Vogt, Michael; Linton, Oliver

Working Paper

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

cemmap working paper, No. CWP23/12

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Vogt, Michael; Linton, Oliver (2012) : Nonparametric estimation of a periodic sequence in the presence of a smooth trend, cemmap working paper, No. CWP23/12, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2012.2312

This Version is available at:
https://hdl.handle.net/10419/64724

cemmap
centre for microdata methods and practice

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

Michael Vogt
Oliver Linton

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP23/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nonparametric Estimation of a Periodic Sequence in the Presence of a Smooth Trend*

Michael Vogt[1] University of Cambridge

Oliver Linton[2] University of Cambridge

September 10, 2012

In this paper, we study 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 as well as the limiting distributions of our estimators of the periodic component and the trend function. The asymptotic results are complemented with a simulation study that investigates the small sample behaviour of our procedure. Finally, we illustrate our method by applying it to a series of global temperature anomalies.

Key words: Nonparametric estimation; penalized least squares; periodic sequence; temperature anomaly data.
AMS 2010 subject classifications: 62G08, 62G20, 62P12.

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 deterministic time trend and a stochastic noise process. Usually, there is not much known about the structure of the periodic and the trend component. It is thus important to have flexible semi- and nonparametric methods at hand to estimate them.

*The authors would like to thank the ERC for financial support.
[1]Address: Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD. Email: mv346@cam.ac.uk.
[2]Address: Faculty of Economics, Austin Robinson Building, Sidgwick Avenue, Cambridge, CB3 9DD. Email: obl20@cam.ac.uk.

In this paper, we develop estimation theory for the periodic and the trend component 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

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

with $\mathbb{E}[\varepsilon_{t,T}] = 0$, where g is a smooth deterministic trend and m is a periodic component with unknown period θ_0. We do not impose any parametric restrictions on m and g. Moreover, we do not assume the noise process $\{\varepsilon_{t,T}\}$ to be stationary but merely put some short-range dependence conditions on it. As usual in nonparametric regression, the time argument of the trend function g is rescaled to the unit interval. We comment on this feature in more detail in Section 2 which discusses 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 $\theta_0 \geq 1$ 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 is an infinite number of functions on $\mathbb{R}$ which take the values $m(t)$ at the points $t \in \mathbb{Z}$. The function which 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 $\frac{\theta_0}{n}$ for some $n \in \mathbb{N}$ as its period. Hence, in our design with equidistant observation points, we are in general neither able to identify the function which 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.

The literature so far has restricted attention to a simplified version of model (1) without a trend function. The latter is given by the equation $Y_t = m(t) + \varepsilon_t$ with error terms ε_t that are assumed to be stationary. The traditional way to estimate the periodic component m in this setup is a trigonometric regression approach. In this approach, the periodic component gets parametrized by a finite number of sinusoids, i.e. the sequence values $\{m(t)\}_{t\in\mathbb{Z}}$ are given as the function values of a linear combination of parameter-dependent sine waves. The underlying function which 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).

As already indicated, we refrain from adopting such a parametric approach as in many cases we have no information whatsoever about the shape of the periodic model part. The same point of view is taken in a recent paper by Sun, Hart & Genton (2012) who

investigate estimating the period of the sequence $\{m(t)\}_{t\in\mathbb{Z}}$ in the model $Y_t = m(t)+\varepsilon_t$ with i.i.d. residuals ε_t. The authors view the issue of estimating the period as a model selection problem and construct a cross-validation 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 one given that the period is not too small. This property is termed "virtual consistency".

A related strand of the literature is concerned with estimating a periodic function 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, $X_1 < X_2 < \ldots < X_T$ are the time points of observation and the residuals ε_t are i.i.d. 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 $(-\frac{1}{2}, \frac{1}{2})$. 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 (2008). Moreover, it allows to identify the function m without imposing any parametric restrictions on it. The reason is that the random design points get scattered all over the cycle of the function 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 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 model (1). Our estimation procedure is introduced in Section 3 and splits up into three steps. In the first step, we estimate the period θ_0 by a penalized least squares method. Given our estimator of θ_0, we then construct 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. Interestingly, our method is completely robust to the presence of a trend. As explained in more detail later on, the trend component g gets "smoothed out" in a certain way 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 g.

