Biometrika (2014), **101**, 1, *pp.* 121–140
Printed in Great Britain
doi: 10.1093/biomet/ast051

Nonparametric estimation of a periodic sequence in the presence of a smooth trend

By MICHAEL VOGT and OLIVER LINTON

Faculty of Economics, University of Cambridge, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD, U.K.
mv346@cam.ac.uk obl20@cam.ac.uk

Summary

We investigate a nonparametric regression model including a periodic component, a smooth trend function, and a stochastic error term. We propose a procedure to estimate the unknown period and the function values of the periodic component as well as the nonparametric trend function. The theoretical part of the paper establishes the asymptotic properties of our estimators. In particular, we show that our estimator of the period is consistent. In addition, we derive the convergence rates and the limiting distributions of our estimators of the periodic component and the trend function. The asymptotic results are complemented with a simulation study and an application to global temperature anomaly data.

Some key words: Nonparametric estimation; Penalized least squares; Periodic sequence; Temperature anomaly data.

1. Introduction

Many time series exhibit a periodic as well as a trending behaviour. Examples come from fields as diverse as astronomy, climatology, population biology and economics. A common way to model such time series is to write them as the sum of a periodic component, a time trend and a noise process. When the periodic model part and the time trend have a parametric form, they can be estimated by a variety of standard methods; see e.g., Brockwell & Davis (1991) and Quinn & Hannan (2001) for overviews. However, usually there is not much known about the structure of the periodic and the trend components. It is thus important to have flexible semi- and nonparametric methods at hand to estimate them.

In this paper, we develop estimation theory for the periodic and the trend components in the following framework. Let $\{Y_{t,T} : t = 1, \ldots, T\}$ be the time series under investigation. The observations are assumed to follow the model

$$Y_{t,T} = g\left(\frac{t}{T}\right) + m(t) + \varepsilon_{t,T} \quad (t = 1, \ldots, T), \tag{1}$$

where $E(\varepsilon_{t,T}) = 0$, the function g is a smooth deterministic trend and m is a deterministic periodic component with unknown period θ_0. We do not impose any parametric restrictions on m and g. Moreover, we allow for nonstationarities and short-range dependence in the noise process $\{\varepsilon_{t,T}\}$. As usual in nonparametric regression, the time argument of the trend function g is rescaled to the unit interval. We comment on this in more detail in § 2 where we discuss the various model components.

The m-component in model (1) is assumed to be periodic in the following sense: the values $\{m(t)\}_{t\in\mathbb{Z}}$ form a periodic sequence with some unknown period θ_0, i.e., $m(t) = m(t+\theta_0)$ for some integer $\theta_0 \geqslant 1$ and all $t \in \mathbb{Z}$. Here and in what follows, θ_0 is implicitly assumed to be the smallest period of the sequence. As can be seen from this definition, we think of the periodic component in model (1) as a sequence rather than a function defined on the real line. The reason for taking this point of view is that there are an infinite number of functions on $\mathbb{R}$ that take the values $m(t)$ at the points $t \in \mathbb{Z}$. The function that generates these values is thus not identified in our framework. Moreover, if this function is periodic, θ_0 need not be its smallest period. It could also have θ_0/n for some $n \in \mathbb{N}$ as its period. Hence, in our design with equidistant observation points, we are in general able to identify neither the function that underlies the sequence values $\{m(t)\}_{t\in\mathbb{Z}}$ nor its smallest period. The best we can do is to work with the sequence $\{m(t)\}_{t\in\mathbb{Z}}$ and extract its periodic behaviour from the data.

So far, only a strongly simplified version of model (1) without a trend function has been considered in the literature. The setting is given by the equation $Y_t = m(t) + \varepsilon_t$, where the error terms ε_t are restricted to be stationary. Indeed, in many studies the errors are even assumed to be independent. The traditional way to estimate the periodic component m in this set-up is a trigonometric regression approach. This approach imposes a parametric structure on m. In particular, m is parameterized by a finite number of sinusoids. The underlying function that generates the sequence $\{m(t)\}_{t\in\mathbb{Z}}$ is thus known up to a finite number of coefficients, which in particular include the period of the function. The vector of parameters can be estimated by frequency domain methods based on the periodogram. Classical articles proceeding along these lines include Walker (1971), Rice & Rosenblatt (1988) and Quinn & Thomson (1991).

Only recently, there has been some work on estimating the periodic component in the setting $Y_t = m(t) + \varepsilon_t$ nonparametrically. Sun et al. (2012) consider the model with independent and identically distributed residuals ε_t and investigate the estimation of the unknown period of the sequence $\{m(t)\}_{t\in\mathbb{Z}}$. They view the issue of estimating the period as a model selection problem and construct a crossvalidation based procedure to solve it. Similar to the Akaike information criterion, their method is not consistent. Nevertheless, it enjoys a weakened version of consistency: roughly speaking, its asymptotic probability of selecting the true period is close to unity provided the period is not too small. This property is termed virtual consistency.

A related strand of the literature is concerned with nonparametrically estimating a periodic function with an unknown period when the observation points are not equally spaced in time. In this case, the model is given by $Y_t = m(X_t) + \varepsilon_t$, where m now denotes a periodic function defined on the real line, $X_1 < \cdots < X_T$ are the time-points of observation and the residuals ε_t are independent and identically distributed. The design points X_t may for example form a jittered grid, i.e., $X_t = t + U_t$ with variables U_t that are independent and uniformly distributed on $(-1/2, 1/2)$. Even though an equidistant design is the most common situation, such a random design is for example suitable for applications in astronomy, as described in Hall et al. (2000). The random design case is very different from that of equidistant observation points, because the function m can be identified without imposing any parametric restrictions on it. Roughly speaking, this is because the random design points are scattered all over the cycle of m as the sample size increases. Estimating the periodic function m in such a random design can be achieved by kernel-based least squares methods, as shown in Hall et al. (2000). Hall & Yin (2003) and Genton & Hall (2007) investigate some variants and extensions of this method. A periodogram-based approach is presented in Hall & Li (2006). Estimation theory for another possible sampling scheme is developed in Gassiat & Lévy-Leduc (2006).

In the models discussed so far, both the trend and the periodic component are deterministic functions of time. A markedly different approach is provided by unobserved components models

from the state space literature; see Harvey (1989) for a comprehensive overview. In these models, both the trend and the periodic component are stochastic in nature. It is hard to compare this approach with ours in theoretical terms, since they are nonnested. From a practical point of view, the two methods offer alternative ways to flexibly estimate the periodic and trend behaviour of a time series. In the unobserved components model, the flexibility comes through small stochastic innovations in the trend and the cycle. Our model in contrast owes its flexibility to the nonparametric nature of the deterministic component functions. An empirical comparison of the two approaches is provided in § 8.

In the following sections, we develop theory for estimating the unknown period θ_0, the sequence values $\{m(t)\}_{t\in\mathbb{Z}}$ and the trend function g in the general framework (1). Our estimation procedure is introduced in § 3 and splits into three steps. In the first step, we introduce a time domain approach to estimate the period θ_0. In particular, we combine least squares methods with an l_0-type penalization to construct an estimator of the period. Given our estimator of θ_0, we set up a least squares type estimator of the sequence values $\{m(t)\}_{t\in\mathbb{Z}}$ in the second step. The first two steps of our estimation procedure are complicated by the fact that the model includes a trend component. Importantly, our method is completely robust to the presence of a trend. As explained later on, the trend component g gets smoothed out by our procedure. We thus do not have to correct for the trend but can completely ignore it when estimating the periodic model part. In the third step of our procedure, we finally set up a kernel-based estimator of the nonparametric trend function g.

The asymptotic properties of our estimators are described in § 4. Our estimator of the period θ_0 is shown to be consistent. Moreover, we derive the convergence rates and asymptotic normality for the estimators of the periodic sequence values and the trend function. As will turn out, our estimator of the periodic sequence values has the same limiting distribution as the estimator in the oracle case where the true period θ_0 is known. A similar oracle property is derived for the estimator of the nonparametric trend function g.

In § 6, we investigate the small sample behaviour of our estimators in a simulation study. In addition, we apply our method to a sample of yearly global temperature anomalies from 1850 to 2011 in § 7. These data exhibit a strong warming trend. As suggested by various articles in climatology, they also contain a cyclical component with a period in the region of 60–70 years. We use our procedure to investigate whether there is in fact evidence for a cyclical component in the data. In addition, we provide estimates of the periodic sequence values and the trend function.

2. Model

Before we turn to our estimation procedure, we have a closer look at model (1) and comment on some of its features. As already seen in § 1, the model equation is

$$Y_{t,T} = g\left(\frac{t}{T}\right) + m(t) + \varepsilon_{t,T} \quad (t = 1, \ldots, T),$$

where $E(\varepsilon_{t,T}) = 0$, the function g is a deterministic trend and $\{m(t)\}_{t\in\mathbb{Z}}$ is a periodic sequence with unknown integer-valued period θ_0. In order to identify the function g and the sequence $\{m(t)\}_{t\in\mathbb{Z}}$, we normalize g to satisfy $\int_0^1 g(u)\,\mathrm{d}u = 0$. As shown in Lemma A2, this uniquely pins down g and $\{m(t)\}_{t\in\mathbb{Z}}$.

