1). APE(≤ 1) gives a period of 21 bi-days; APE($\leq T$) gives a period of 19 bi-days, which is almost exactly the average period of the observed cycles. We have also postulated a SETAR model. With APE($\leq T$), the SETAR model can also capture the observed period very well, but again this is not the case with APE(≤ 1). To investigate how the cycles change with the time needed by the fly to grow to maturity, we vary the time τ from 4 to 100 bi-days. The corresponding cycles (in bi-days) are shown in the last two panels of Figure 8. APE($\leq T$) shows a clear linearly increasing trend in the cycle-periods as τ increases, while APE(≤ 1) shows strange excursions that are difficult to interpret. The linear relationship suggested by APE($\leq T$) may be helpful in throwing some light on the important but not completely resolved cycle problem of animal populations. We have also tried APE($\leq m$) with m equal to twice or thrice the cycle-period. Their results are similar to those of APE($\leq T$).

6.4 Measles Dynamics in London

It is well known that the continuous-time susceptible-infected-recovered (SIR) model using a set of ordinary differential equations can describe qualitatively the behavior of epidemics quite well. However, it is difficult to use it for real data modeling when the observations are made in discrete time. To bridge the gap between the theoretical model and real data fitting, several discrete-time or chain models have been introduced. The Nicholson–Bailey host-parasite model (Nicholson and Bailey, 1935) is an early example. Bailey (1957), Bartlett (1960) and Finkenstädt and Grenfell (2000) proposed different types of discrete-time epidemic models. A general discrete-time or chain model can be written as follows:

$$\begin{cases} S_{t+1} = S_t + B_t - I_{t+1}, \\ I_{t+1} = S_t P(I_t), \end{cases} \tag{6.1}$$

where I_t, S_t and B_t are respectively the number of the infectious, the number of the susceptible and the number of births, all at the tth time unit. There are many possible functional forms for the (probability) $P(I_t)$. Examples are $1-(1-r_0/N)^{I_t}$ (Bartlett, 1960), $1-\exp(-r_0 I_t/N)$ (Bartlett, 1956), $r_0 I_t/N$ (Baily, 1957) and $R_0 I_t^{\alpha}/N$ (Liu, Hethcote and Levin, 1987; Finkenstädt and Grenfell, 2000), where N is the effective population of hosts, and r_0 is the basic reproductive rate.

Next, we postulate the following (deterministic) discrete-time SIR model for the transmission of measles:

$$I_{t+1} = \exp(\delta_{t,k}\beta_k) S_t I_t,$$
$$S_{t+1} = S_t + b_t - I_{t+1} = S_0 + \sum_{\tau=0}^{t} B_\tau - \sum_{\tau=1}^{t+1} I_\tau,$$

where $\exp(\delta_{t,k}\beta_k)$ is employed to indicate the seasonality force, with $\delta_{t,k} = 1$ if time t is at the kth season, 0 otherwise. For measles, the time unit for t is bi-weekly, based on the infection procedure of measle; see Finkenstädt and Grenfell (2000). Now, $k = 1, \ldots, 26$ bi-weeks corresponds to about 54 weeks in a year. Finkenstädt and Grenfell (2000) considered the same model but with the first equation being $I_{t+1} = \exp(\delta_{t,k}\beta_k) S_t I_t^{\alpha}$. Here, we take $\alpha = 1$ for two reasons. (1) If $\alpha < 1$, Finkenstädt and Grenfell (2000) were unable to use the model to explain the dynamics of measles in the massive vaccination era. (2) Experience with statistical modeling of ecological populations suggests that α can be taken as 1 with improved interpretation; see Bjønstad, Finkenstädt and Grenfell (2002). In practice, I_t may not be observed directly; what can be observed is a random variable, say y_t, that has mean I_t. For this observable y_t, we postulate a model x_t that follows a Poisson distribution with mean I_t.

There are some problems with the data. There is nonnegligible observation error in the data due to the under-reporting rate, which can be as high as 50%; see Finkenstädt and Grenfell (2000), where a method was proposed to recover the data. Following their method, the data were adjusted for the under-reporting rate. The adjusted data are shown in dashed lines in panels 1 and 2 of Figure 9. It is known that the role of vaccination is equivalent to the reduction of the birth rate (Earn et al., 2000). Thus, we adjust the number of births by multiplying it by the un-vaccination rate, that is, 1−(vaccination rate). We show the adjusted births in the third panel of Figure 9. Another problem with the data is that the susceptible S_t is unknown, which can also be reconstructed by the method of Finkenstädt and Grenfell (2000).

The estimates of the model by APE(≤ 1) are listed in Table 3. To ease the calculation of APE($\leq T$), we simplify the model by taking $\beta_k = \bar{\beta} + \lambda(\beta_{k,1} - \bar{\beta})$, where $\beta_{1,1}, \ldots, \beta_{26,1}$ are the estimates of APE(≤ 1) and $\bar{\beta}$ is their average. Consequently, only λ and S_0 need to be estimated in implementing APE($\leq T$). The skeletons based on models fitted by APE(≤ 1) and APE($\leq T$) are shown in solid red lines in panel 1 and panel 2 of Figure 9, respectively. APE($\leq T$) shows a much better match than APE(≤ 1) in terms of outbreak scale and cycle period. The periodogram is also much better matched by APE($\leq T$) than by APE(≤ 1); see the last two panels of Figure 9. We have also tried APE($\leq m$) with m being twice or thrice the cycle period (i.e., 26 bi-weeks). The results are similar to APE($\leq T$).

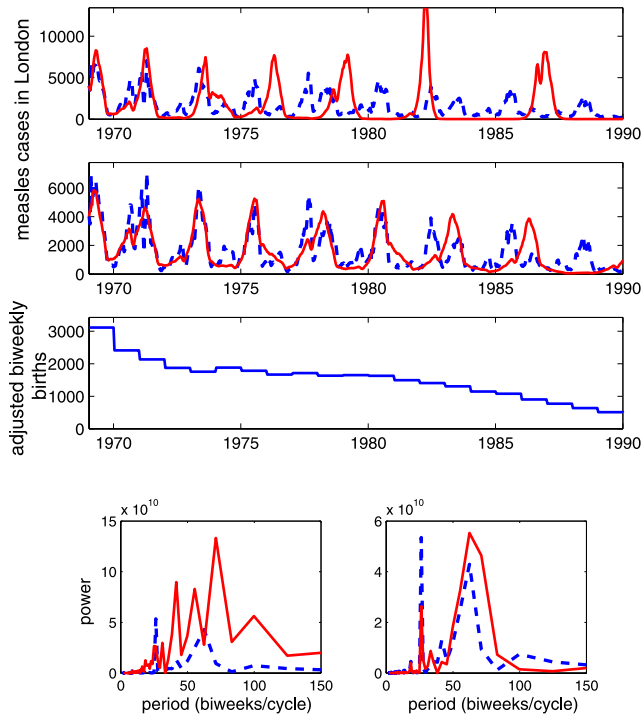

FIG. 9. *Results for modeling the measles incidents in London. The dashed lines in panels* 1 *and* 2 *are the recovered incidents of measles*; *the solid lines are the realizations of the model based on* APE(≤ 1) *and* APE($\leq T$), *respectively. Panel* 3 *is the adjusted birth rate by removing the vaccinated*; *in the bottom panels, the dashed lines are the periodograms of the data and the red lines are those of the matching skeleton by* APE(≤ 1) *and* APE($\leq T$), *respectively.*

An important feature in the measles transmission is that there were some big annual outbreaks in the 1950s when the birth rate was very high after the second world war, and some big bi-annual outbreaks in the middle of the 1960s when the birth rate was relatively low. The dynamics before the massive vaccination in the late 1960s was modeled very well by a time series model in Finkenstädt and Grenfell (2000). The theory that relates population cycle length to birth rate has been well accepted in epidemiology and ecology. In epidemiology, the relationship will either prolong or shorten the cumulation procedure of susceptibles for a big outbreak. Observations from the other sources have lent support to this theory. For example, the measles in New York have a three-year or four-year cycle when the birth rate is very low. As another supporting piece of evidence, in the vaccination era, the cycles lasted longer, to four or five years because vaccination is equivalent to the reduction of birth rate in the transmission of disease. However, the dynamics after the massive vaccination is difficult to model