The asymptotic properties of our estimators are described in Section 4. To start with, our estimator of the period θ_0 is shown to be consistent. Moreover, we derive the convergence rates and asymptotic normality results 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.
To complement the asymptotic analysis of the paper, we investigate the small sample behaviour of our estimators by a simulation study in Section 6. Moreover, we apply our method to a sample of yearly global temperature anomalies from 1850 to 2011 in Section 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 the introduction, the model equation is given by

$$Y_{t,T} = g\Big(\frac{t}{T}\Big) + m(t) + \varepsilon_{t,T} \quad \text{for } t = 1, \dots, T$$

with $\mathbb{E}[\varepsilon_{t,T}] = 0$, where 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 $\int_0^1 g(u)du = 0$. As shown in Lemma A2 in the appendix, this uniquely pins down g and $\{m(t)\}_{t\in\mathbb{Z}}$.
The trend function g in model (1) depends on rescaled time $\frac{t}{T}$ 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), Dahlhaus (1997), Dahlhaus & Subba Rao (2006) and Zhou & Wu (2009) among many 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. This allows us to do reasonable asymptotics in this framework.
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 $\{1, \dots, \theta_0\}$. 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.

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

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

where $I_s(t) = I(t = k\theta_0 + s \text{ for some } k)$ and $I(\cdot)$ is an 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

$$Y = g + X_{\theta_0}\beta + \varepsilon, \tag{3}$$

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

3 Estimation Procedure

Our estimation procedure splits up into three steps. In the first step, we estimate the unknown period θ_0. The estimation of the sequence values $\{m(t)\}_{t\in\mathbb{Z}}$ is addressed in the second step. In the final step, we provide an estimator of the nonparametric trend component g.

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 $1 \leq \theta \leq \Theta_T$. Here, the upper bound Θ_T is not fixed but 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

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

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

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

where $\|x\| = (\sum_{t=1}^{T} x^2)^{1/2}$ denotes the usual l_2-norm for vectors $x \in \mathbb{R}^T$.
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.
One way to overcome this problem is to add a regularization term to the residual sum of squares which 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 defined as the minimizer

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

where the upper bound Θ_T may tend to infinity as the sample size T increases. In Section 4.2, we discuss the exact rates at which Θ_T is allowed to diverge.
Note that 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 lag selection, see e.g. Hannan & Quinn (1979). 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 amongst these. Similar to Sun, Hart & Genton (2012), we thus look at estimating the period θ_0 from the perspective of model selection. Nevertheless, our selection method strongly differs from their cross-validation 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 e.g.

the original LASSO article by Tibshirani (1996) and the book by 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 completely ignore the trend function g. As will become clear from our technical arguments, this is possible because g is "smoothed out" in a certain way: At many points of the proofs, g shows up in sums of the form $\frac{1}{T}\sum_{t=1}^{T} g(\frac{t}{T})$ which approximate the integral $\int_0^1 g(u)du$. As this integral is equal to zero by our normalization of g, these sums converge to zero and can be effectively neglected. In this sense, the function 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.

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

The estimator $\hat{m}(t)$ of the sequence value $m(t)$ at time point t is then defined by writing $\hat{\beta}_{\hat{\theta}} = (\hat{m}(1), \dots, \hat{m}(\hat{\theta}))$ and letting $\hat{m}(s + k\hat{\theta}) = \hat{m}(s)$ for all $s = 1, \dots, \hat{\theta}$ and all k. Hence, by construction, $\hat{m}$ is a periodic sequence with period $\hat{\theta}$. Note that 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.