The trend function g in model (1) depends on rescaled time t/T rather than on real time t. This rescaling device is quite common in the literature. It is for example used in nonparametric regression and in the analysis of locally stationary processes; see Robinson (1989) and Dahlhaus

(1997), among others. The main reason for rescaling time to the unit interval is to obtain a framework for a reasonable asymptotic theory. If we defined g in terms of real time, we would not get additional information on the shape of g locally around a fixed time-point t as the sample size increases. Within the framework of rescaled time, in contrast, the function g is observed on a finer and finer grid of rescaled time-points on the unit interval as T grows. Thus, we obtain more and more information on the local structure of g around each point in rescaled time.

In contrast to g, we let the periodic component m depend on real time t. This allows us to exploit its periodic character when doing asymptotics: let s be a time-point in $\{1, \ldots, \theta_0\}$. As m is periodic, it has the same value at s, $s + \theta_0$, $s + 2\theta_0$, $s + 3\theta_0$, and so on. Hence, the number of time-points in our sample at which m has the value $m(s)$ increases as the sample size grows. This gives us more and more information about the value $m(s)$ and thus allows us to do asymptotics.

Rescaling the time argument of the trend component while letting the periodic part depend on real time is a rather natural way to formulate the model. It captures the fact that the trend function is much smoother and varies more slowly than the periodic component. An analogous model formulation has for example been used in Atak et al. (2011) who apply a model with a rescaled time trend and a seasonal component to a panel of temperature data from the UK. A discussion of different time series models that feature a nonparametric trend and a seasonal component can be found in Chapter 6 of Fan & Yao (2005).

Even though we do not impose any parametric restrictions on the sequence $\{m(t)\}_{t\in\mathbb{Z}}$, it can be represented by a vector of θ_0 parameters due to its periodic character. In particular, it is fully determined by the tuple of values $\beta = (\beta_1, \ldots, \beta_{\theta_0}) = \{m(1), \ldots, m(\theta_0)\}$. As a consequence, we can rewrite model (1) as

$$Y_{t,T} = g\left(\frac{t}{T}\right) + \sum_{s=1}^{\theta_0} \beta_s I_s(t) + \varepsilon_{t,T}, \tag{2}$$

where $I_s(t) = I(t = k\theta_0 + s \text{ for some } k)$ and $I(\cdot)$ is the indicator function. Model (1) can thus be regarded as a semiparametric regression model with indicator functions as regressors and the parameter vector β. In matrix notation, (2) becomes

$$Y = g + X_{\theta_0}\beta + \varepsilon,$$

where slightly abusing notation, $Y = (Y_{1,T}, \ldots, Y_{T,T})^{\mathrm{T}}$ is the vector of observations, $g = \{g(1/T), \ldots, g(T/T)\}^{\mathrm{T}}$ is the trend component, $X_{\theta_0} = (I_{\theta_0}, I_{\theta_0}, \ldots)^{\mathrm{T}}$ is the design matrix with I_{θ_0} being the $\theta_0 \times \theta_0$ identity matrix, and $\varepsilon = (\varepsilon_{1,T}, \ldots, \varepsilon_{T,T})^{\mathrm{T}}$ is the vector of residuals.

3. Estimation procedure

3·1. *Estimation of the period* θ_0

Roughly speaking, the period θ_0 is estimated as follows: to start with, we construct an estimator of the periodic sequence $\{m(t)\}_{t\in\mathbb{Z}}$ for each candidate period θ with $1 \leqslant \theta \leqslant \Theta_T$. Here, the upper bound Θ_T is allowed to grow with the sample size at a rate to be specified later on. Based on a penalized residual sum of squares criterion, we then compare the resulting estimators in terms of how well they fit the data. Finally, the true period θ_0 is estimated by the period corresponding to the estimator with the best fit.

More formally, for each candidate period θ, define the least squares estimate $\hat{\beta}_\theta$ as

$$\hat{\beta}_\theta = (X_\theta^{\mathrm{T}} X_\theta)^{-1} X_\theta^{\mathrm{T}} Y,$$

where $X_\theta = (I_\theta, I_\theta, \ldots)^{\mathrm{T}}$ is the design matrix with I_θ denoting the $\theta \times \theta$ identity matrix. In addition, let the residual sum of squares for the model with period θ be given by

$$\mathrm{RSS}(\theta) = \|Y - X_\theta \hat{\beta}_\theta\|^2,$$

where $\|x\| = (\sum_{t=1}^{\mathrm{T}} x_t^2)^{1/2}$ denotes the usual l_2-norm for vectors $x = (x_1, \ldots, x_T) \in \mathbb{R}^{\mathrm{T}}$.

At first glance, it may appear to be a good idea to take the minimizer of the residual sum of squares $\mathrm{RSS}(\theta)$ as an estimate of the period θ_0. However, this approach is too naive. In particular, it does not yield a consistent estimate of θ_0. The main reason is that each multiple of θ_0 is a period of the sequence m as well. Thus, model (2) may be represented by using a multiple of θ_0 parameters and a corresponding number of indicator functions. Intuitively, employing a larger number of regressors to explain the data yields a better fit, thus resulting in a smaller residual sum of squares than that obtained for the estimator based on the true period θ_0. This indicates that minimizing the residual sum of squares will usually overestimate the true period. In particular, it will notoriously tend to select multiples of θ_0 rather than θ_0 itself.

To overcome this problem, we add a regularization term to the residual sum of squares that penalizes choosing large periods. In particular, we base our estimation procedure on the penalized residual sum of squares

$$Q(\theta, \lambda_T) = \mathrm{RSS}(\theta) + \lambda_T \theta,$$

where the regularization parameter λ_T diverges to infinity at an appropriate rate to be specified later on. Our estimator $\hat{\theta}$ of the true period θ_0 is now defined as the minimizer

$$\hat{\theta} = \arg \min_{1 \leqslant \theta \leqslant \Theta_T} Q(\theta, \lambda_T).$$

In this definition, the upper bound Θ_T may tend to infinity as the sample size T increases. From a practical point of view, this means that we allow the period to be fairly large compared to the number of observations. In § 4·2, we discuss the exact rates at which Θ_T is allowed to diverge.

The regularization term $\lambda_T \theta$ can be regarded as an l_0-penalty: recalling the formulation (2) of our model, θ can be seen to equal the number of model parameters. In the literature, methods based on l_0-penalties have been employed to deal with model selection problems such as variable or lag selection in linear regression; see, e.g., Hannan & Quinn (1979), Nishii (1984) or Claeskens & Hjort (2008) for an overview. Indeed, the issue of estimating the period θ_0 can also be regarded as a model selection problem: for each candidate period θ, we have a model of the form (2) with a different set of regressors and model parameters. The aim is to pick the correct model among these. Similar to Sun et al. (2012), we thus look at estimating the period θ_0 from the perspective of model selection. Nevertheless, our selection method strongly differs from their crossvalidation approach.

Importantly, our l_0-penalized method is computationally not very costly, as we only have to calculate the criterion function $Q(\theta, \lambda_T)$ for Θ_T different choices of θ with Θ_T being of much smaller order than the sample size T. This contrasts with various problems in high-dimensional statistics, where an l_0-penalty turns out to be computationally too burdensome. To obtain computationally feasible methods, convex regularizations have been employed in this context instead. In particular, the l_1-regularization and the corresponding lasso approach have become very popular in recent years. See the original lasso article by Tibshirani (1996), and Bühlmann & van de Geer (2011) for a comprehensive overview.

When applying our penalized least squares procedure to estimate the period θ_0, we do not correct for the presence of a trend but ignore the trend function g. This is possible because g is smoothed out by our estimation procedure: for a given candidate period θ, the least squares

estimator of the periodic component at the time-point s can essentially be written as a sample average of the observations at the time-points $t \in \{1, \ldots, T\}$ with $t = s + (k-1)\theta$ for some $k \in \mathbb{Z}$. The trend g shows up in averages of the form $K^{-1} \sum_{k=1}^{K} g[\{s + (k-1)\theta\}/T]$ in this estimator, where K is the number of time points $t = s + (k-1)\theta$ in the sample. These averages approximate the integral $\int_0^1 g(u)\,\mathrm{d}u$, which is equal to zero by our normalization of g. Hence, they converge to zero and can effectively be neglected. In this sense, g gets smoothed or integrated out.

3·2. *Estimation of the periodic component m*

Given the estimate $\hat{\theta}$ of the true period θ_0, it is straightforward to come up with an estimator of the periodic sequence $\{m(t)\}_{t \in \mathbb{Z}}$. We simply define the estimator of the sequence values β as the least squares estimate $\hat{\beta}_{\hat{\theta}}$ that corresponds to the estimated period $\hat{\theta}$, i.e.,

$$\hat{\beta}_{\hat{\theta}} = (X_{\hat{\theta}}^{\mathrm{T}} X_{\hat{\theta}})^{-1} X_{\hat{\theta}}^{\mathrm{T}} Y.$$

The estimator $\hat{m}(t)$ of the sequence value $m(t)$ at time-point t is then defined by writing $\hat{\beta}_{\hat{\theta}} = \{\hat{m}(1), \ldots, \hat{m}(\hat{\theta})\}$ and letting $\hat{m}(s + k\hat{\theta}) = \hat{m}(s)$ for all $s = 1, \ldots, \hat{\theta}$ and all k. Hence, by construction, $\hat{m}$ is a periodic sequence with period $\hat{\theta}$. As in the previous estimation step, we completely ignore the trend function g when estimating the periodic sequence values. This is possible for exactly the same reasons as outlined in the previous subsection.