TABLE 3
Parameters in the measles transmission model

Method	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
APE(≤ 1)	−11.92	−12.00	−11.88	−11.99	−11.89	−11.81	−11.89	−11.97
APE($\le T$)	−11.95	−12.00	−11.93	−11.99	−11.93	−11.89	−11.93	−11.98
	β_9	β_{10}	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}
APE(≤ 1)	−11.92	−11.99	−12.05	−12.01	−11.93	−11.96	−11.98	−12.04
APE($\le T$)	−11.95	−11.99	−12.03	−12.00	−11.96	−11.98	−11.99	−12.02
	β_{17}	β_{18}	β_{19}	β_{20}	β_{21}	β_{22}	β_{23}	β_{24}
APE(≤ 1)	−11.95	−12.15	−12.28	−12.40	−12.21	−11.99	−11.79	−11.87
APE($\le T$)	−11.97	−12.08	−12.16	−12.23	−12.12	−11.99	−11.87	−11.92
	β_{25}	β_{26}	S_0					
APE(≤ 1)	−11.99	−11.98	17,8280					
APE($\le T$)	−11.99	−11.98	16,8190					

due to the quickly changing birth rate. The method of Finkenstädt and Grenfell (2000) has failed to capture this change of cycles in the vaccination era. It is therefore worth noting that our modified model, with the aid of APE($\le m$) with $m > 1$, shows satisfactory matching. To investigate further how the cycles change with the birth rate, for each fixed number of births we run the estimated model and depict its periodogram and highlight the peaks by color-coding (brighter color for higher power). The peaks with the brightest points correspond to the cycles of the postulated model. Figure 10 shows clearly that when the birth rate is high (from about 5,000 upward) the cycle is annual, but when the birth rate is medium at about 3,000 to 4,000, the cycles become two-year cycles. As the birth rate gets lower, the model shows that cycles become three-year cycles or even five-year cycles. It seems that by fitting a substantive model with the catch-all approach, we have obtained perhaps the first *discrete-time* model that is capable of revealing the complete function linking birth-rates to the cyclicity of measles epidemics, thereby lending support to the general theory developed by Earn et al. (2000), which was based on differential SIR equations in continuous time.

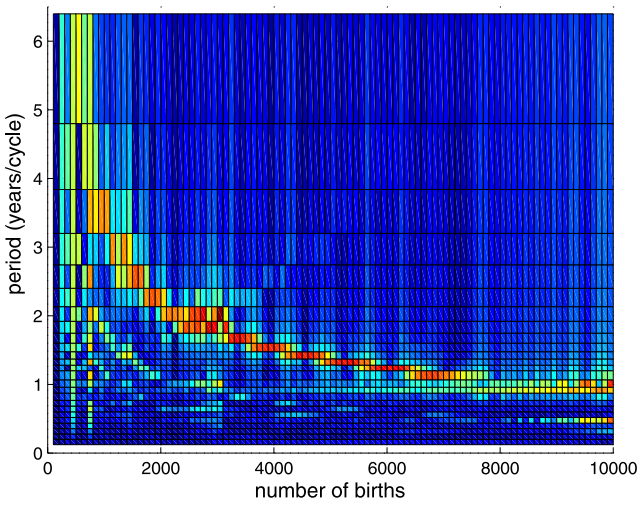

FIG. 10. *Measles transmission. Each vertical column is the periodogram with the values color-coded, brighter color corresponding to higher power.* (*Dark blue is considered a dull color.*)

7. CONCLUDING REMARKS AND FURTHER PROBLEMS

In this paper, we adhere to Box's dictum and abandon, right from the very beginning, the assumption of either the postulated parametric model being true or the observations being error-free. Instead, we focus on ways to improve the feature matching of a postulated parametric model to the observable time series. We have introduced the notion of an optimal parameter in the absence of a true model and defined a new form of consistency. In particular, we have synthesized earlier attempts into a systematic approach of estimation of the optimal parameter, by reference to up-to-m-step-ahead predictions of the postulated model. We have also developed some general results with proofs.

Conventional methods of estimation are typically based on just the one-step-ahead prediction. Our analysis, simulation study and real applications have convinced us that they are often found wanting in many situations, for example, the absence of a true model, short data sets, observation errors, highly cyclical data and others. Our stated primary objective is feature matching. Prediction is secondary here. However, we have

evidence to suggest that a model with good feature matching can stand a better chance of enjoying good medium- to long-term prediction. Of course, if the aim is prediction with a *specified* horizon, say m_0, then we simply set $w_{m_0} = 1$ and the rest zero. In that case, our catch-all approach really offers nothing new.

Let us now take another look at the difference between APE($\le m$) with $m > 1$ and APE(≤ 1). Suppose we postulate the model $x_t = g_\theta(X_{t-1}) + \varepsilon_t$ where $X_{t-1} = (x_{t-1}, \ldots, x_{t-p})$ to match an observable y-time series. Given data $\{y_1, y_2, \ldots, y_T\}$, APE($\le m$) with $m > 1$ and with a constant $w_j > 0$, all j, estimates θ by minimizing the objective function

$$L_m(\theta) = \sum_{t=p+1}^{T} \sum_{k=1}^{\min(m,T-t)} \{y_{t-1+k} - g_\theta^{[k]}(Y_{t-1})\}^2 = L_1(\theta) + L_1^+(\theta),$$

where $Y_{t-1} = (y_{t-1}, \ldots, y_{t-p})$ and

$$L_1(\theta) = \sum_{t=p+1}^{T} \{y_t - g_\theta^{[k]}(Y_{t-1})\}^2,$$

$$L_1^+(\theta) = \sum_{t=p+1}^{T} \sum_{k=2}^{\min(m,T-t)} \{y_{t-1+k} - g_\theta^{[k]}(Y_{t-1})\}^2.$$

Note that $L_1(\theta)$ is the commonly used objective function for APE(≤ 1), while $L_1^+(\theta)$ is the extra information provided by the dynamics. In terms of samples, $L_1(\theta)$ is based on sample $\{y_t, Y_{t-1} : t = p + 1, \ldots, T\}$. The extra term $L_1^+(\theta)$ is associated with the extra pseudo designed samples $\{y_{t-1+k}, Y_{t-1} : t = p + 1, \ldots, T, k = 1, \ldots, m\}$. If the data are actually generated by the postulated model (a rare event), then under some general conditions such as ε_t are i.i.d. normal, $L_1(\theta)$ will include all the information about θ. In that case, estimation based on $L_1(\theta)$ alone is the most efficient and the extra term $L_1^+(\theta)$ can provide no additional information. However, if the data are not exactly generated by the postulated model (a common event), the extra information provided by $L_1^+(\theta)$ can indeed be very helpful and should be exploited.

Despite evidence, both theoretical and practical, of the utility of the catch-all approach, much more remains to be done. Our paper should be seen as the first word on feature matching. Although we have provided some concrete approaches, such as the catch-all-conditional-mean approach, the catch-all-ACF approach, which can easily be generalized to catch-all-mth-order moments and others, there are outstanding issues. For example, we can, at present, offer no theoretical guidance on the specification of the weights, $\{w_m\}$. We have only offered some practical suggestions based on our experience. It would be interesting to investigate further possible connections with a prior in Bayesian statistics.

We have been quite fortunate with our real examples using the APE method, thanks to our long-standing collaboration with ecologists and epidemiologists. However, we are conscious of the need for the accumulation of further experience. We are convinced that, especially in the area of substantive modeling, guidance by relevant subject scientists is paramount. Relevant references include He, Ionides and King (2010), King et al. (2008), Laneri et al. (2010) and others.

Last but not least, future research should include at least the following: other weaker forms of (2.1), choice of a suitable weaker form in a specific application, other criteria for model comparison, non-additive and/or heteroscedastic measurement errors, the relaxation of stationarity, the effect of prefiltering of data, multiple time series, model selection among a set of wrong models (each fitted by the catch-all method; perhaps the idea of model calibration in econometrics might be useful here), possible extension to other types of dependent data, for example, spatial data.

APPENDIX: OUTLINES OF THEORETICAL JUSTIFICATION

We need the following assumptions. However, these assumptions can be relaxed with more complicated theoretical derivation.

(C1) Time series $\{y_t\}$ is a strictly stationary and strongly mixing sequence with exponentially decreasing mixing-coefficients.

(C2) The moments $\mathbf{E}\|y_t\|^{2\delta}$, $\mathbf{E}\|g_\vartheta^{[k]}(y_t, \ldots, y_{t-p})\}\|^{2\delta}$, $\mathbf{E}\|\partial g_\vartheta^{[k]}(y_t, \ldots, y_{t-p})/\partial\theta\|^\delta$ and $\mathbf{E}\|\partial^2 g_\vartheta^{[k]}(y_t)/(\partial\theta\,\partial\theta^\top)\|^\delta$ exist for some $\delta > 2$.

(C3) The functions $\partial g_\theta^{[k]}(y_t)/\partial\theta$ and $\partial^2 g_\theta^{[k]}(y_t)/(\partial\theta\,\partial\theta^\top)$ are continuous in $\theta \in \Theta$ and

$$\Omega \stackrel{\text{def}}{=} \mathbf{E} \sum_{k=1}^{m} w_k \left\{ \frac{\partial g_\vartheta^{[k]}(y_t)}{\partial\theta} \frac{\partial g_\vartheta^{[k]}(y_t)}{\partial\theta^\top} - \big[y_{t+k} - g_\vartheta^{[k]}(y_t)\big] \frac{\partial^2 g_\vartheta^{[k]}(y_t)}{\partial\theta\,\partial\theta^\top} \right\}$$

is nonsingular.

(C4) The function $\sum_{k=1}^{m} w_k E[y_{t+k} - g_\theta^{[k]}(Y_t)]^2$ has a unique minimum point for θ in the parameter space Θ.