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 was 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

$$Z_{t,T} = g\Big(\frac{t}{T}\Big) + \varepsilon_{t,T} \tag{4}$$

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

$$\begin{bmatrix} \tilde{g}(u) \\ \partial\tilde{g}(u)/\partial u \end{bmatrix} = \underset{(g_0,g_1)\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{t=1}^{T} \Big(Z_{t,T} - g_0 - g_1\Big(\frac{t}{T} - u\Big)\Big)^2 K_h\Big(u - \frac{t}{T}\Big), \tag{5}$$

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. Here, h denotes the bandwidth and K is a kernel function with $K_h(x) = K(x/h)/h$.

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 (5) by the approximations $\hat{Z}_{t,T}$ yields an estimator $\hat{g}$ which can be computed from the data. Standard calculations show that $\hat{g}(u)$ has the closed form solution

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

with

$$w_{t,T}(u) = K_h\Big(u - \frac{t}{T}\Big)\Big[S_{T,2}(u) - \Big(\frac{t}{T} - u\Big)S_{T,1}(u)\Big]$$

and $S_{T,j}(u) = \sum_{t=1}^{T} K_h(u - \frac{t}{T})(\frac{t}{T} - u)^j$ for $j = 1, 2$.
Note that alternatively to the above local linear estimator, we could have used a somewhat simpler Nadaraya-Watson smoother to estimate the function g. It is however well-known that Nadaraya-Watson smoothing notoriously suffers from boundary problems. To circumvent these issues, we have decided to employ a local linear smoother.

4 Asymptotics

In this section, we describe the asymptotic properties of our estimators. The first subsection lists the assumptions needed for our analysis. The following subsections state the main asymptotic results, with each subsection dealing with a separate step of our estimation procedure.

4.1 Assumptions

We impose the following regularity conditions.

(C1) The error process $\{\varepsilon_{t,T}\}$ is strongly mixing with mixing coefficients $\alpha(k)$ satisfying $\alpha(k) \leq Ca^k$ for some positive constants C and $a < 1$.

(C2) It holds that $\mathbb{E}[|\varepsilon_{t,T}|^{4+\delta}] \leq C$ for some small $\delta > 0$ and a positive constant $C < \infty$.

(C3) The function g is twice continuously differentiable on $[0, 1]$.

(C4) The kernel K is bounded, symmetric about zero and has compact support. Moreover, it fulfills the Lipschitz condition that there exists a positive constant L with $|K(u) - K(v)| \leq L|u - v|$.

We quickly 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. Note that we do not necessarily require exponentially decaying mixing rates as assumed in (C1). These could alternatively be replaced by slower polynomial rates (at the cost of having stronger restrictions on the penalty parameter λ_T later on). To keep the notation and structure of the proofs as clear as possible, we stick to exponential mixing rates throughout. Also note that the smoothness condition (C3) is only needed for the third estimation step, i.e. for establishing the asymptotic properties of the trend function 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 (C1)–(C3) be fulfilled and assume that $\Theta_T \leq CT^{2/5-\delta}$ 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*

$$\hat{\theta} \xrightarrow{P} \theta_0,$$

i.e. $\hat{\theta}$ is a consistent estimator of θ_0.

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 off to infinity, the stronger restrictions have to 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 to an order of almost $T^{3/5}$ if Θ_T diverges at the highest possible rate.

4.3 Asymptotics for the Estimator $\hat{m}$

The next result provides the convergence rate and the limiting distribution of the estimator $\hat{m}$ of the periodic model component. To simplify notation, define

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

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

Theorem 2. *Let the conditions of Theorem 1 be satisfied. Then it holds that*

$$\max_{1 \leq t \leq T} |\hat{m}(t) - m(t)| = O_p\Big(\frac{1}{\sqrt{T}}\Big).$$