In many applications, the periodic component may be expected to have a fairly smooth shape. For this reason, it may be useful to work with a smoothed version of the estimator $\hat{m}$ in practice. In particular, we may define

$$\hat{m}_{\mathrm{smooth}}(s) = \frac{\sum_{t=1}^{\mathrm{T}} K_h(t-s)\hat{m}(t)}{\sum_{t=1}^{\mathrm{T}} K_h(t-s)},$$

where h is the bandwidth, K is a kernel function and $K_h(x) = K(x/h)/h$. This estimator provides a smoothed out picture of the periodic component that is easier to interpret in many cases, in particular when the estimated period is large. However, even though smoothing is a useful tool in practice, it does not make much difference from a theoretical point of view. To see this, let K be a kernel with bounded support. For $\hat{m}_{\mathrm{smooth}}(s)$ to be a consistent estimate of $m(s)$, the bandwidth must shrink to zero. For small values of the bandwidth, however, the kernel weight $K_h(t-s)$ contains only the point s itself. As the sample size grows large, we thus obtain that $\hat{m}_{\mathrm{smooth}}(s) = \hat{m}(s)$ at any time-point s.

3·3. *Estimation of the trend component g*

We finally tackle the problem of estimating the trend function g. Let us first consider an infeasible estimator of g. If the periodic component m were known, we could observe the variables $Z_{t,T} = Y_{t,T} - m(t)$. In this case, the trend component g could be estimated from the equation

$$Z_{t,T} = g\left(\frac{t}{T}\right) + \varepsilon_{t,T}$$

by standard procedures. One could for example use a local linear estimator defined by the minimization problem

$$\begin{bmatrix} \tilde{g}(u) \\ \partial\tilde{g}(u)/\partial u \end{bmatrix} = \underset{(g_0,g_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{t=1}^{\mathrm{T}} \left\{ Z_{t,T} - g_0 - g_1\left(\frac{t}{T} - u\right) \right\}^2 K_h\left(u - \frac{t}{T}\right), \qquad (3)$$

where $\tilde{g}(u)$ is the estimate of g at time-point u and $\partial\tilde{g}(u)/\partial u$ is the estimate of the first derivative of g at u. As in the previous subsection, h denotes the bandwidth and K is a kernel function with $\int K(v)\,\mathrm{d}v = 1$ and $K_h(x) = K(x/h)/h$.

Even though we do not observe the variables $Z_{t,T}$, we can approximate them by $\hat{Z}_{t,T} = Y_{t,T} - \hat{m}(t)$. This allows us to come up with a feasible estimator of the trend function g: simply replacing the variables $Z_{t,T}$ in (3) by the approximations $\hat{Z}_{t,T}$ yields an estimator $\hat{g}$ that can be computed from the data. Standard calculations show that $\hat{g}(u)$ has the closed form solution

$$\hat{g}(u) = \frac{\sum_{t=1}^{\mathrm{T}} w_{t,T}(u)\hat{Z}_{t,T}}{\sum_{t=1}^{\mathrm{T}} w_{t,T}(u)}$$

with $w_{t,T}(u) = K_h(u - t/T)\{S_{T,2}(u) - (t/T - u)S_{T,1}(u)\}$ and $S_{T,j}(u) = \sum_{t=1}^{\mathrm{T}} K_h(u - t/T)(t/T - u)^j$ for $j = 1, 2$. Alternatively to the above local linear estimator, we could have used a simple Nadaraya–Watson smoother. However, as Nadaraya–Watson smoothing notoriously suffers from boundary problems, we work with a local linear estimator throughout.

4. Asymptotics

4·1. *Assumptions*

To derive the asymptotic properties of our estimators, we impose the following conditions.

Condition 1. The error process $\{\varepsilon_{t,T}\}$ is strongly mixing with mixing coefficients $\alpha(k)$ satisfying $\alpha(k) \leqslant Ca^k$ for some positive constants C and $a < 1$.

Condition 2. It holds that $E(|\varepsilon_{t,T}|^{4+\delta}) \leqslant C$ for some small $\delta > 0$ and a positive constant $C < \infty$.

Condition 3. The function g is twice continuously differentiable on $[0, 1]$.

Condition 4. The kernel K is bounded, symmetric about zero and has compact support. Moreover, it is Lipschitz, i.e., there exists a positive constant L with $|K(u) - K(v)| \leqslant L|u - v|$.

We briefly give some remarks on the above conditions. Most importantly, we do not assume the error process $\{\varepsilon_{t,T}\}$ to be stationary. We merely put some restrictions on its dependence structure. In particular, we assume the array $\{\varepsilon_{t,T}\}$ to be strongly mixing. We do not necessarily require exponentially decaying mixing rates, as assumed in Condition 1. These could be replaced by slower polynomial rates, at the cost of having somewhat stronger restrictions on the penalty parameter λ_T later on. To keep the notation and structure of the proofs as clear as possible, however, we stick to exponential mixing rates throughout. Condition 3 is needed only for the third estimation step, i.e., for establishing the asymptotic properties of the trend g. If we restrict attention to the first two steps of our procedure, i.e., to estimating the periodic model component, it suffices to assume that g is of bounded variation.

4·2. *Asymptotics for the period estimator* $\hat{\theta}$

The next theorem characterizes the asymptotic behaviour of the estimator $\hat{\theta}$. To formulate the result in a neat way, we introduce the following notation: for any two sequences $\{v_T\}$ and $\{w_T\}$ of positive numbers, we write $v_T \ll w_T$ to mean that $v_T = o(w_T)$.

THEOREM 1. *Let Conditions* 1–3 *be fulfilled and assume that* $\Theta_T \leqslant CT^{2/5-\delta}$ *for some small* $\delta > 0$ *and a finite constant* C. *Moreover, choose the regularization parameter* λ_T *to satisfy* $(\log T)\Theta_T^{3/2} \ll \lambda_T \ll T$. *Then* $\hat{\theta} = \theta_0 + o_p(1)$, *as* $T \to \infty$.

The theorem shows that we get consistency under rather general conditions on the upper bound Θ_T. In particular, Θ_T is allowed to grow at a rate of almost $T^{2/5}$. Clearly, the faster Θ_T goes to infinity, the stronger restrictions must be imposed on the regularization parameter λ_T. If Θ_T is a fixed number, then it suffices to choose λ_T of slightly larger order than $\log T$. This contrasts with an order of almost $T^{3/5}$ if Θ_T diverges at the highest possible rate.

4·3. *Asymptotics for the estimator* $\hat{m}$

We now provide the convergence rate and the limiting distribution of the estimator $\hat{m}$ of the periodic model component. To simplify notation, define

$$V_{t_0,T} = \frac{\theta_0^2}{T} \sum_{k,k'=1}^{K_{t_0,T}} \operatorname{cov}(\varepsilon_{t_0+(k-1)\theta_0,T}, \varepsilon_{t_0+(k'-1)\theta_0,T})$$

for each time-point t, with $t_0 = t - \theta_0 \lfloor (t-1)/\theta_0 \rfloor$ and $K_{t_0,T} = 1 + \lfloor (T - t_0)/\theta_0 \rfloor$.

THEOREM 2. *Let the conditions of Theorem* 1 *be satisfied. Then*

$$\max_{1 \leqslant t \leqslant T} \left| \hat{m}(t) - m(t) \right| = O_p\left(T^{-1/2}\right), \quad T \to \infty.$$

In addition, assume that the limit $V_{t_0} = \lim_{T\to\infty} V_{t_0,T}$ *exists. Then for each time-point* $t = 1, \ldots, T$,

$$T^{1/2} \left\{ \hat{m}(t) - m(t) \right\} \to N(0, V_{t_0}), \quad T \to \infty,$$

in distribution.

The limit V_{t_0} exists under quite general assumptions on the error process $\{\varepsilon_{t,T}\}$. If the latter is stationary, then V_{t_0} simplifies to $V_{t_0} = \theta_0 \sum_{k=-\infty}^{\infty} \operatorname{cov}(\varepsilon_{0,T}, \varepsilon_{k\theta_0,T})$ and can be estimated by classical methods discussed in Hannan (1957). Estimating the long-run variance V_{t_0} in a more general setting, which allows for nonstationarities in the errors, is studied in de Jong & Davidson (2000) among others. Inspecting the proof of Theorem 2, one can see that the estimator $\hat{m}$ has the same limiting distribution as the estimator in the oracle case where the true period θ_0 is known. In particular, it has the same asymptotic variance V_{t_0}. Hence, the error of estimating the period θ_0 does not become visible in the limiting distribution of $\hat{m}$.

4·4. *Asymptotics for the estimator* $\hat{g}$

We finally derive the asymptotic properties of the local linear smoother $\hat{g}$. To do so, define

$$V_{u,T} = \frac{h}{T} \sum_{s,t=1}^{\mathrm{T}} K_h\left(u - \frac{s}{T}\right) K_h\left(u - \frac{t}{T}\right) E[\varepsilon_{s,T}\varepsilon_{t,T}].$$

The next theorem specifies the uniform convergence rate and the asymptotic distribution of the smoother $\hat{g}$.

THEOREM 3. *Suppose that the conditions of Theorem* 1 *are satisfied and that the kernel* K *fulfils Condition* 4.

(i) *If the bandwidth h shrinks to zero and fulfils $T^{1/2-\delta}h \to \infty$ for some small $\delta > 0$, then it holds that*

$$\sup_{u\in[0,1]} \left|\hat{g}(u) - g(u)\right| = O_p\left\{\left(\frac{\log T}{Th}\right)^{1/2} + h^2\right\}, \ T \to \infty.$$

(ii) *Consider a fixed point $u \in (0, 1)$ and assume that the limit $V_u = \lim_{T\to\infty} V_{u,T}$ exists. Moreover, let $Th^5 \to c_h$ for some constant $c_h \geqslant 0$. Then it holds that*

$$(Th)^{1/2}\left\{\hat{g}(u) - g(u) - h^2 B_u\right\} \to N(0, V_u), \ T \to \infty,$$

in distribution with $B_u = (1/2)\{\int v^2 K(v)\,\mathrm{d}v\}g''(u)$.