THEOREM A. *Suppose that $\{x_t(\theta)\}$ and $\{y_t\}$ have the same marginal distribution and each has second-order moments. Then*

$$D_C(y_t, x_t(\theta)) \le C_1 \tilde{Q}(y_t, x_t(\theta)),$$
$$D_F(y_t, x_t(\theta)) \le C_2 \tilde{Q}(y_t, x_t(\theta))$$

for some positive constants C_1 and C_2. Moreover, if $\{x_t(\theta)\}$ and $\{y_t\}$ are linear AR *models, then there are some positive constants C_3 and C_4 such that*

$$\tilde{Q}(y_t, x_t(\theta)) \le C_3 D_C(y_t, x_t(\theta)),$$
$$\tilde{Q}(y_t, x_t(\theta)) \le C_4 D_F(y_t, x_t(\theta)).$$

PROOF. By the condition on the marginal distributions, we have

$$\mathbf{E}(y_{t+m}) = \mathbf{E}(x_{t+m}). \tag{A.1}$$

Since $\mathbf{E}[y_t\{y_{t+m} - \mathbf{E}(y_{t+m}|y_t)\}] = 0$, we have

$$\begin{aligned}\mathbf{E}(y_t y_{t+m}) &= \mathbf{E}\{y_t g_\theta^{[m]}(y_t)\} + \mathbf{E}[y_t\{y_{t+m} - g^{[m]}(y_t)\}] \\ &= \mathbf{E}\{y_t g_\theta^{[m]}(y_t)\} \\ &\quad + \mathbf{E}[y_t\{\mathbf{E}(y_{t+m}|y_t) - g_\theta^{[m]}(y_t)\}].\end{aligned}$$

By the assumption on the marginal distribution, we have

$$\begin{aligned}\mathbf{E}\{y_t g_\theta^{[m]}(y_t)\} &= \mathbf{E}\{x_t g_\theta^{[m]}(x_t)\} \\ &= \mathbf{E}\{x_t \mathbf{E}(x_{t+m}|x_t)\} = \mathbf{E}(x_t x_{t+m}).\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{E}(y_t y_{t+m}) &= \mathbf{E}(x_t x_{t+m}) \\ &\quad + \mathbf{E}[y_t\{\mathbf{E}(y_{t+m}|y_t) - g_\theta^{[m]}(y_t)\}].\end{aligned} \tag{A.2}$$

It follows from (A.1) and (A.2) that

$$\gamma_y(m) = \gamma_x(m) + \Delta_m,$$

where $\Delta_m = \mathbf{E}[y_t\{\mathbf{E}(y_{t+m}|y_t) - g_\theta^{[m]}(y_t)\}]$. By the Hölder inequality, we have

$$|\Delta_m| \le \{\mathbf{E}y_t^2\}^{1/2}\{\mathbf{E}\{\mathbf{E}(y_{t+m}|y_t) - g_\theta^{[m]}(y_t)\}^2\}^{1/2}.$$

Therefore,

$$\begin{aligned}&D_C(x_t(\theta), y_t) \\ &\le \sup_{\{w_k\}} \sum_{k=0}^{\infty} w_k \{\mathbf{E}y_t^2\}^{1/2} \\ &\qquad \cdot \{\mathbf{E}\{\mathbf{E}(y_{t+k}|y_t) - g_\theta^{[k]}(y_t)\}^2\}^{1/2} \\ &\le C_1 \tilde{Q}(\theta),\end{aligned}$$

where $C_1 = \{\mathbf{E}y_t^2\}^{1/2}$. This is the first inequality of Theorem A.

For ease of exposition, assume that $\{y_t\}$ and $\{x_t(\theta)\}$ are given by AR models with the same order, P. Otherwise we take the order as the larger of the two orders. So $y_t = \beta_1 y_{t-1} + \cdots + \beta_P y_{t-P} + \varepsilon_t$ and $x_t = \theta_1 x_{t-1} + \cdots + \theta_P x_{t-P} + \eta_t$.

Let $e_1 = (1, 0, \ldots, 0)^\top$, $Y_{t-1} = (y_{t-1}, \ldots, y_{t-P})^\top$, $X_{t-1} = (x_{t-1}, \ldots, x_{t-P})^\top$, $\mathcal{E}_t = (\varepsilon_t, 0, \ldots, 0)^\top$ and

$$\Gamma_0 = \begin{pmatrix} \beta_1 & \beta_2 & \cdots & \beta_{P-1} & \beta_P \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

$$\Gamma = \begin{pmatrix} \theta_1 & \theta_2 & \cdots & \theta_{P-1} & \theta_P \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

Then $Y_{t-1+m} = e_1^\top \Gamma_0^m Y_{t-1} + e_1^\top(\mathcal{E}_{t-1+m} + \Gamma_0 \mathcal{E}_{t-2+m} + \cdots + \Gamma_0^m \mathcal{E}_t)$. It follows that

$$\begin{aligned}&[\gamma_y(m), \gamma_y(m+1), \ldots, \gamma_y(m+P-1)] \\ &\quad = \mathbf{E}(y_{t-1+m} Y_{t-1}^\top) = e_1^\top \Gamma_0^m \Sigma_0,\end{aligned} \tag{A.3}$$

where $\Sigma_0 = E(Y_{t-1} Y_{t-1}^\top) = (\gamma_y(|i-j|))_{1 \le i,j \le P}$. Similarly, we have

$$\begin{aligned}&[\gamma_x(m), \gamma_x(m+1), \ldots, \gamma_x(m+P-1)] \\ &\quad = \mathbf{E}(x_{t-1+m} X_{t-1}^\top) = e_1^\top \Gamma^m \Sigma,\end{aligned} \tag{A.4}$$

where $\Sigma = E(X_{t-1} X_{t-1}^\top) = (\gamma_x(|i-j|))_{1 \le i,j \le P}$.

Assuming ε_t, η_t are independent sequences of i.i.d. random variables, we have

$$\mathbf{E}(y_{t-1+m}|Y_{t-1}) = e_1^\top \Gamma_0^m Y_{t-1},$$
$$\mathbf{E}(x_{t-1+m}|X_{t-1} = Y_{t-1}) = e_1^\top \Gamma^m Y_{t-1}.$$

(Note: The i.i.d. assumption can be relaxed at the expense of a much lengthier proof.) It follows that

$$\begin{aligned}&\mathbf{E}\{\mathbf{E}(y_{t-1+m}|Y_{t-1}) - \mathbf{E}(x_{t-1+m}|X_{t-1} = Y_{t-1})\}^2 \\ &= e_1^\top(\Gamma_0^m - \Gamma^m)\Sigma_0(\Gamma_0^m - \Gamma^m)^\top e_1 \\ &= e_1^\top[\Gamma_0^m \Sigma_0 - \Gamma^m \Sigma + \Gamma^m(\Sigma - \Sigma_0)] \\ &\quad \cdot \Sigma_0^{-1}[\Gamma_0^m \Sigma_0 - \Gamma^m \Sigma + \Gamma^m(\Sigma - \Sigma_0)]^\top e_1 \\ &\le \lambda_{\min}^{-1}(\Sigma_0)\|[\gamma_y(m) - \gamma_x(m), \\ &\qquad \gamma_y(m+1) - \gamma_x(m+1), \ldots,\end{aligned}$$

$$\gamma_y(m+P-1)-\gamma_x(m+P-1)] + e_1^\top \Gamma^m(\Sigma-\Sigma_0)\|^2$$
$$\le \lambda_{\min}^{-1}(\Sigma_0)\sum_{k=m}^{m+P-1}\{\gamma_y(k)-\gamma_x(k)\}^2 + \lambda_{\min}^{-1}(\Sigma_0)\lambda_{\max}^m(\Gamma)P\sum_{k=0}^{P-1}\{\gamma_y(k)-\gamma_x(k)\}^2,$$

where $\lambda_{\min}(\Sigma_0)$ and $\lambda_{\max}(\Gamma)$ are the minimum eigenvalue of Σ_0 and the maximum eigenvalue of Γ, respectively. Note that $\lambda_{\max}(\Gamma)<1$. Therefore,

$$\tilde{Q}(\theta) \le P\lambda_{\min}^{-1}(\Sigma_0)\sum_{m=0}^{\infty} w_m\{\gamma_y(k)-\gamma_x(k)\}^2 = C_3 D_c(x_t(\theta), y_t),$$

for some $w_m \ge 0$. The proof is completed. □

THEOREM B. *Under assumptions* (C1) *and* (C2), *we have in distribution*

$$\sqrt{n}\{\hat{\theta}_{\{m\}}-\vartheta\} \to N(0,\tilde{\Sigma}_m),$$

where $\vartheta=(\tilde{\Gamma}_m^\top\tilde{\Gamma}_m)^{-1}\tilde{\Gamma}_m^\top\tilde{\Upsilon}_m$ *and* $\tilde{\Sigma}_m$ *is a positive definite matrix. As a special case, if* $y_t=x_t+\eta_t$ *with* $\mathrm{Var}(\varepsilon_t)>0$ *and* $\mathrm{Var}(\eta_t)=\sigma_\eta^2>0$, *then the above asymptotic result holds with* $\vartheta=\theta+\sigma_\eta^2(\Gamma_m^\top\Gamma_m+2\sigma_\eta^2\Gamma_p+\sigma_\eta^4 I)^{-1}(\Gamma_p+\sigma_\eta^2 I)\theta$.