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

$$\sqrt{T}(\hat{m}(t) - m(t)) \xrightarrow{d} N(0, V_{t_0}).$$

Note that the limit expression V_{t_0} exists in a wide range of cases, e.g. when imposing some local stationarity assumptions on the error process $\{\varepsilon_{t,T}\}$. Moreover, if the error process is stationary, then V_{t_0} simplifies to $V_{t_0} = \theta_0 \sum_{k=-\infty}^{\infty} \mathrm{Cov}(\varepsilon_{0,T}, \varepsilon_{k\theta_0,T})$. In this case, the long-run variance V_{t_0} can be estimated by classical methods as discussed in Hannan (1957). Estimating the long-run variance in a more general setting which allows for nonstationarities in the data is studied in Newey & West (1987) and 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

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

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 *fulfills (C4).*

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

$$\sup_{u\in[0,1]} \big|\hat{g}(u) - g(u)\big| = O_p\Big(\sqrt{\frac{\log T}{Th}} + h^2\Big).$$

(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* $c_h \geq 0$. *Then it holds that*

$$\sqrt{Th}(\hat{g}(u) - g(u) - h^2 B_u) \xrightarrow{d} N(0, V_u)$$

with $B_u = \frac{1}{2}(\int v^2 K(v)dv) g''(u)$.

Similarly to Theorem 2, the limit V_u exists under rather general conditions, e.g. when imposing some locally stationary structure on the process $\{\varepsilon_{t,T}\}$. If the error process is stationary, then the asymptotic variance V_u simplifies to $V_u = (\int K^2(v)dv) \sum_{l=-\infty}^{\infty} \gamma_\varepsilon(l)$ with $\gamma_\varepsilon(l) = \mathrm{Cov}(\varepsilon_{0,T}, \varepsilon_{l,T})$. For methods to estimate V_u, we again refer to the papers by Hannan (1957), Newey & West (1987) 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 cross-validation or plug-in rules can be employed. Note however that 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 Altman (1990) and 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 certain range of rates, in particular $(\log T)\Theta_T^{3/2} \ll \lambda_T \ll T$. Thus 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 argument how to choose λ_T in a suitable way. To make the argument as clear as possible, we consider a simplified version of model (1). In particular, we analyze the setting

$$Y_t = m(t) + \varepsilon_t,$$

where the errors ε_t are assumed to be i.i.d. with $\mathbb{E}[\varepsilon_t^2] = \sigma^2$. We thus drop the trend component from the model and assume that there is no serial dependence at all 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 rather than θ_0 itself. We thus focus attention on periods θ which are multiples of θ_0, i.e. $\theta = r\theta_0$ for some r. Let $\hat{\beta}_\theta = (\hat{\beta}_{\theta,1}, \ldots, \hat{\beta}_{\theta,\theta})$ be the least squares estimator based on the period θ. For ease of notation, we define the shorthand $I_s(t) = I(t = k\theta + s \text{ for some } k)$ and write $(\beta_1, \ldots, \beta_\theta)$ with $\beta_s = \beta_{s-\theta_0\lfloor s/\theta_0 \rfloor}$. With this, it holds that

$$\frac{\mathrm{RSS}(\theta)}{T} = \frac{1}{T}\sum_{t=1}^{T}\big(Y_t - \hat{\beta}_{\theta,1}I_1(t) - \ldots - \hat{\beta}_{\theta,\theta}I_\theta(t)\big)^2.$$

As $Y_t - \beta_1 I_1(t) - \ldots - \beta_\theta I_\theta(t) = \varepsilon_t$ for $\theta = r\theta_0$, we further obtain

$$\frac{\mathrm{RSS}(\theta)}{T} = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 + \frac{2}{T}\sum_{t=1}^{T}\varepsilon_t\big[(\beta_1 - \hat\beta_{\theta,1})I_1(t) + \ldots + (\beta_\theta - \hat\beta_{\theta,\theta})I_\theta(t)\big] + \frac{1}{T}\sum_{t=1}^{T}\big[(\beta_1 - \hat\beta_{\theta,1})I_1(t) + \ldots + (\beta_\theta - \hat\beta_{\theta,\theta})I_\theta(t)\big]^2.$$