Similarly to Theorem 2, the limit V_u exists under rather general conditions on the process $\{\varepsilon_{t,T}\}$. If the process is stationary, then the asymptotic variance V_u simplifies to $V_u = \{\int K^2(v)\,\mathrm{d}v\}\sum_{l=-\infty}^{\infty}\mathrm{cov}(\varepsilon_{0,T}, \varepsilon_{l,T})$. For methods to estimate V_u, we again refer to Hannan (1957) and de Jong & Davidson (2000).

Inspecting the proof of Theorem 3, one can see that the smoother $\hat{g}$ asymptotically behaves in the same way as the oracle estimator $\tilde{g}$, which is constructed under the assumption that the periodic component m is known. In particular, replacing $\hat{g}$ by $\tilde{g}$ results in an error of the order $O_p(T^{-1/2})$ uniformly over u and h. As a consequence, $\hat{g}$ has the same limiting distribution as $\tilde{g}$. Thus, the need to estimate the periodic sequence m is not reflected in the limit law of $\hat{g}$.

As the difference between $\hat{g}$ and the standard smoother $\tilde{g}$ is of the asymptotically negligible order $O_p(T^{-1/2})$, the bandwidth of $\hat{g}$ can be selected by the same techniques as used for the smoother $\tilde{g}$. In particular, standard methods like crossvalidation or plug-in rules can be employed. However, these techniques may perform very poorly when the errors are correlated. To achieve reasonable results, they have to be adjusted as shown for example in Hart (1991).

5. Selecting the regularization parameter λ_T

As shown in Theorem 1, our procedure to estimate the period θ_0 is asymptotically valid for all sequences of regularization parameters λ_T within a fairly wide range of rates. Hence, from an asymptotic perspective, we have a lot of freedom to choose the regularization parameter. In finite samples, a totally different picture arises. There, different choices of λ_T may result in completely different estimates of the period θ_0. Selecting the regularization parameter λ_T in an appropriate way is thus a crucial issue in small samples.

In what follows, we provide a heuristic discussion on how to choose λ_T in a suitable way. The argument is similar to that for deriving the classical final prediction error criterion of Akaike (1969). To make the argument as clear as possible, we consider a simplified version of model (1). In particular, we analyse the setting $Y_t = m(t) + \varepsilon_t$, where the errors ε_t are assumed to be independent and identically distributed with $E(\varepsilon_t^2) = \sigma^2$. We thus drop the trend component from the model and assume that there is no serial dependence in the error terms.

As can be seen from the proof of Theorem 1, the main role of the penalty term $\lambda_T\theta$ is to avoid selecting multiples of the true period θ_0. For this reason, we focus attention on periods θ that are multiples of θ_0, i.e., $\theta = r\theta_0$ for some r. It is not difficult to show that

$$E\{\mathrm{RSS}(r\theta_0)\} + \sigma^2 r\theta_0 = E\{\mathrm{RSS}(\theta_0)\} + \sigma^2\theta_0. \tag{4}$$

For completeness, the proof is provided in the Supplementary Material. Formula (4) suggests selecting the penalty parameter λ_T larger than σ^2 in order to avoid choosing multiples of the true

period θ_0 rather than θ_0 itself. On the other hand, λ_T should not be picked too large. Otherwise we add a strong penalty to the residual sum of squares $\mathrm{RSS}(\theta_0)$ of the true period θ_0, thus making the criterion function at θ_0 rather large, in particular larger than the criterion function at 1. As a result, our procedure would yield the estimate $\hat{\theta} = 1$, i.e., it would suggest a model without a periodic component.

The above heuristics suggest to select the penalty parameter λ_T slightly larger than σ^2. In particular, we propose to choose it as

$$\lambda_T = \sigma^2 \kappa_T \tag{5}$$

with some sequence $\{\kappa_T\}$ that slowly diverges to infinity. More specifically, $\{\kappa_T\}$ should grow slightly faster than $\{\log T\}$ to meet the conditions of the asymptotic theory from Theorem 1.

As the error variance σ^2 is unknown in general, we cannot take the formula (5) at face value but must replace σ^2 with an estimator. This can be achieved as follows: to start with, define $\check{\theta} = \min_{1 \leqslant \theta \leqslant \Theta_T} \mathrm{RSS}(\theta)$. As already noted in § 3·1, minimizing the residual sum of squares without a penalty does not yield a consistent estimate of θ_0. Inspecting the proof of Theorem 1, it can however be seen that $\mathrm{pr}(\check{\theta} = k\theta_0 \text{ for some } k \in \mathbb{N}) \to 1$ as $T \to \infty$. Hence, with probability approaching one, $\check{\theta}$ is equal to a multiple of the period θ_0. Since multiples of θ_0 are periods of m, the least squares estimator $\hat{\beta}_{\check{\theta}}$ can be used as a preliminary estimator of the periodic sequence values. Let us denote the resulting estimator of $m(t)$ at time-point t by $\check{m}(t)$. Given this estimator, we can repeat the third step of our procedure to obtain an estimator $\check{g}$ of the trend function g. Finally, subtracting the estimators $\check{m}(t)$ and $\check{g}(t/T)$ from the observations Y_t yields approximations $\check{\varepsilon}_t$ of the residuals ε_t. These can be used to construct the standard-type estimator $\check{\sigma}^2 = T^{-1} \sum_{t=1}^{\mathrm{T}} \check{\varepsilon}_t^2$ of the error variance σ^2.

So far, we have restricted attention to the case of independent error terms. Repeating our heuristic argument with serially correlated errors, the variance σ^2 in (4) is replaced by some type of long-run variance that incorporates covariance terms of the errors. Our selection rule for the penalty parameter λ_T does not take into account this effect of the dependence structure. Nevertheless, this does not mean that our rule becomes useless when the error terms are correlated. As long as the correlation is not too strong, σ^2 will dominate the long-run variance. Hence, our heuristic rule should still yield an appropriate penalty parameter λ_T. This consideration is confirmed by our simulations in the next section, where the error terms are assumed to follow an AR(1) process.

6. SIMULATION

In this section, we examine the finite sample behaviour of our procedure in a Monte Carlo experiment. To do so, we simulate the model (1) with a periodic sequence of the form

$$m(t) = \sin\left(\frac{2\pi}{\theta_0} t + \frac{3\pi}{2}\right)$$

and a period $\theta_0 = 60$. The trend function g is given by $g(u) = 2u^2$. The error terms ε_t of the simulated model are drawn from the AR(1) process $\varepsilon_t = 0{\cdot}45\varepsilon_{t-1} + \eta_t$, where η_t are independent and identically distributed variables following a normal distribution with mean zero and variance σ_η^2. We will choose different values for σ_η^2 later on, thus altering the signal-to-noise ratio in the model. The simulation set-up is chosen to mimic the situation in the real-data example investigated in § 7.

We simulate the model $N = 1000$ times for three different sample sizes $T = 160$, 250, 500 and three different values of the residual variance $\sigma_\eta^2 = 0{\cdot}2$, $0{\cdot}4$, $0{\cdot}6$. The sample size $T = 160$ corresponds to the situation in the application later on where we have 162 data points. The

Table 1. *Empirical probabilities that* $\hat{\theta} = 60$*, first three columns, and that* $55 \leqslant \hat{\theta} \leqslant 65$*, last three columns, for different choices of* T *and* σ^2

	$\text{pr}(\hat{\theta} = 60)$			$\text{pr}(55 \leqslant \hat{\theta} \leqslant 65)$		
	$T = 160$	$T = 250$	$T = 500$	$T = 160$	$T = 250$	$T = 500$
$\sigma^2 = 0{\cdot}25$	0·102	0·247	0·587	0·994	0·932	1·000
$\sigma^2 = 0{\cdot}5$	0·089	0·178	0·539	0·942	0·979	1·000
$\sigma^2 = 0{\cdot}75$	0·087	0·143	0·472	0·854	0·950	1·000

values of σ_η^2 translate into error variances $\sigma^2 = E(\varepsilon_t^2)$ of approximately 0·25, 0·5, and 0·75, respectively. To get a rough idea of the noise level in our set-up, we consider the ratio $\overline{\varepsilon^2}/\overline{Y^2} = (\sum_{t=1}^{\mathrm{T}} \varepsilon_t^2)/(\sum_{t=1}^{\mathrm{T}} Y_t^2)$, which gives the fraction of variation in the data that is due to the variation in the error terms. More exactly, we report the values of the ratio $\overline{\hat{\varepsilon}^2}/\overline{Y^2}$ with $\hat{\varepsilon}_t$ being the estimated residuals. This makes it easier to compare the noise level in the simulations to that in the real-data example later on. For $\sigma^2 = 0{\cdot}25, 0{\cdot}5, 0{\cdot}75$, we obtain $\overline{\hat{\varepsilon}^2}/\overline{Y^2} \approx 0{\cdot}12, 0{\cdot}2, 0{\cdot}26$. These numbers are a bit higher than the value 0·07 obtained in the real-data example, indicating that the noise level is somewhat higher in the simulations.

The regularization parameter is chosen as $\lambda_T = \check{\sigma}^2 \kappa_T$ with $\kappa_T = \log T$. Here, $\check{\sigma}^2$ is the estimator of the error variance σ^2 introduced in § 5. We thus pick λ_T according to the heuristic idea described there. From a theoretical perspective, we should have chosen κ_T to diverge slightly faster than $\log T$. However, as the rate of κ_T may become arbitrarily close to $\log T$, we neglect this technicality and simply choose κ_T to equal $\log T$.