PROOF. To simplify the range of summation in the triangular array due to the lags with fixed m as $T\to\infty$, we introduce $\cong$ to denote the fact that the quantities on both sides of it have negligible difference. By Theorem 3.1 of Romano and Thombs (1996), in an enlarged probability space we have

$$\hat{\Gamma}_m=\Gamma_m+n^{-1/2}\mathcal{U}_m+o_p(n^{-1/2}),$$
$$\hat{\Upsilon}_m=\Upsilon_m+n^{-1/2}\mathcal{V}_m+o_p(n^{-1/2}),$$

where $\mathcal{U}_m$ and $\mathcal{V}_m$ have the same structure as Γ_m and Υ_m, respectively, but with $\gamma(k)$ being replaced by v_k and $(v_{i+1},\dots,v_{i+j})$ for any i,j being jointly normal, with variance–covariance matrix given by Romano and Thombs (1996). Therefore, we have

$$\hat{\theta}_m=\vartheta+n^{-1/2}\mathcal{W}+o_P(n^{-1/2}),$$

where $\mathcal{W}=(\Gamma_m^\top\Gamma_m)^{-1}\mathcal{U}_m^\top\Upsilon_m+(\Gamma_m^\top\Gamma_m)^{-1}\Gamma_m^\top\mathcal{V}_m-(\Gamma_m^\top\Gamma_m)^{-1}\{\Gamma_m^\top\mathcal{U}_m+\mathcal{U}_m^\top\Gamma_m\}(\Gamma_m^\top\Gamma_m)^{-1}\Gamma_m^\top\Upsilon_m$ is a linear combination of $\{v_k\}$. Thus, $\mathcal{W}$ is normally distributed with mean 0. This is the first part of Theorem B.

If $y_t=x_t+\eta_t$, let $\gamma_x(k)=n^{-1}\sum_{i=1}^n x_t x_{t+k}$; it is easy to see that

$$\hat{\gamma}_y(k)\cong\hat{\gamma}_x(k)+D_k+E_k,\quad k=0,1,\dots,$$

where $D_k=n^{-1}\sum_{t=1}^n(x_{t+k}+x_{t-k})\eta_t$ and $E_k=n^{-1}\sum_{t=1}^n\eta_t\eta_{t+k}$. By the central limit theorem and Theorem 3.1 of Romano and Thombs (1996), in an enlarged probability space there are random variables ξ_k,ζ_k and δ_k such that $\hat{\gamma}_x(k)=\gamma_x(k)+n^{-1/2}\xi_k+o_p(n^{-1/2})$, $D_k=n^{-1/2}\zeta_k+o_p(n^{-1/2})$ and

$$E_k=\begin{cases}\sigma_\eta^2+n^{-1/2}\delta_k+o_p(n^{-1/2}), & \text{if } k=0,\\ n^{-1/2}\delta_k+o_p(n^{-1/2}), & \text{if } k>0,\end{cases}$$

where $\xi_0,\xi_1,\dots,\{\zeta_k,k=0,1,\dots\},\delta_0,\delta_1,\dots$ are mutually independent and $\xi_k=\gamma_x(k)\{\mathbf{E}\varepsilon_t^4-1\}^{1/2}W_0+\sum_{j=1}^\infty\{\gamma_x(j+k)+\gamma_x(j-k)\}W_j$. Here $W_0,W_1,\dots$ are i.i.d. $N(0,1)$, $\zeta_k\sim N(0,2(\gamma_y(0)+\gamma_y(2k)))$, $\mathrm{Cov}(\zeta_k,\zeta_\ell)=2(\gamma_y(k-\ell)+\gamma_y(k+\ell))$ and $\delta_k\sim N(0,\sigma_\eta^4)$ if $k>0$ and $\delta_0\sim N(0,\mathbf{E}(\eta^2-1)^2)$. Define Ξ_k, Z_k and Δ_k similarly as Γ_k with $\gamma_x(k)$ being replaced by ξ_k,ζ_k and δ_k, respectively. Let B_k be a $k\times p$ matrix with the first $p\times p$ submatrix being $\sigma_\eta^2 I_p$ and all the others 0. We have

$$\hat{\Gamma}_k=\Gamma_k+B_k+n^{-1/2}\mathcal{E}_k+o_p(n^{-1/2}),$$

where $\mathcal{E}_k=\Xi_k+Z_k+\Delta_k$,

$$\hat{\Upsilon}_k=\Upsilon_k+n^{-1/2}\Psi_k+o_p(n^{-1/2})=\Gamma_k\theta+n^{-1/2}\Psi_k+o_p(n^{-1/2}),$$

and $\Psi_k=(\xi_1,\dots,\xi_k)^\top$. It follows that

$$\begin{aligned}\hat{\theta}_m=&[\Gamma_m^\top\Gamma_m+2\sigma_\eta^2\Gamma_p+\sigma_\eta^4 I\\ &+n^{-1/2}\{(\Gamma_m+B_m)^\top\mathcal{E}_m+\mathcal{E}_m^\top(\Gamma_m+B_m)\}\\ &+o_p(n^{-1/2})]^{-1}\\ &\cdot[\Gamma_m^\top\Gamma_m+\sigma_\eta^2\Gamma_p\\ &+n^{-1/2}\{(\Gamma_m+B_m)^\top\Psi_m+\mathcal{E}_m^\top\Gamma_m\theta\}\\ &+o_p(n^{-1/2})]\\ =&(\Gamma_m^\top\Gamma_m+2\sigma_\eta^2\Gamma_p+\sigma_\eta^4 I)^{-1}(\Gamma_m^\top\Gamma_m+\sigma_\eta^2\Gamma_p)\theta\\ &+n^{-1/2}\mathcal{W}_n+o(n^{-1/2})\\ =&\theta-\sigma_\eta^2(\Gamma_m^\top\Gamma_m+2\sigma_\eta^2\Gamma_p+\sigma_\eta^4 I)^{-1}(\Gamma_p+\sigma_\eta^2 I)\theta\\ &+n^{-1/2}\mathcal{W}_n+o(n^{-1/2}),\end{aligned}$$

where $\mathcal{W}_m=(\Gamma_m^\top\Gamma_m+2\sigma_\eta^2\Gamma_p+\sigma_\eta^4 I)^{-1}\{(\Gamma_m+B_m)^\top\Psi_m+\mathcal{E}_m^\top\Gamma_m\theta\}-(\Gamma_m^\top\Gamma_m+2\sigma_\eta^2\Gamma_p+\sigma_\eta^4 I)^{-2}\cdot\{(\Gamma_m+B_m)^\top\mathcal{E}_m+\mathcal{E}_m^\top(\Gamma_m+B_m)\}(\Gamma_m^\top\Gamma_m+\sigma_\eta^2\Gamma_p)$ is normally distributed. We have proved the second part. □

THEOREM C. *Suppose the system* $\{x_t = g_{\theta_0}(x_{t-1}, \ldots, x_{t-p})\}$ *has a finite-dimensional state–space and admits only limit cycles, but* x_t *is observed as* $y_t = x_t + \eta_t$, *where* $\{\eta_t\}$ *are independent with mean* 0. *Suppose that the function* $g_\theta(v_1, \ldots, v_p)$ *has bounded derivatives in both* θ *in the parameter space* Θ *and* $v_1, \ldots, v_p$ *in a neighborhood of the state–space. Suppose that the system* $z_t = g_\theta(z_{t-1}, \ldots, z_{t-p})$ *has only negative Lyapunov exponents in a small neighborhood of* $\{x_t\}$ *and in* $\theta \in \Theta$. *Let* $X_t = (x_t, x_{t-1}, \ldots, x_{t-p})$ *and* $Y_t = (y_t, y_{t-1}, \ldots, y_{t-p})$.

1. *If the observed* $Y_0 = X_0 + (\eta_0, \eta_{-1}, \ldots, \eta_{-p})$ *is taken as the initial values of* $\{x_t\}$, *then for any* n,
$$f(y_{m+1}, \ldots, y_{m+n}|X_0) - f(y_{m+1}|X_0 = Y_0) \cdots f(y_{m+n}|X_0 = Y_0) \to 0$$
as $m \to \infty$.
2. *Suppose the equation* $\sum_{X_{t-1}} \{g_\theta(X_{t-1}) - x_t\}^2 = 0$ *has a unique solution in* θ, *where the summation is taken over all limiting states. Let* $\theta_{\{m\}} = \arg\min_\theta m^{-1} \sum_{k=1}^m \mathbf{E}\{\, y_{t-1+k} - g_\theta^{[k]}(Y_{t-1})\}^2$. *If the noise takes value in a small neighborhood of the origin, then* $\theta_{\{m\}} \to \theta_0$ *as* $m \to \infty$.