Inspecting the definition of the least squares estimator $\hat\beta_\theta$, it can be seen that $\hat\beta_{\theta,s} = (K_{s,T}^{[\theta]})^{-1}\sum_{t=1}^{T} I_s(t)Y_t$ with $K_{s,T}^{[\theta]} = 1 + \lfloor (T-s)/\theta \rfloor$. Thus

$$\beta_s - \hat\beta_{s,\theta} = -\frac{1}{K_{s,T}^{[\theta]}}\sum_{t=1}^{T} I_s(t)\varepsilon_t.$$

Using this, some straightforward calculations yield that

$$\frac{\mathrm{RSS}(\theta)}{T} = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 - \sum_{s=1}^{\theta}\frac{1}{T}\Big(\frac{1}{K_{s,T}^{[\theta]}}\sum_{t,t'=1}^{T} I_s(t)I_s(t')\varepsilon_t\varepsilon_{t'}\Big)$$

and hence

$$\mathbb{E}\Big[\frac{\mathrm{RSS}(\theta)}{T}\Big] = \sigma^2 - \frac{\sigma^2\theta}{T}.$$

As a result,

$$\mathbb{E}\Big[\frac{\mathrm{RSS}(r\theta_0)}{T}\Big] = \sigma^2 - \frac{\sigma^2(r\theta_0)}{T} < \sigma^2 - \frac{\sigma^2\theta_0}{T} = \mathbb{E}\Big[\frac{\mathrm{RSS}(\theta_0)}{T}\Big]$$

or put differently,

$$\mathbb{E}\big[\mathrm{RSS}(r\theta_0)\big] + \sigma^2 r\theta_0 = \mathbb{E}\big[\mathrm{RSS}(\theta_0)\big] + \sigma^2\theta_0. \tag{6}$$

Formula (6) 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.

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

$$\lambda_T = \sigma^2\kappa_T \tag{7}$$

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.

Repeating our heuristic argument with serially correlated errors, the variance σ^2 gets replaced by some type of long-run variance which incorporates covariance terms of the errors. Our selection rule of the penalty parameter λ_T does not take into account this effect of the dependence structure at all. Nevertheless, this does not mean that it becomes useless when the error terms are correlated. As long as the correlation is not too strong, σ^2 will be the dominant term in the long-run variance. Hence, our heuristic rule should still yield an appropriate penalty parameter λ_T. This consideration is confirmed by our simulations later on, where the error terms are assumed to follow an AR(1) process.

As the error variance σ^2 is unknown in general, we cannot take the formula (7) at face value but have to replace σ^2 with an estimator. This can be achieved as follows: To start with, define

$$\check{\theta} = \min_{1 \leq \theta \leq \Theta_T} \mathrm{RSS}(\theta).$$

As already noted in Subsection 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

$$\mathbb{P}\big(\check{\theta} = k\theta_0 \text{ for some } k \in \mathbb{N}\big) \to 1$$

as $T \to \infty$. Thus, 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 estimate $\hat{\beta}_{\check{\theta}}$ 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 estimates $\check{m}(t)$ and $\check{g}(\frac{t}{T})$ 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 $\check{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} \check{\varepsilon}_t^2$ of the error variance σ^2.

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

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

and a period $\theta_0 = 60$. Moreover, the trend function g is given by

$$g(u) = 2u^2.$$

The functions m and g are depicted in Figure 1. The error terms ε_t of the simulated model are drawn from the AR(1) process $\varepsilon_t = 0.45\varepsilon_{t-1} + \eta_t$, where η_t are i.i.d. 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 setup is chosen to mimic the situation in the real data example investigated in the subsequent section.