In our simulation exercise, we focus on the estimation of the period θ_0. This is the crucial step in our estimation scheme as the finite sample behaviour of the estimators $\hat{m}$ and $\hat{g}$ hinges on how well $\hat{\theta}$ approximates the true period θ_0. If the period θ_0 is known, $\hat{m}$ simplifies to a standard least squares estimator. Moreover, if the periodic model component m as a whole is observed, then $\hat{g}$ turns into an ordinary local linear smoother. The finite sample properties of these estimators have been extensively studied and are well known. Given a good approximation of θ_0, our estimators $\hat{m}$ and $\hat{g}$ can be expected to perform similarly to these standard estimators. For this reason, we concentrate on the properties of $\hat{\theta}$ in what follows.

The simulation results are presented in Table 1. For each choice of the sample size T and the error variance σ^2, we have performed 1000 simulations, where periods θ with $1 \leqslant \theta \leqslant T/2$ have been taken into account for the estimation. The first three columns of the table give the probabilities with which the estimator $\hat{\theta}$ hits the true value $\theta_0 = 60$; the last three columns are the probabilities of $\hat{\theta}$ taking values between 55 and 65. Overall, Table 1 suggests that the estimator $\hat{\theta}$ performs well in small samples. Even at a sample size of $T = 160$, where we observe only a bit less than three full cycles of the periodic component, the estimates strongly cluster around the true period θ_0. Clearly, at this small sample size, the estimator $\hat{\theta}$ does not exactly hit the true period in many cases. Nevertheless, it gives a reasonable approximation to it most of the time. The performance of the estimator quickly improves as we observe more and more cycles of the periodic component. Moving to a sample size of $T = 500$, it already hits the true value θ_0 in around 50–60% of the simulations and always takes values between 55 and 65.

A graphical presentation of the simulation results for the sample size $T = 250$ is given in Fig. 1. Each panel shows the distribution of $\hat{\theta}$ for a specific choice of σ^2. The figure makes visible some additional features of the simulation results: (a) in addition to the main cluster of estimates around the true period θ_0, smaller clusters can be found around multiples of θ_0. As can be seen from the proof of Theorem 1, this behaviour of $\hat{\theta}$ is suggested by the asymptotic

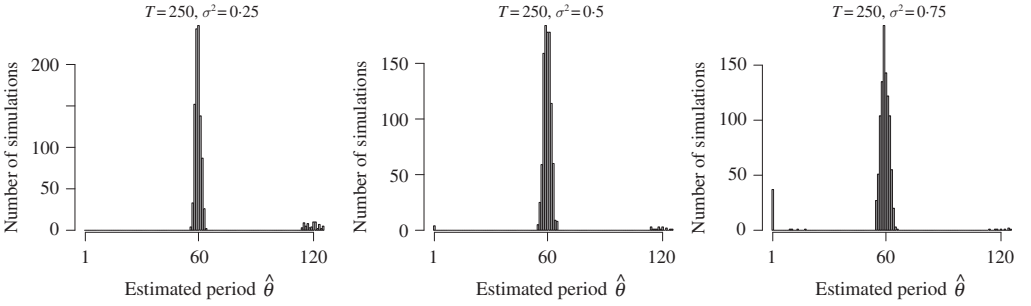

Fig. 1. Histograms of the simulation results for $T = 250$ and different choices of σ^2. The bars give the number of simulations, out of a total of 1000, in which a certain value $\hat{\theta}$ is obtained.

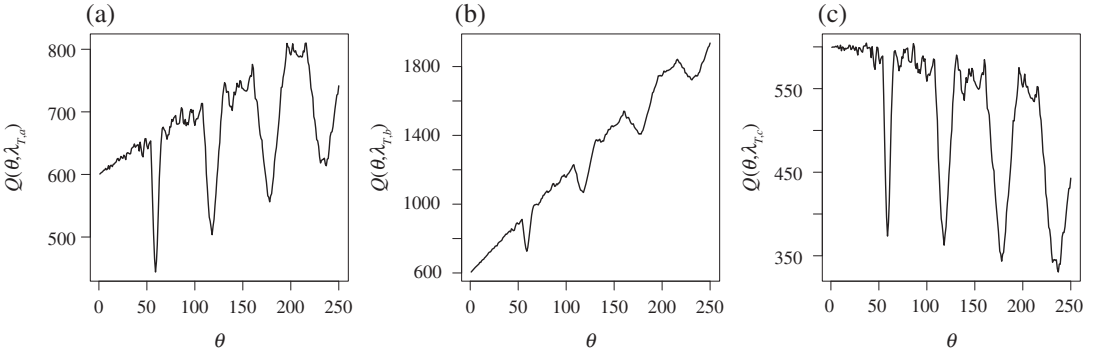

Fig. 2. Plot of the criterion function for a typical simulation with $T = 500$, $\sigma^2 = 0{\cdot}5$ and three choices of λ_T. In particular, λ_T is given by $\lambda_{T,a} = \check{\sigma}^2 \log T$, $\lambda_{T,b} = 4\lambda_{T,a}$ and $\lambda_{T,c} = \lambda_{T,a}/4$ in the three panels.

theory. (b) For $\sigma^2 = 0{\cdot}75$, $\hat{\theta}$ is equal to 1 in a nonnegligible number of simulations. This is a finite sample effect, which vanishes as the sample size increases. As will become clear below, this has to do with the choice of the penalty parameter λ_T. In particular, we could considerably lower the number of simulations with $\hat{\theta} = 1$ by decreasing the penalty λ_T slightly.

In what follows, we have a closer look at what happens when the regularization parameter λ_T is varied. Since the effect of varying λ_T is better visible for larger sample sizes, we consider a situation with $T = 500$. Figure 2 presents the criterion function $Q(\theta, \lambda_T)$ for a typical simulation with $T = 500$, $\sigma^2 = 0{\cdot}5$ and three different choices of λ_T. In panel (a), we have chosen the regularization parameter as before, i.e., $\lambda_{T,a} = \check{\sigma}^2 \log T$. In panel (b), we pick it a bit larger, $\lambda_{T,b} = 4\lambda_{T,a}$, and in panel (c), we choose it somewhat smaller, $\lambda_{T,c} = \lambda_{T,a}/4$.

As can be seen from the plots, the main features of the criterion function are the downward spikes around the true period θ_0 and multiples thereof. The parameter λ_T influences the overall upward or downward movement of the criterion function, because the penalty $\lambda_T\theta$ is linear in θ with slope parameter λ_T. If λ_T is too large, then the criterion function increases too quickly, and the global minimum does not lie at the first downward spike around θ_0 but at $\theta = 1$; see panel (b). If λ_T is chosen too small, then the criterion function decreases, taking its global minimum not at the first downward spike but at a subsequent one; see panel (c).

Our heuristic rule for selecting λ_T can be regarded as a guideline to choose the right order of magnitude for the penalty term. Nevertheless, we may still pick λ_T a bit too large or small, thus ending up in a similar situation as in panels (b) or (c). When applying our method to real data, it is thus important to examine the criterion function. If it exhibits large downward spikes at a certain

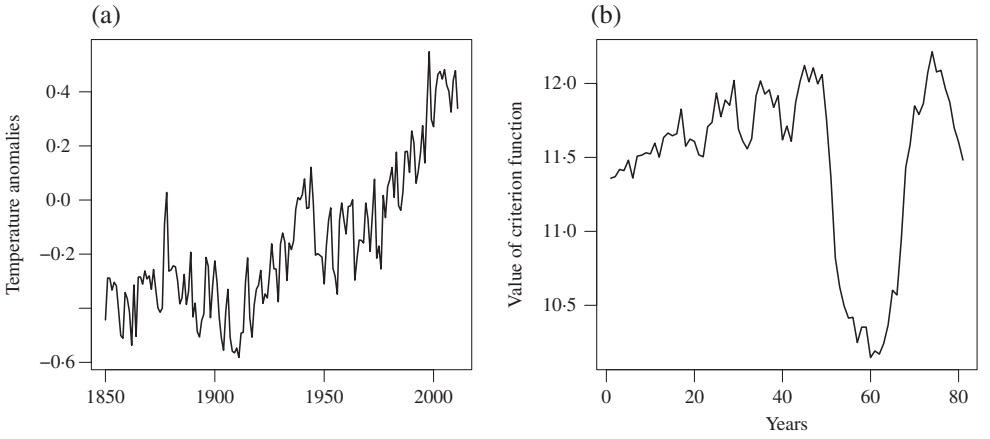

Fig. 3. (a) Yearly global temperature anomalies from 1850 to 2011, measured in °C; (b) a plot of the criterion function $Q(\theta, \lambda_T)$.

value and at multiples thereof, this is strong evidence for there being a periodic component in the data. In particular, the true period should lie in the region of the first downward spike. If our procedure yields a completely different estimate of the period, one should treat this result with caution; it may be due to an inappropriate choice of the penalty parameter.

7. Application

Global mean temperature records over the last 150 years suggest that there has been a significant upward trend in the temperatures; see Bloomfield (1992) or Hansen et al. (2002), among others. This global warming trend is also visible in the time series presented in Fig. 3(a). The depicted data are yearly global temperature anomalies from 1850 to 2011. By anomalies we mean the departure of the temperature from some reference value or a long-term average. In particular, the data at hand are temperature deviations from the average 1961–1990 measured in degrees Celsius. The dataset is called HadCRUT3 and can be obtained from the Climatic Research Unit of the University of East Anglia, England. A detailed description of the data is given by Brohan et al. (2006).