PROOF. Let $Y_{t-1} = (y_{t-1}, \ldots, y_{t-p})$, $\mathcal{E}_{t-1} = (\eta_{t-1}, \ldots, \eta_{t-p})$ and $X_{t-1} = (x_{t-1}, \ldots, x_{t-p})$. By the condition, we have $x_t = g_{\theta_0}(X_{t-1})$. Write

$$\begin{aligned}
&\mathbf{E}\big[\{g_\theta^{[k]}(Y_{t-1}) - x_{t-1+k}\}^2\big] \\
&\quad = \{g_\theta^{[k]}(X_{t-1}) - g_{\theta_0}^{[k]}(X_{t-1})\}^2 \\
&\qquad - 2\{g_\theta^{[k]}(X_{t-1}) - g_{\theta_0}^{[k]}(X_{t-1})\} \\
&\qquad\quad \cdot \mathbf{E}\{g_\theta^{[k]}(X_{t-1} + \mathcal{E}_{t-1}) - g_\theta^{[k]}(X_{t-1})\} \\
&\qquad + \mathbf{E}\big[\{g_\theta^{[k]}(X_{t-1} + \mathcal{E}_{t-1}) - g_\theta^{[k]}(X_{t-1})\}^2\big].
\end{aligned}$$

Note that by the definition of the Lyapunov exponent,

$$\begin{aligned}
&|g_\theta^{[k]}(X_t + \mathcal{E}_t) - g_\theta^{[k]}(X_t)| \\
&\qquad \le \exp(k\lambda)\{\mathbf{E}\|\mathcal{E}_t\|\}^{1/2}.
\end{aligned} \tag{A.5}$$

Starting from $X_0 = Y_0$, the system at the kth step is $g^{[k]}(Y_0)$. Since the Lyapunov exponent is negative, we have

$$\begin{aligned}
&\big(g_{\theta_0}^{[m+1]}(Y_0), \ldots, g_{\theta_0}^{[m+n]}(Y_0)\big) \\
&\quad = \big(g_{\theta_0}^{[m+1]}(X_0), \ldots, g_{\theta_0}^{[m+n]}(X_0)\big) \\
&\qquad + (\delta_{m+1}, \ldots, \delta_{m+n}),
\end{aligned}$$

where $\delta_k = g_{\theta_0}^{[k]}(Y_0) - g_{\theta_0}^{[k]}(X_0)$, with $|\delta_k| \le \exp(k\lambda) \cdot \{\mathbf{E}\|\mathcal{E}_0\|\}^{1/2}$. Therefore,

$$\begin{aligned}
&(y_{m+1}, \ldots, y_{m+n})|(X_0 = Y_0) \\
&\quad = (y_{m+1}, \ldots, y_{m+n})|X_0 + (\delta_{m+1}, \ldots, \delta_{m+n}).
\end{aligned}$$

Note that $(y_{m+1}, \ldots, y_{m+n})|X_0 = (g_{\theta_0}^{[m+1]}(X_0), \ldots, g_{\theta_0}^{[m+n]}(X_0)) + (\eta_{m+1}, \ldots, \eta_{m+n})$ and that $\eta_{m+1}, \ldots, \eta_{m+n}$ are independent. Therefore the first part of Theorem C follows.

By (A.5), we have

$$\begin{aligned}
&\big|\mathbf{E}\big[\{g_\theta^{[k]}(Y_t) - x_{t+k}\}^2\big] - \{g_\theta^{[k]}(X_t) - g_{\theta_0}^{[k]}(X_t)\}^2\big| \\
&\quad \le C\exp(k\lambda)\{\mathbf{E}\|\mathcal{E}_t\|\}^{1/2}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&\Big| m^{-1} \sum_{k=1}^m \mathbf{E}\{x_{t-1+k} - g_\theta^{[k]}(Y_t)\}^2 \\
&\qquad - m^{-1} \sum_{k=1}^m \{g_\theta^{[k]}(X_t) - g_{\theta_0}^{[k]}(X_t)\}^2 \Big| \\
&\qquad \le C\{\mathbf{E}\|\mathcal{E}_t\|\}^{1/2} m^{-1} \sum_{k=1}^m \exp(k\lambda) \\
&\qquad \equiv \Delta(m) \to 0 \quad \text{as } m \to \infty.
\end{aligned}$$

That is,

$$\begin{aligned}
&m^{-1} \sum_{k=1}^m \{g_\theta^{[k]}(X_t) - g_{\theta_0}^{[k]}(X_t)\}^2 - \Delta(m) \\
&\qquad \le m^{-1} \sum_{k=1}^m \mathbf{E}\{x_{t-1+k} - g_\theta^{[k]}(Y_t)\}^2 \\
&\qquad \le m^{-1} \sum_{k=1}^m \{g_\theta^{[k]}(X_t) - g_{\theta_0}^{[k]}(X_t)\}^2 + \Delta(m).
\end{aligned} \tag{A.6}$$

By the second inequality of (A.6) and the continuity, we have as $\theta \to \theta_0$ and $m \to \infty$,

$$m^{-1} \sum_{k=1}^m \mathbf{E}\{x_{t-1+k} - g_\theta^{[k]}(Y_{t-1})\}^2 \to 0. \tag{A.7}$$

Next, we show that if $\|\theta - \theta_0\| \ge \delta > 0$, then as $m \to \infty$ there exists $\delta' > 0$ such that

$$m^{-1} \sum_{k=1}^m \{g_\theta^{[k]}(X_t) - g_{\theta_0}^{[k]}(X_t)\}^2 \ge \delta' > 0. \tag{A.8}$$

We prove (A.8) by contradiction. Suppose the period of the limit cycle is π. For continuous dynamics, the as-

sumption of a unique solution is equivalent to the statement that as $i \to \infty$,

$$\sum_{k=i+1}^{i+\pi} \{g_\theta(X_{k-1}) - x_k\}^2 \to 0 \iff \theta \to \theta_0. \tag{A.9}$$

If (A.8) does not hold, that is, there is a ϑ such that

$$m^{-1} \sum_{k=1}^{m} \mathbf{E}\{g_\vartheta^{[k]}(X_t) - x_{t-1+k}\}^2 \to 0,$$

then there must be a sequence $\{i_j : j = 1, 2, \ldots\}$ with $i_j \to \infty$ as $j \to \infty$ and

$$\sum_{k=i_j-p}^{i_j+\pi} \{g_\vartheta^{[k]}(X_t) - x_{t+k}\}^2 \to 0 \quad \text{as } j \to \infty. \tag{A.10}$$

Let $z_{t+k} = g_\vartheta^{[k]}(X_t)$ and $e_{t+k} = z_{t+k} - x_{t+k}$. It follows from (A.10) that for $k = i_j - p, \ldots, i_j + \pi$,

$$|e_{t+k}| \to 0 \quad \text{as } j \to \infty,$$

and that

$$\sum_{k=i_j+1}^{i_j+\pi} \{g_\vartheta(x_{t+k-1} + e_{t+k-1}, \ldots, x_{t+k-p} + e_{t+k-p}) - x_{t+k}\}^2 \to 0.$$

By the same argument leading to (A.6), we have

$$\begin{aligned}
&\sum_{k=i_j+1}^{i_j+\pi} \{g_\vartheta(x_{t+k-1}, \ldots, x_{t+k-p}) - x_{t+k}\}^2 \\
&\quad \geq \sum_{k=i_j+1}^{i_j+\pi} \{g_\vartheta(x_{t+k-1} + e_{t+k-1}, \ldots, x_{t+k-p} + e_{t+k-p}) - x_{t+k}\}^2 \\
&\qquad - C(e_{t+i_j-p}^2 + \cdots + e_{t+i_j+\pi}^2) \\
&\quad \to 0
\end{aligned}$$

for some $C > 0$. Let $j \to \infty$; we have $\sum_{k=i_j+1}^{i_j+\pi} \{g_\vartheta(x_{t+k-1}, \ldots, x_{t+k-p}) - x_{t+k}\}^2 = 0$, which contradicts the assumption of a unique solution (A.9).