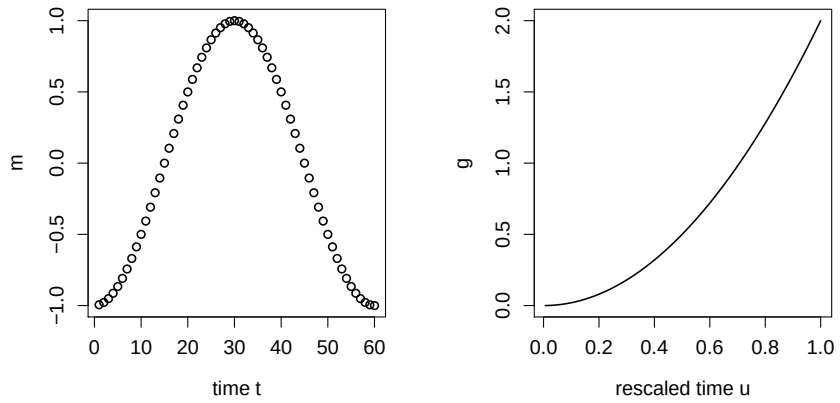

Figure 1: Plot of the functions m and g in our simulation setup.

We simulate the model $N = 1000$ times for three different sample sizes $T = 160$, 250, 500 and three different values of the residual variance $\sigma_\eta^2 = 0.2$, 0.4, 0.6. Note that these values of σ_η^2 translate into an error variance $\sigma^2 = \mathbb{E}[\varepsilon_t^2]$ of approximately 0.25, 0.5, and 0.75, respectively. To get a rough idea of the noise level in our setup, we consider the ratio $\overline{\varepsilon^2}/\overline{Y^2} := (\sum_{t=1}^T \varepsilon_t^2)/(\sum_{t=1}^T Y_t^2)$, 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 $\sigma^2 = 0.25$, 0.5, 0.75, we obtain $\overline{\hat{\varepsilon}^2}/\overline{Y^2} \approx 0.12$, 0.2, 0.26. Note that 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 an estimator of the error variance σ^2 which is constructed as explained at the end of Section 5. We thus pick λ_T according to the heuristic idea described there. Note that 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 $\hat{g}$ strongly 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.