The issue of global warming has received considerable attention over the last decades. From a statistical point of view, the challenge is to come up with methods to reliably estimate the warming trend. Providing such methods is complicated by the fact that the global mean temperatures may contain not only a trend but also a long-run oscillatory component. Various research articles in climatology suggest that the global temperature system possesses an oscillation with a period in the region between 60 and 70 years; see Schlesinger & Ramankutty (1994), Delworth & Mann (2000) and Mazzarella (2007), among others. The presence of such a periodic component obviously creates problems when estimating the trend function. In particular, an estimation procedure is required that is able to accurately separate the periodic and the trend components. Otherwise, an inaccurate picture of the global warming trend emerges. Moreover, a precise estimate of both components is required to reliably predict future temperature changes.

In what follows, we apply our three-step procedure to the temperature anomalies from Fig. 3. We thus fit the model (1) to the sample of global anomaly data $\{Y_{t,T}\}$ and estimate the unknown period θ_0, the values of the periodic sequence $\{m(t)\}_{t\in\mathbb{Z}}$, and the nonparametric trend function g.

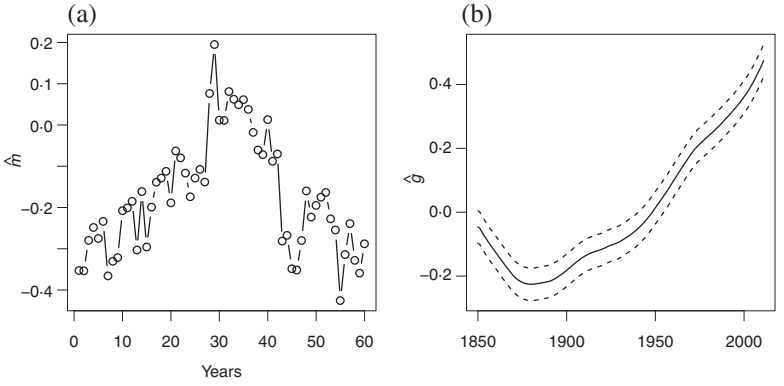

Fig. 4. (a) The estimator $\hat{m}$ of the periodic component. The solid line in (b) depicts the estimator $\hat{g}$ of the trend function, the dashed lines are pointwise 95% confidence bands.

To estimate the period θ_0, we employ our penalized least squares method with the penalty parameter $\lambda_T = \check{\sigma}^2 \log T$. As in the simulations, $\check{\sigma}^2$ is an estimate of the error variance that is constructed as described in § 5. Selecting the penalty parameter in this way, the criterion function $Q(\theta, \lambda_T)$ is minimized at $\hat{\theta} = 60$. We thus detect an oscillation in the temperature data with a period in the same region as in the climatological studies cited above. The criterion function $Q(\theta, \lambda_T)$ is plotted in Fig. 3(b). Its most dominant feature is the enormous downward spike with a minimum at 60 years. As discussed in the simulations, this kind of spike is characteristic for the presence of a periodic component in the data. The spike being very pronounced, the shape of the criterion function provides strong evidence for there being an oscillation in the region of 60 years.

We next turn to the estimation of the periodic component m and the trend function g. The estimator $\hat{m}$ is presented in Fig. 4(a) over a full cycle of 60 years. The solid curve in 4(b) shows the local linear smoother $\hat{g}$, the dashed lines are the corresponding 95% pointwise confidence bands. For the estimation, we have used an Epanechnikov kernel and have chosen the bandwidth to equal $h = 0{\cdot}15$. To check the robustness of our results, we have additionally repeated the analysis for various choices of the bandwidth. As the results are fairly stable, we report the findings only for the bandwidth $h = 0{\cdot}15$.

Figure 5 depicts the time series of the estimated residuals $\hat{\varepsilon}_{t,T} = Y_{t,T} - \hat{g}(t/T) - \hat{m}(t)$ together with its sample autocorrelation function. The residuals do not exhibit a strong periodic or trending behaviour. This suggests that our procedure has done a good job in extracting the trend and periodic component from the data. Moreover, inspecting the sample autocorrelations, the residuals do not appear to be strongly dependent over time. The sample autocorrelation at the first lag has the value 0·45 and equals the parameter estimate obtained from fitting an AR(1) process to the residuals. This value was used as a guideline in the design of the error terms in the simulations.

8. COMPARISON WITH UNOBSERVED COMPONENTS MODELS

In this section, we analyse the temperature anomaly data by means of an unobserved components model and compare the empirical findings with those obtained by our method. We fit the following version of the unobserved components model to the data:

$$Y_t = \mu_t + \psi_t + \varepsilon_t,$$

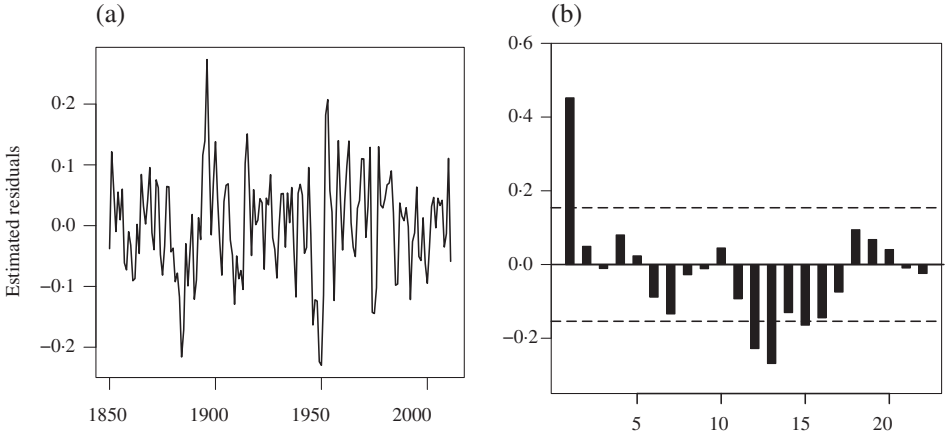

Fig. 5. Time series of the estimated residuals (a) and its sample autocorrelation function (b). The dashed lines show the Bartlett bounds $\pm 1{\cdot}96T^{-1/2}$.

where Y_t are the observed data points, μ_t is the trend, ψ_t denotes the cyclical component and ε_t is the error term. The trend component is modelled by the equations

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t$$
$$\beta_t = \beta_{t-1} + \zeta_t,$$

where η_t and ζ_t are independent white noise disturbances with variances σ_η^2 and σ_ζ^2, respectively. The trend is thus a random walk with a drift β_t. The dynamics of the cyclical component is described by

$$\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix} = \rho \begin{bmatrix} \cos\lambda & \sin\lambda \\ -\sin\lambda & \cos\lambda \end{bmatrix} \begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix},$$

where κ_t and κ_t^* are independent white noise disturbances with a common variance σ_κ^2 and ψ_t^* is an auxiliary variable appearing by construction. The parameter ρ is a dampening factor satisfying $0 \leqslant \rho \leqslant 1$ and λ is the frequency of the cycle. Most importantly, the parameter $\vartheta = 2\pi/\lambda$ is the counterpart of the period θ_0 in our model. The amount of smoothness in the trend and the cyclical component depends on the values of the variances σ_η^2, σ_ζ^2 and σ_κ^2, which are commonly called the hyperparameters of the model. To obtain a rather smooth estimate of the trend component, we work with an integrated random walk trend, i.e., we set the variance σ_η^2 of the disturbances η_t in the stochastic trend specification to zero. The hyperparameters and the model components can be estimated by maximum likelihood methods once the model has been brought into state space form; see Harvey (1989) or Durbin & Koopman (2001) for details.

Figure 6 compares the estimation results of our method with those obtained from the unobserved components approach. The latter have been produced by the STAMP software of Koopman et al. (1999) The solid line is the sum of the estimated trend and periodic component in our model, the dashed line is the corresponding estimate in the unobserved components framework. In order to make it easier to compare the curves visually, we have applied a small amount of smoothing to our estimator of the periodic component as described at the end of § 3·2. Specifically, we have used an Epanechnikov kernel along with a bandwidth of 5·5. As can be seen, the fits are fairly similar and indeed become almost indiscernible if we make our estimator of the periodic component even smoother.

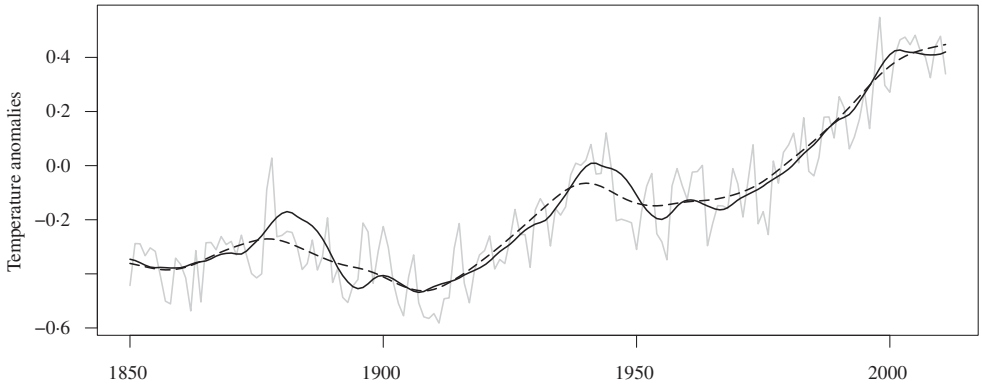

Fig. 6. Data fits produced by our method and the unobserved components approach. The solid line is the sum of the estimated trend and periodic component in our model, the dashed line is the corresponding estimate in the unobserved components model. The grey line in the background shows the time series of temperature anomalies.