By (A.6), (A.7) and (A.8), we have completed the proof of Theorem C. □

THEOREM D. *Recall the notation in Section 3.2 and let* $\mathcal{E}_t = (\varepsilon_t, 0, \ldots, 0)^\top$ *and* $\mathcal{N}_t = (\eta_t, \ldots, \eta_{t-p+1})^\top$. *For the nonlinear skeleton, we further assume that* $g_\theta(x)$ *has bounded second-order derivative with respect to* θ *in neighbor of* ϑ *for all possible values of* y_t. *Suppose that the assumptions* (C1)–(C4) *hold. Then*

$$T^{-1/2}(\tilde{\theta}_{\{m\}} - \vartheta_{m,\mathbf{w}}) \xrightarrow{D} N(0, \Omega^{-1}\Lambda(\Omega^{-1})^\top).$$

Specifically, for model (3.1) *and* $y_t = x_t + \eta_t$, *if* $E|\varepsilon_t|^\delta < \infty$ *and* $E|\eta_t|^\delta < \infty$ *for some* $\delta > 4$, *then*

$$\begin{aligned}
\Lambda &= \operatorname{Cov}(\Delta_t, \Delta_t) \\
&\quad + \sum_{k=1}^{\infty} \{\operatorname{Cov}(\Delta_t, \Delta_{t-k}) + \operatorname{Cov}(\Delta_{t-k}, \Delta_t)\}, \\
\Omega &= \sum_{k=1}^{m} w_k \mathbf{E}\Big[\frac{\partial g_\vartheta^{[k]}(y_t)}{\partial \theta} \frac{\partial g_\vartheta^{[k]}(y_t)}{\partial \theta^\top} \\
&\qquad - e_1^\top(\Phi^k - \Psi^k) X_t \frac{\partial^2 g_\vartheta^{[k]}(X_t)}{\partial \theta \partial \theta^\top} \\
&\qquad + e_1^\top \Phi^k \mathcal{N}_t \frac{\partial^2 g_\vartheta^{[k]}(\mathcal{N}_t)}{\partial \theta \partial \theta^\top} \Big]
\end{aligned}$$

with $\Delta_t = \sum_{k=1}^m w_k \{\sum_{j=0}^{k-1} e_1^\top \Phi^j e_1 \varepsilon_{t+k-j} + \eta_{t+k}\} \cdot \partial g_\vartheta^{[k]}(y_t)/\partial\theta + \{e_1^\top(\Phi^k - \Psi^k) X_t - e_1^\top \Phi^k \mathcal{N}_t\} \cdot \partial g_\vartheta^{[k]}(y_t)/\partial\theta$. *For the nonlinear model* (3.5) *and* $y_t = x_t + \eta_t$,

$$\begin{aligned}
\Lambda &= \operatorname{Var}\Big[\sum_{k=1}^{m} w_k \frac{\partial g_\vartheta^{[k]}(y_{t-k})}{\partial \theta} \eta_t \\
&\qquad + \sum_{k=1}^{m} w_k [g_{\theta_0}^{[k]}(X_t) - g_\vartheta^{[k]}(y_t)] \frac{\partial g_\vartheta^{[k]}(y_t)}{\partial \theta} \Big]
\end{aligned}$$

and

$$\Omega = \sum_{k=1}^{m} w_k \mathbf{E}\Big\{ \frac{\partial g_\vartheta^{[k]}(y_t)}{\partial \theta} \frac{\partial g_\vartheta^{[k]}(y_t)}{\partial \theta^\top} \Big\}.$$

PROOF. Let $Q(\theta) = \sum_{k=1}^m w_k \mathbf{E}[y_{t+k} - g_\theta^{[k]}(Y_t)]^2$ and

$$\begin{aligned}
Q_n(\theta) &= \sum_{k=1}^{m} w_k T^{-1} \sum_{t=1}^{T} [y_{t+k} - g_\theta^{[k]}(Y_t)]^2 \\
&\cong \sum_{k=1}^{m} w_k \frac{1}{T-k} \sum_{t=1}^{T-k} [y_{t+k} - g_\theta^{[k]}(Y_t)]^2 \\
&\stackrel{\text{def}}{=} Q_n(\theta).
\end{aligned}$$

Let $\tilde{\theta}_{\{m\}} = \arg\min_{\theta\in\Theta} Q_n(\theta)$. We denote this by $\tilde{\theta}$ and $\vartheta_{m,w}$ by ϑ, for simplicity. It is easy to see that $Q_n(\theta) \to Q(\theta)$. Following the same argument of Wu (1981), we have $\tilde{\theta} \to \vartheta$ in probability.

By the definition of $\tilde{\theta}$, we have $\partial Q_n(\tilde{\theta})/\partial\theta = 0$. By Taylor expansion, we have

$$\begin{aligned} 0 &= \frac{\partial Q_n(\tilde{\theta})}{\partial\theta} \\ &= \frac{\partial Q_n(\vartheta)}{\partial\theta} + \frac{\partial^2 Q_n(\theta^*)}{\partial\theta\partial\theta^\top}(\tilde{\theta}-\vartheta), \end{aligned} \tag{A.11}$$

where θ^* is a vector between $\tilde{\theta}$ and ϑ, and

$$\begin{aligned} &\frac{\partial Q_n(\vartheta)}{\partial\theta} \\ &\quad= -2T^{-1}\sum_{k=1}^{m} w_k \sum_{t=1}^{T}\bigl[y_{t+k} - g_\vartheta^{[k]}(Y_t)\bigr] \\ &\qquad\qquad \cdot \frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta} \\ &\quad= 2T^{-1}\sum_{t=1}^{T}\xi_{t,m}, \end{aligned} \tag{A.12}$$

where $\xi_{t,m} = \sum_{k=1}^{m} w_k[y_{t+k} - g_\vartheta^{[k]}(Y_t)]\partial g_\vartheta^{[k]}(Y_t)/\partial\theta$. By the definition of ϑ, we have $\partial Q(\vartheta)/\partial\theta = 0$, that is,

$$E\xi_{t,m} = 0. \tag{A.13}$$

Since y_t is a strongly mixing process with exponential decreasing mixing coefficients, so is $\xi_{t,m}$. By (C2), we have $\mathbf{E}\|\xi_{t,m}\|^\delta < \infty$. It follows from Theorem 2.21 of Fan and Yao [(2003), page 75] that

$$\sum_{t=1}^{T}\Delta_t/\sqrt{T} \xrightarrow{D} N\Biggl(0, \sum_{k=0}^{\infty}\Gamma_\Delta(k)\Biggr).$$

On the other hand, we have by (C3) and Proposition 2.8 of Fan and Yao [(2003), page 74]

$$\begin{aligned} &\frac{\partial^2 Q_n(\theta^*)}{\partial\theta\partial\theta} \\ &\cong 2T^{-1}\sum_{t=1}^{T}\sum_{k=1}^{m} w_k\biggl\{\frac{\partial g_{\theta^*}^{[k]}(Y_t)}{\partial\theta}\frac{\partial g_{\theta^*}^{[k]}(Y_t)}{\partial\theta^\top} \\ &\qquad\qquad - \bigl[y_{t+k} - g_\vartheta^{[k]}(Y_t)\bigr]\frac{\partial^2 g_{\theta^*}^{[k]}(Y_t)}{\partial\theta\partial\theta^\top}\biggr\} \\ &\to 2\Omega. \end{aligned}$$

For model (3.1), we have $X_{t+1} = \Phi X_t + \mathcal{E}_{t+1}$ and $X_{t+k} = \Phi^k X_t + (\mathcal{E}_{t+k} + \Phi\mathcal{E}_{t+k-1} + \cdots + \Phi^{k-1}\mathcal{E}_{t+1})$. Let Ψ be the matrix Φ when $\theta = \vartheta$, respectively. Note that $Y_t = X_t + \mathcal{N}_t$. It follows that

$$\begin{aligned} &y_{t+k} - \Psi^k Y_t \\ &\quad= (x_{t+k} + \mathcal{N}_{t+k}) - \Phi^k(X_t + \mathcal{N}_t) + (\Phi^k - \Psi^k)Y_k \\ &\quad= (\mathcal{E}_{t+k} + \Phi\mathcal{E}_{t+k-1} + \cdots + \Phi^{k-1}\mathcal{E}_{t+1}) \\ &\qquad + (\mathcal{N}_{t+k} - \Phi^k\mathcal{N}_t) + (\Phi^k - \Psi^k)Y_t \end{aligned}$$

and

$$\begin{aligned} &y_{t+k} - e_1^\top\Psi^k Y_t \\ &\quad= \sum_{j=0}^{k-1} e_1^\top\Phi^j e_1\varepsilon_{t+k-j} + \eta_{t+k} \\ &\qquad + e_1^\top(\Phi^k - \Psi^k)X_t - e_1^\top\Phi^k\mathcal{N}_t. \end{aligned} \tag{A.14}$$

It follows from (A.13) and (A.14) that

$$2\sum_{k=1}^{m} w_k\mathbf{E}e_1\biggl[\{(\Phi^k - \Psi^k)Y_t - \Phi^k\mathcal{N}_t\}\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta}\biggr] = 0.$$

We have

$$\begin{aligned} &\frac{\partial Q_n(\vartheta)}{\partial\theta} \\ &\cong -2T^{-1}\sum_{t=1}^{T}\sum_{k=1}^{m} w_k\Biggl[\Biggl\{\sum_{j=0}^{k-1} e_1^\top\Phi^j e_1\varepsilon_{t+k-j} + \eta_{t+k}\Biggr\} \\ &\qquad\qquad \cdot \frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta} \\ &\qquad + \Biggl\{\{e_1^\top(\Phi^k - \Psi^k)X_t - e_1^\top\Phi^k\mathcal{N}_t\}\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta} \\ &\qquad\qquad - \mathbf{E}\biggl[\{e_1^\top(\Phi^k - \Psi^k)X_t - e_1^\top\Phi^k\mathcal{N}_t\} \\ &\qquad\qquad\qquad \cdot \frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta}\biggr]\Biggr\}\Biggr] \\ &\stackrel{\text{def}}{=} -2T^{-1}\sum_{t=1}^{T}\Delta_t \end{aligned}$$

and that $\mathbf{E}\Delta_t = 0$. Let $\tilde{\partial}_k = \partial(e_1^\top\Psi^k)/\partial\theta$. We further have

$$\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta} = \frac{\partial e_1^\top\Psi^k}{\partial\theta}Y_t = \tilde{\partial}_k(X_t + \mathcal{N}_t).$$