Fig. 7. Time series of the estimated slope parameters β_t in the unobserved components model.

Even though the two methods provide similar fits to the data, they produce two very different decompositions into a trend and a periodic component. Using our method, we find a period of 60 years and a great deal of the fluctuations in the data are assigned to the periodic component. As noted in the previous section, this is in accordance with a variety of results from the climatology literature. When fitting an unobserved components model to the data in contrast, everything gets absorbed in the trend component and the cycle is effectively dropped from the model. Thus, the estimated model contains only a trend, which is depicted by the dashed line in Fig. 6.

Although the fitted unobserved components model does not involve a cycle, the periodic character of the data remains visible on a more hidden level in the trend component. Figure 7 plots the time series of the estimated slope parameters β_t. As can be seen, they exhibit a fairly cyclical behaviour with roughly three cycles over the observed time span. Thus, an oscillation of 60–70 years appears to be present in the slope coefficient. A similar observation has been made by Harvey (2006, § 2.7) in the context of US GDP data. There, some long-term upward and downward swings of the economy are captured by the trend component rather than the cycle and are also visible in the cyclical shape of the slope coefficients.

ACKNOWLEDGEMENT

We thank Andrew Harvey for his advice on the empirical comparison of our method with unobserved components models, and an associate editor and two referees for their constructive comments and suggestions. Financial support by the European Research Council is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs and technical details that are omitted in the paper. In addition, we discuss some extensions of our approach and continue the empirical comparison with unobserved components models from § 8.

APPENDIX

In this appendix, we prove the main theoretical results of the paper. We use the following notation: the symbol C denotes a generic constant that may take a different value on each occurrence. We implicitly suppose that C does not depend on any model parameters, in particular it is independent of the candidate period θ and the sample size T. Moreover, we let

$$\Pi_\theta = X_\theta (X_\theta^{\mathrm{T}} X_\theta)^{-1} X_\theta^{\mathrm{T}}$$

be the projection matrix onto the subspace $\{X_\theta b : b \in \mathbb{R}^\theta\}$. As the design matrix X_θ is orthogonal, Π_θ has a rather simple structure. In particular,

$$\Pi_\theta = X_\theta D_\theta X_\theta^{\mathrm{T}} = \begin{pmatrix} D_\theta & D_\theta & \cdots \\ D_\theta & \ddots & \\ \vdots & & \end{pmatrix}, \quad D_\theta = \begin{pmatrix} 1/K_{1,T}^{[\theta]} & & 0 \\ & \ddots & \\ 0 & & 1/K_{\theta,T}^{[\theta]} \end{pmatrix},$$

where $K_{s,T}^{[\theta]} = 1 + \lfloor (T-s)/\theta \rfloor$ for $s = 1, \ldots, \theta$. $K_{s,T}^{[\theta]}$ is the number of time-points t in the sample that satisfy $t = s + (k-1)\theta$ for some $k \in \mathbb{N}$. It equals either $\lfloor T/\theta \rfloor$ or $\lfloor T/\theta \rfloor + 1$; in particular it holds that $K_{s,T}^{[\theta]} = O(T/\theta)$. Finally, rewriting the residual sum of squares in terms of Π_θ yields $\mathrm{RSS}(\theta) = Y^{\mathrm{T}}(I - \Pi_\theta)Y$.

We first state an auxiliary lemma that is repeatedly used in the proofs later on.

LEMMA A1. *Let θ be any natural number with $1 \leqslant \theta \leqslant \Theta_T$ and let $s \in \{1, \ldots, \theta\}$. Then*

$$\left| \frac{1}{K_{s,T}^{[\theta]}} \sum_{k=1}^{K_{s,T}^{[\theta]}} g\left\{ \frac{s + (k-1)\theta}{T} \right\} - \int_0^1 g(u)\,\mathrm{d}u \right| \leqslant \frac{C}{K_{s,T}^{[\theta]}}$$

for some constant C that is independent of s, θ, and T.

The proof is straightforward and thus omitted. We next provide a result on the identification of the model components g and m.

LEMMA A2. *The sequence m and the function g in model (1) are uniquely identified if g is normalized to satisfy $\int_0^1 g(u)\,\mathrm{d}u = 0$. More precisely, let $\bar{g}$ be a smooth trend function with $\int_0^1 \bar{g}(u)\,\mathrm{d}u = 0$ and $\bar{m}$ a periodic sequence with, smallest, period $\bar{\theta}_0$. If*

$$\bar{g}\left(\frac{t}{T}\right) + \bar{m}(t) = g\left(\frac{t}{T}\right) + m(t)$$

for all $t = 1, \ldots, T$ and all $T = 1, 2, \ldots$, then $\bar{g} = g$ and $\bar{m} = m$ with $\bar{\theta}_0 = \theta_0$.

The proof can be found in the Supplementary Material. We now turn to the proofs of the main theorems.

Proof of Theorem 1. Our arguments are based on the inequality

$$\mathrm{pr}(\hat{\theta} \neq \theta_0) \leqslant \sum_{\substack{1 \leqslant \theta \leqslant \Theta_T \\ \theta \neq \theta_0}} \mathrm{pr}\{Q(\theta, \lambda_T) \leqslant Q(\theta_0, \lambda_T)\}. \tag{A1}$$

In the sequel, we show that the right-hand side of (A1) converges to zero as T grows large. This can be achieved by bounding the probabilities $\mathrm{pr}\{Q(\theta, \lambda_T) \leqslant Q(\theta_0, \lambda_T)\}$ for each fixed $\theta \neq \theta_0$ in an appropriate way. To do so, write

$$\mathrm{pr}\{Q(\theta, \lambda_T) \leqslant Q(\theta_0, \lambda_T)\} = \mathrm{pr}\{V_\theta \leqslant -B_\theta - 2S_\theta^\varepsilon - 2S_\theta^g + 2W_\theta^\varepsilon + W_\theta^g + \lambda_T(\theta_0 - \theta)\} \tag{A2}$$

with

$$\begin{aligned} V_\theta &= \varepsilon^{\mathrm{T}}(\Pi_{\theta_0} - \Pi_\theta)\varepsilon, \quad & B_\theta &= (X_{\theta_0}\beta)^{\mathrm{T}}(I - \Pi_\theta)(X_{\theta_0}\beta), \\ S_\theta^\varepsilon &= \varepsilon^{\mathrm{T}}(I - \Pi_\theta)X_{\theta_0}\beta, \quad & S_\theta^g &= g^{\mathrm{T}}(I - \Pi_\theta)X_{\theta_0}\beta, \\ W_\theta^\varepsilon &= \varepsilon^{\mathrm{T}}(\Pi_\theta - \Pi_{\theta_0})g, \quad & W_\theta^g &= g^{\mathrm{T}}(\Pi_\theta - \Pi_{\theta_0})g. \end{aligned}$$

We proceed in two steps. In the first, we analyse the asymptotic behaviour of the terms V_θ, B_θ, S_θ^ε, S_θ^g, W_θ^ε and W_θ^g one after the other. In the second, we combine the results on the various terms to obtain an appropriate bound on the probabilities $\mathrm{pr}\{Q(\theta, \lambda_T) \leqslant Q(\theta_0, \lambda_T)\}$.

The overall proof strategy outlined above is similar to that known from other problems based on l_0-penalties such as variable or lag selection in a linear regression model; see for example the proofs in Nishii (1984), Zhang (1992) or Zheng & Loh (1995). Nevertheless, the specific arguments of our proof are rather different. The main reasons are as follows: first of all, we have to accommodate terms that result from incorporating a nonparametric trend function in the model. More importantly, the models corresponding to different candidate periods in our framework are not nested. Even worse, there are no two models that have a regressor in common: for any pair of models, the two sets of regressors, i.e., the two sets of indicator functions, are disjoint. The problem of selecting the true period θ_0 is thus rather different from the problem of selecting the true subset of variables in a regression model.

In what follows, we give a brief summary of the two main steps of the proof. The technical details, in particular the proofs of Lemmas A3 and A4, can be found in the Supplementary Material. To examine the asymptotic behaviour of the various terms in the first step, we distinguish between two cases:

Case A: $\theta \neq \theta_0$ and θ is not a multiple of θ_0.
Case B: $\theta \neq \theta_0$ and θ is a multiple of θ_0.

We first consider the terms B_θ, S_θ^ε, S_θ^g, W_θ^ε and W_θ^g. The following lemma characterizes their asymptotic properties, in particular their tail behaviour for large sample sizes. To formulate it, we let $\{\nu_T\}$ be an arbitrary sequence of positive numbers that diverges to infinity and $c > 0$ a fixed constant that is sufficiently small. Moreover, $n = n(\theta)$ is a natural number with $n \leqslant \theta^\times$, where $\theta^\times$ denotes the least common multiple of θ and θ_0. More specifically, $n = \#\mathcal{S}$, where $\mathcal{S}$ is the subset of indices $s \in \{1, \ldots, \theta^\times\}$ for which $\zeta_s = m(s) - \theta_0^{-1}\sum_{k=1}^{\theta_0} m\{(k-1)\theta + s_\theta\} \neq 0$ and $s_\theta = s - \theta\lfloor (s-1)/\theta \rfloor$. The motivation behind this fairly technical definition will become clearer in the proof. Whereas n varies with the period θ, the constants c, C, and T_0 in the lemma depend neither on θ nor on the sample size T.