Since $(X_t, \varepsilon_t, \eta_t)$ is a stationary process and a strongly mixing sequence (Pham and Tran, 1985) with exponentially decreasing mixing coefficients, and Δ_t is a function of $\{(X_\tau, \varepsilon_\tau, \eta_\tau) : \tau = t, t-1, \ldots, t-m\}$, it is easy

to see that Δ_t is also a strongly mixing sequence with exponentially decreasing mixing coefficients. Note that $\mathbf{E}\Delta_t = 0$ and $\mathbf{E}|\Delta_t|^\delta < \infty$ for some $\delta > 2$. By Theorem 2.21 of Fan and Yao [(2003), page 75], we have

$$\sum_{t=1}^{T} \Delta_t/\sqrt{T} \xrightarrow{D} N\Bigg(0, \sum_{k=0}^{\infty} \Gamma_\Delta(k)\Bigg).$$

On the other hand, we have in probability

$$\begin{aligned}
&\frac{\partial^2 Q_n(\vartheta)}{\partial\theta\partial\theta^\top} \\
&\quad = 2T^{-1}\sum_{k=1}^{m} w_k \sum_{t=1}^{T} \frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta}\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta^\top} \\
&\qquad - 2T^{-1}\sum_{k=1}^{m} w_k \sum_{t=1}^{T}\Bigg\{\sum_{j=0}^{k-1} e_1^\top \Phi^j e_1 \varepsilon_{t+k-j} + \eta_{t+k} \\
&\qquad\quad + e_1^\top(\Phi^k - \Psi^k)X_t - e_1^\top \Phi^k \mathcal{N}_t\Bigg\}\frac{\partial^2 g_\vartheta^{[k]}(Y_t)}{\partial\theta\partial\theta^\top} \\
&\quad \to 2\sum_{k=1}^{m} w_k \mathbf{E}\Bigg\{\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta}\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta^\top}\Bigg\} \\
&\qquad - 2\sum_{k=1}^{m} w_k \mathbf{E}\Bigg[e_1^\top(\Phi^k - \Psi^k)X_t \frac{\partial^2 g_\vartheta^{[k]}(X_t)}{\partial\theta\partial\theta^\top}\Bigg] \\
&\qquad + 2\sum_{k=1}^{m} w_k \mathbf{E}\Bigg[e_1^\top \Phi^k \mathcal{N}_t \frac{\partial^2 g_\vartheta^{[k]}(\mathcal{N}_t)}{\partial\theta\partial\theta^\top}\Bigg] \\
&\quad \stackrel{\text{def}}{=} 2\Omega.
\end{aligned}$$

Therefore, it follows from (A.11) that

$$T^{-1/2}(\tilde\theta - \vartheta) \xrightarrow{D} N\Bigg\{0, \Omega^{-1}\sum_{k=0}^{\infty}\Gamma_\Delta(k)(\Omega^{-1})^\top\Bigg\}.$$

Next, consider model (3.5). Note that $\eta_{t+k} = y_{t+k} - g_{\theta_0}^{[k]}(X_t)$. We have from (A.12) that

$$\begin{aligned}
&\frac{\partial Q_n(\vartheta)}{\partial\theta} \\
&\quad = -2T^{-1}\sum_{k=1}^{m} w_k \sum_{t=1}^{T} \eta_{t+k}\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta} \\
&\qquad -2T^{-1}\sum_{k=1}^{m} w_k \sum_{t=1}^{T}\big[g_{\theta_0}^{[k]}(X_t) - g_\vartheta^{[k]}(Y_t)\big] \\
&\qquad\qquad \cdot \frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta} \\
&\quad \cong -2T^{-1}\sum_{t=1}^{T}\Bigg\{\Bigg[\sum_{k=1}^{m} w_k \frac{\partial g_\vartheta^{[k]}(Y_{t-k})}{\partial\theta}\Bigg]\eta_t \\
&\qquad\qquad + \sum_{k=1}^{m} w_k\big[g_{\theta_0}^{[k]}(X_t) - g_\vartheta^{[k]}(Y_t)\big] \\
&\qquad\qquad \cdot \frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta}\Bigg\}.
\end{aligned}$$

Let

$$\begin{aligned}
C_m(x_{t-k}, \eta_{t-k}) &= \Bigg[\sum_{k=1}^{m} w_k \frac{\partial g_\vartheta^{[k]}(Y_{t-k})}{\partial\theta}\Bigg], \\
B_m(x_t, \eta_t) &= \sum_{k=1}^{m} w_k\big[g_{\theta_0}^{[k]}(X_t) - g_\vartheta^{[k]}(Y_t)\big] \\
&\quad \cdot \frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta}.
\end{aligned}$$

By (A.13), we have $\mathbf{E}B_m(X_t, \eta_t) = 0$. Thus $B_m(x_t, \eta_t)$ are independent with expectation 0. It is easy to see that $\xi_{m,t} = C_m(X_{t-k}, \eta_{t-k})\eta_t + B_m(X_t, \eta_t)$ is a martingale difference. The Lyapunov's condition is satisfied. Thus, we have

$$T^{-1/2}\sum_{t=1}^{T}\xi_{t,m} \xrightarrow{D} N\{0, \mathbf{E}(\xi_{m,t}\xi_{m,t}^\top)\}. \tag{A.15}$$

Similarly to $\partial Q_n(\vartheta)/\partial\theta$ above, we have

$$\begin{aligned}
&\frac{\partial^2 Q_n(\theta^*)}{\partial\theta\partial\theta} \\
&\quad \cong -2T^{-1}\sum_{t=1}^{T}\Bigg[\sum_{k=1}^{m} w_k \frac{\partial^2 g_{\theta^*}^{[k]}(Y_{t-k})}{\partial\theta\partial\theta^\top}\Bigg]\eta_t \\
&\qquad + 2T^{-1}\sum_{t=1}^{T}\sum_{k=1}^{m} w_k \frac{\partial g_{\theta^*}^{[k]}(Y_t)}{\partial\theta}\frac{\partial g_{\theta^*}^{[k]}(Y_t)}{\partial\theta^\top}
\end{aligned} \tag{A.16}$$

$$\begin{aligned}
&\quad \to 2\sum_{k=1}^{m} w_k \mathbf{E}\Bigg\{\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta}\frac{\partial g_\vartheta^{[k]}(Y_t)}{\partial\theta^\top}\Bigg\} \\
&\quad \stackrel{\text{def}}{=} 2\Omega.
\end{aligned} \tag{A.17}$$

Finally, from (A.11), (A.15) and (A.16) we have

$$T^{-1/2}(\tilde\theta - \vartheta) \xrightarrow{D} N\{0, \Omega^{-1}\mathbf{E}(\xi_{m,t}\xi_{m,t}^\top)\Omega^{-1}\}.$$

We have completed the proof. □

ACKNOWLEDGMENTS

Yingcun Xia's research is supported in part by a grant from the Risk Management Institute, National

University of Singapore. Howell Tong gratefully acknowledges partial support from the National University of Singapore (Saw Swee Hock Professorship) and the University of Hong Kong (Distinguished Visiting Professorship). We are grateful to the Executive Editor and two anonymous referees for constructive comments. We are also grateful to the Institute of Mathematical Science, National University of Singapore, for giving us the opportunity to present our work at their Workshop on Nonlinear Time Series Analysis in February, 2011.

REFERENCES

AKAIKE, H. (1978). On the likelihood of a time series model. *The Statistician* **27** 217–235.

ALLIGOOD, K. T., SAUER, T. D. and YORKE, J. A. (1997). *Chaos*: *An Introduction to Dynamical Systems*. Springer, New York. MR1418166

ANDERSON, R. M. and MAY, R. M. (1991). *Infectious Diseases of Humans*: *Dynamics and Control*. Oxford Univ. Press, Oxford.

BAILEY, N. T. J. (1957). *The Mathematical Theory of Epidemics*. Hafner Publishing Co., New York. MR0095085

BARTLETT, M. S. (1956). Deterministic and stochastic models for recurrent epidemics. In *Proc. Third Berkeley Symp. Math. Statist. Probab.* **IV** 81–109. Univ. California Press, Berkeley. MR0084932

BARTLETT, M. S. (1957). Measles periodicity and community size. *J. Roy. Statist. Soc. Ser. A* **120** 48–70.

BARTLETT, M. S. (1960). The critical Community size for measles in the United States. *J. Roy. Statist. Soc. Ser. A* **123** 37–44.

BHANSALI, R. J. and KOKOSZKA, P. S. (2002). Computation of the forecast coefficients for multistep prediction of long-range dependent time series. *Int. J. Forecasting* **18** 181–206.

BJØNSTAD, O. N., FINKENSTÄDT, B. and GRENFELL, B. T. (2002). Dynamics of measles epidemics: Estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs* **72** 169–184.

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

BOX, G. E. P. and JENKINS, G. M. (1970). *Times Series Analysis. Forecasting and Control*. Holden-Day, San Francisco, CA. MR0272138

BROCKWELL, P. J. and DAVIS, R. A. (1991). *Time Series*: *Theory and Methods*, 2nd ed. Springer, New York. MR1093459

CANOVA, F. (2007). *Methods for Applied Macroeconomic Research*. Princeton Univ. Press, Princeton.

CHAN, K.-S. and TONG, H. (2001). *Chaos*: *A Statistical Perspective*. Springer, New York. MR1851668

CHAN, K.-S., TONG, H. and STENSETH, N. C. (2009). Analyzing short time series data from periodically fluctuating rodent populations by threshold models: A nearest block bootstrap approach (with discussion). *Sci. China Ser. A* **52** 1085–1112. MR2520564

CHEN, R., YANG, L. and HAFNER, C. (2004). Nonparametric multistep-ahead prediction in time series analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 669–686. MR2088295

CHENG, B. and TONG, H. (1992). On consistent nonparametric order determination and chaos (with discussion). *J. Roy. Statist. Soc. Ser. B* **54** 427–474. MR1160478

COX, D. R. (1961). Prediction by exponentially weighted moving averages and related methods. *J. Roy. Statist. Soc. Ser. B* **23** 414–422. MR0137160

DURBIN, J. and KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. *Oxford Statistical Science Series* **24**. Oxford Univ. Press, Oxford. MR1856951

DYE, C. and GAY, N. (2003). Modeling the SARS epidemic. *Science* **300** 1884–1885.

EARN, D. J. D., ROHANI, P., BOLKER, B. M. and GRENFELL, B. T. (2000). A simple model for complex dynamical transitions in epidemics. *Science* **287** 667–670.

ELLNER, S. P., SEIFU, Y. and SMITH, R. H. (2002). Fitting population-dynamic models to time-series data by gradient matching. *Ecology* **83** 2256–2270.

FAN, J. and YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

FAN, J. and ZHANG, W. (2004). Generalised likelihood ratio tests for spectral density. *Biometrika* **91** 195–209. MR2050469

FINKENSTÄDT, B. F. and GRENFELL, B. T. (2000). Time series modelling of childhood diseases: A dynamical systems approach. *J. Roy. Statist. Soc. Ser. C* **49** 187–205. MR1821321

FRIEDLANDER, B. and SHARMAN, K. C. (1985). Performance evaluation of the modified Yule-Walker estimator. *IEEE Trans. Acoust., Speech, Signal Process.* **33** 719–725.

GEORGIOU, T. T. (2007). Distances and Riemannian metrics for spectral density functions. *IEEE Trans. Signal Process.* **55** 3995–4003. MR2464411

GLASS, K., XIA, Y. and GRENFELL, B. T. (2003). Interpreting time-series analyses for continuous-time biological models—Measles as a case study. *J. Theoret. Biol.* **223** 19–25. MR2069237

GRENFELL, B. T., BJØRNSTAD, O. N. and FINKENSTÄDT, B. (2002). Dynamics of measles epidemics: Scaling noise, determinism and predictability with the TSIR model. *Ecological Monographs* **72** 185–202.

GUO, M., BAI, Z. and AN, H. Z. (1999). Multi-step prediction for nonlinear autoregression models based on empirical distributions. *Statist. Sinica* **9** 559–570. MR1707854

GURNEY, W. S. C., BLYTHE, P. B. and NISBET, R. M. (1980). Nicholson's Blowflies revisited. *Nature* **287** 17–21.

HALL, A. R. (2005). *Generalized Method of Moments*. Oxford Univ. Press, Oxford. MR2135106

HE, D., IONIDES, E. L. and KING, A. A. (2010). Plug-and-play inference for disease dynamics: Measles in large and small towns as a case study. *J. Roy. Soc. Interface* **7** 271–283.

HETHCOTE, H. W. (1976). Qualitative analyses of communicable disease models. *Math. Biosci.* **28** 335–356. MR0401216

ISHAM, V. and MEDLEY, G. (2008). *Models for Infectious Human Diseases*: *Their Structure and Relation to Data*. Cambridge Univ. Press, Cambridge.

KEELING, M. J. and GRENFELL, B. T. (1997). Disease extinction and community size: Modeling the persistence of measles. *Science* **275** 65–67.

KING, A. A., IONDIDES, E. L., PASCUAL, M. and BOUMA, M. J. (2008). Inapparent infections and cholera dynamics. *Nature* **454** 877–880.

KYDLAND, F. E. and PRESCOTT, E. C. (1996). The computational experiment: An econometric tool. *J. Economic Perspectives* **10** 69–85.

LANERI, K., BHADRA, A., IONIDES, E. L., BOUMA, M., YADAV, R., DHIMAN, R. and PASCUAL, M. (2010). Forcing versus feedback: Epidemic malaria and monsoon rains in NW India. *PLoS Comput. Biol.* **6** e1000898.

LIU, W. M., HETHCOTE, H. W. and LEVIN, S. A. (1987). Dynamical behavior of epidemiological models with nonlinear incidence rates. *J. Math. Biol*. **25** 359–380. MR0908379

MAN, K. S. (2002). Long memory time series and short tem forecasts. *Int. J. Forecasting* **19** 477–491.

MAY, R. M. (1976). Simple mathematical models with very complicated dynamics. *Nature* **261** 459–467.

NICHOLSON, A. J. and BAILEY, V. A. (1935). The balance of animal populations. Part 1. *Proc. Zool. Soc. London* **1** 551–598.

OSTER, G. and IPAKTCHI, A. (1978). Population cycles. In *Periodicitie in Chemistry and Biology* (H. Eyring, ed.) 111–132. Academic Press, New York.

PARZEN, E. (1962). *Stochastic Processes*. Holden-Day, San Francisco, CA. MR0139192

PHAM, T. D. and TRAN, L. T. (1985). Some mixing properties of time series models. *Stochastic Process. Appl*. **19** 297–303.

ROHANI, P., GREEN, C. J., MANTILLA-BENIERS, N. B. and GRENFELL, B. T. (2003). Ecological interference between fatal diseases. *Nature* **422** 885–888.

ROMANO, J. P. and THOMBS, L. A. (1996). Inference for autocorrelations under weak assumptions. *J. Amer. Statist. Assoc*. **91** 590–600. MR1395728

SAKAI, H., SOEDA, T. and TOKUMARU, H. (1979). On the relation between fitting autoregression and periodogram with applications. *Ann. Statist*. **7** 96–107. MR0515686

SLUTSKY, E. (1927). The summation of random causes as the source of cyclic processes. *Econometrica* **5** 105–146.

STAUDENMAYER, J. and BUONACCORSI, J. P. (2005). Measurement error in linear autoregressive models. *J. Amer. Statist. Assoc*. **100** 841–852. MR2201013

STOICA, P., MOSES, R. L. and LI, J. (1991). Optimal higher-order Yule-Walker estimation of sinusoidal frequencies. *IEEE Trans. Signal Process.* **39** 1360–1368.

STOKES, T. G., GURNEY, W. S. C., NISBET, R. M. and BLYTHE, S. P. (1988). Parameter evolution in a laboratory insect population. *Theor. Pop. Biol*. **34** 248–265.

TIAO, G. C. and XU, D. (1993). Robustness of maximum likelihood estimates for multi-step predictions: The exponential smoothing case. *Biometrika* **80** 623–641. MR1248027

TONG, H. (1990). *Nonlinear Time Series*: *A Dynamical System Approach*. *Oxford Statistical Science Series* **6**. Oxford Univ. Press, New York. MR1079320

TONG, H. and LIM, K. S. (1980). Threshold autoregression, limit cycles and cyclical data (with discussion). *J. Roy. Statist. Soc. Ser. B* **42** 245–292.

TSAY, R. S. (1992). Model checking via parametric bootstraps in time series analysis. *J. Roy. Statist. Soc. Ser. C* **41** 1–15.

VARLEY, G. C., GRADWELL, G. R. and HASSELL, M. P. (1973). *Insect Population Ecology*. Univ. California Press, Berkeley.

WALKER, A. M. (1960). Some consequences of superimposed error in time series analysis. *Biometrika* **47** 33–43. MR0114284

WHITTLE, P. (1962). Gaussian estimation in stationary time series. *Bull. Inst. Internat. Statist*. **39** 105–129. MR0162345

WOOD, S. N. (2001). Partially specified ecological models. *Ecological Monographs* **71** 1–25.

WU, C. F. J. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist*. **9** 501–513.

YULE, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philos. Trans. R. Soc. Lond. Ser. A* **226** 267–298.