LEMMA A3. *There exists a natural number T_0 such that for all $T \geqslant T_0$, we have the following results:*

$$\begin{aligned} &\textit{Case} \text{ A}: \quad B_\theta \geqslant c\left(\frac{nT}{\theta}\right), \quad \mathrm{pr}\left\{|S_\theta^\varepsilon| > \nu_T\left(\frac{nT}{\theta}\right)^{1/2}\right\} \leqslant C\nu_T^{-2}, \quad |S_\theta^g| \leqslant Cn, \\ &\textit{Case} \text{ B}: \quad B_\theta = 0, \quad S_\theta^\varepsilon = 0, \quad S_\theta^g = 0. \end{aligned}$$

Moreover, $|W_\theta^g| \leqslant C$ *and* $\mathrm{pr}(|W_\theta^\varepsilon| > \nu_T) \leqslant C\nu_T^{-2}$ *in both Cases* A *and* B.

For the reasons mentioned above, we cannot simply appeal to standard results from variable selection in linear regression models to derive Lemma A3. However, we have the advantage of knowing the explicit structure of the projection matrix Π_θ. Our proof strategy heavily draws on exploiting this specific structure. We finally note that the expression V_θ can be written as $V_\theta = V_{\theta,1} - V_{\theta,2}$ with

$$V_{\theta,1} = \sum_{l=1}^{\mathrm{T}} \frac{1}{K_{l_{\theta_0},T}^{[\theta_0]}} \sum_{k=1}^{K_{l_{\theta_0},T}^{[\theta_0]}} \varepsilon_{(k-1)\theta_0+l_{\theta_0},T}\varepsilon_{l,T}, \qquad V_{\theta,2} = \sum_{l=1}^{\mathrm{T}} \frac{1}{K_{l_\theta,T}^{[\theta]}} \sum_{k=1}^{K_{l_\theta,T}^{[\theta]}} \varepsilon_{(k-1)\theta+l_\theta,T}\varepsilon_{l,T}$$

and $l_\theta = l - \theta\lfloor (l-1)/\theta \rfloor$. The structure of V_θ is thus similar to that of a U-statistic.

With the help of the above remarks on the terms V_θ, B_θ, S_θ^ε, S_θ^g, W_θ^ε and W_θ^g, we can now bound the probabilities $\mathrm{pr}\{Q(\theta,\lambda_T) \leqslant Q(\theta_0,\lambda_T)\}$. Specifically, we can derive the following result.

LEMMA A4. *There exists a natural number T_0 such that for all $T \geqslant T_0$ and for all $\theta \neq \theta_0$ with $1 \leqslant \theta \leqslant \Theta_T$,*

$$\mathrm{pr}\left\{Q(\theta,\lambda_T) \leqslant Q(\theta_0,\lambda_T)\right\} \leqslant C(\rho_T\Theta_T)^{-1}, \tag{A3}$$

where $\{\rho_T\}$ is a sequence of positive numbers that slowly diverges to infinity, e.g., $\rho_T = \log\log T$.

Inserting the bound (A3) into (A1), it immediately follows that $\mathrm{pr}(\hat{\theta} \neq \theta_0) = o(1)$, which in turn yields that $\hat{\theta} = \theta_0 + o_p(1)$, thus completing the proof of Theorem 1. The proof of Lemma A4 is given in the Supplementary Material. Here, we are content with sketching its idea. Using (A2) together with the tail properties summarized in Lemma A3, we can show that

$$\mathrm{pr}\left\{Q(\theta,\lambda_T) \leqslant Q(\theta_0,\lambda_T)\right\} \leqslant \mathrm{pr}\left(V_\theta \leqslant -Cr_T\right) + C(\rho_T\Theta_T)^{-1}$$

with a certain sequence $\{r_T\}$ that diverges to infinity as T grows large. Thus, the tail behaviour of the terms B_θ, S_θ^ε, S_θ^g, W_θ^g and W_θ^ε allows us to replace them by the deterministic sequence $\{-Cr_T\}$, introducing an error term of the order $(\rho_T\Theta_T)^{-1}$. It now remains to bound the tail probability $\mathrm{pr}\left(V_\theta \leqslant -Cr_T\right)$. To do so, we exploit the U-statistic-like structure of V_θ. In particular, we first apply Chebychev's inequality and then use a covariance inequality for mixing variables to bound the resulting sum of higher moments. This completes the proof of Lemma A4. □

Proof of Theorem 2. The overall idea of the proof is as follows: we first compare $\hat{m}$ with an oracle estimator of m that knows the true period θ_0 and show that the difference between these two estimators is asymptotically negligible. This allows us to replace $\hat{m}$ with the oracle estimator whose asymptotic properties can be derived by standard arguments. The technical details are given in the Supplementary Material. □

Proof of Theorem 3. Similarly to the proof of Theorem 2, we replace $\hat{g}$ with an oracle estimator and then derive the properties of the latter. The details are again deferred to the Supplementary Material. □

REFERENCES

AKAIKE, H. (1969). Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.* **21**, 243–7.

ATAK, A., LINTON, O. & XIAO, Z. (2011). A semiparametric panel model for unbalanced data with application to climate change in the United Kingdom. *J. Economet.* **164**, 92–115.

BLOOMFIELD, P. (1992). Trends in global temperature. *Climatic Change* **21**, 1–16.

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

BROHAN, P., KENNEDY, J. J., HARRIS, I., TETT, S. F. B. & JONES, P. D. (2006). Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.* **111**, D12106.

BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*. Berlin: Springer.

CLAESKENS, G. & HJORT, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

DAHLHAUS, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.* **25**, 1–37.

DELWORTH, T. L. & MANN, M. E. (2000). Observed and simulated multidecadal variability in the northern hemisphere. *Climate Dynam.* **16**, 661–76.
DURBIN, J. & KOOPMAN, S. J. (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.
FAN, J. & YAO, Q. (2005). *Nonlinear Time Series*. Berlin: Springer.
GASSIAT, E. & LÉVY-LEDUC, C. (2006). Efficient semiparametric estimation of the periods in a superposition of periodic functions with unknown shape. *J. Time Ser. Anal.* **27**, 877–910.
GENTON, M. G. & HALL, P. (2007). Statistical inference for evolving periodic functions. *J. R. Statist. Soc.* B **69**, 643–57.
HALL, P. & YIN, J. (2003). Nonparametric methods for deconvolving multiperiodic functions. *J. R. Statist. Soc.* B **65**, 869–86.
HALL, P. & LI, M. (2006). Using the periodogram to estimate period in nonparametric regression. *Biometrika* **93**, 411–24.
HALL, P., REIMANN, J. & RICE, J. (2000). Nonparametric estimation of a periodic function. *Biometrika* **87**, 545–57.
HANNAN, E. J. (1957). The variance of the mean of a stationary process. *J. R. Statist. Soc.* B **19**, 282–5.
HANNAN, E. J. & QUINN, B. G. (1979). The determination of the order of an autoregression. *J. R. Statist. Soc.* B **41**, 190–5.
HANSEN, J., RUEDY, R., SATO, M. & LO, K. (2002). Global warming continues. *Science* **295**, 275.
HART, J. D. (1991). Kernel regression estimation with time series errors. *J. R. Statist. Soc.* B **53**, 173–87.
HARVEY, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
HARVEY, A. C. (2006). Forecasting with unobserved components time series models. In *Handbook of Economic Forecasting*, vol. 1, Ed. G. Elliott, C. W. J. Granger and A. Timmermann, pp. 327–412, Amsterdam: North-Holland.
DE JONG, R. M. & DAVIDSON, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica* **68**, 407–23.
KOOPMAN, S. J., HARVEY, A. C., DOORNIK, J. A. & SHEPHARD, N. (1999). *Structural Time Series Analysis, Modelling, and Prediction using STAMP*. London: Timberlake Consultants Press.
MAZZARELLA, A. (2007). The 60-year solar modulation of global air temperature: The Earth's rotation and atmospheric circulation connection. *Theor. Appl. Climatol.* **88**, 193–9.
NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12**, 758–65.
QUINN, B. G. & HANNAN, E. J. (2001). *The Estimation and Tracking of Frequency*. Cambridge: Cambridge University Press.
QUINN, B. G. & THOMSON, P. J. (1991). Estimating the frequency of a periodic function. *Biometrika* **78**, 65–74.
RICE, J. A. & ROSENBLATT, M. (1988). On frequency estimation. *Biometrika* **75**, 477–84.
ROBINSON, P. M. (1989). Nonparametric estimation of time-varying parameters. In *Statistical Analysis and Forecasting of Economic Structural Change,* Ed. P. Hackl, pp. 253–64. Berlin: Springer.
SCHLESINGER, M. E. & RAMANKUTTY, N. (1994). An oscillation in the global climate system of period 65–70 years. *Nature* **367**, 723–6.
SUN, Y., HART, J. D. & GENTON, M. G. (2012). Nonparametric inference for periodic sequences. *Technometrics* **54**, 83–96.
TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
WALKER, A. M. (1971). On the estimation of a harmonic component in a time series with stationary independent residuals. *Biometrika* **58**, 21–36.
ZHANG, P. (1992). On the distributional properties of model selection criteria. *J. Am. Statist. Assoc.* **87**, 732–7.
ZHENG, X. & LOH W.-Y. (1995). Consistent variable selection in linear models. *J. Am. Statist. Assoc.* **90**, 151–6.

[*Received November* 2012. *Revised April* 2013]