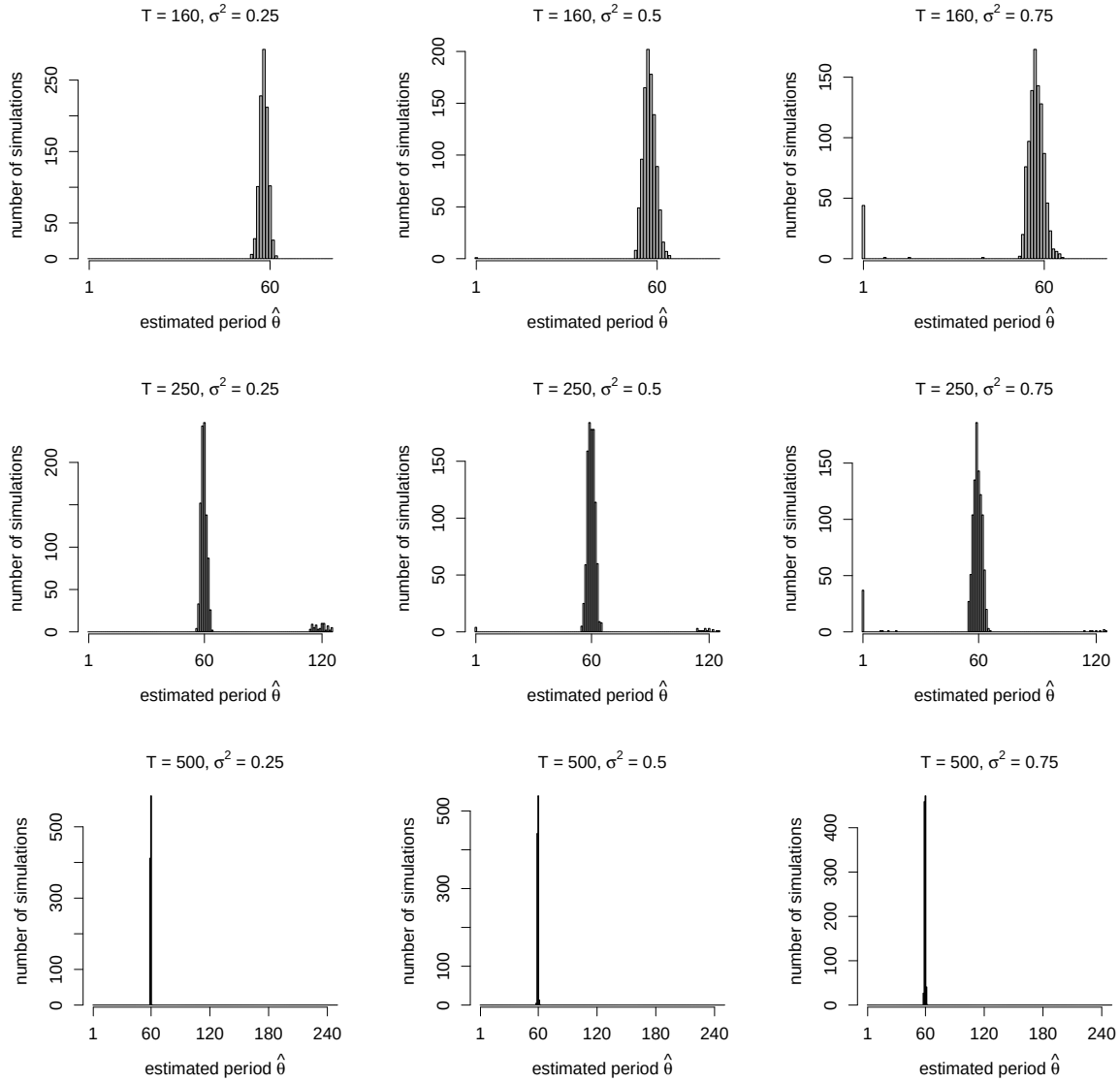

Figure 2: Simulation results for different choice of the sample size T and the error variance σ^2. The bars give the number of simulations (out of a total of 1000) in which a certain value $\hat{\theta}$ is obtained.

The simulation results are presented in Figure 2. Each panel shows the distribution of $\hat{\theta}$ for a specific choice of T and σ^2. The bars give the number of simulations (out of a total of 1000) in which a certain value $\hat{\theta}$ is obtained. For each sample size, we take into account periods θ with $1 \leq \theta \leq T/2$. We now summarize the most important features of the results:

(a) The estimates $\hat{\theta}$ cluster around the true period θ_0. In addition, smaller clusters can be found around multiples of the period θ_0. As can be seen from the proof of Theorem 1, this behaviour of $\hat{\theta}$ is suggested by the asymptotic theory.

(b) For smaller sample sizes, in particular for $T = 160$, the clusters are not exactly centered around the true period θ_0 but are somewhat biased towards smaller values. As turns out, this finite sample effect is due to the trend component in the model. When repeating the simulations without a trend function, the bias can be seen to vanish completely.

(c) The clusters become more dispersed when moving towards larger values of the error variance σ^2. This intuitively makes sense as the signal-to-noise ratio deteriorates with increasing σ^2, making it harder to estimate θ_0.

(d) Inspecting the results for $\sigma^2 = 0.75$, one can see that $\hat{\theta}$ is equal to 1 in a non-negligible number of simulations. This is a finite sample effect which is strongest for $T = 160$ and vanishes as the sample size increases. As will become clear in the discussion at the end of this section, this effect 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 making the penalty λ_T a bit smaller.

Overall, the simulations suggest that the estimator $\hat{\theta}$ performs well in small samples. Even at a sample size of $T = 160$ where we only observe 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 a high percentage of the simulations and misses the true value only very slightly throughout.

Before we close this section, we have a closer look at what happens when the regularization parameter λ_T is varied. Figure 3 presents the criterion function $Q(\theta, \lambda_T)$ for a typical simulation with $T = 500$, $\sigma^2 = 0.5$ and three different choices of λ_T. In panel (a), we have chosen the regularization parameter as above, 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 feature 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. This is due to the fact that the penalty $\lambda_T\theta$ is a linear function in θ with the slope parameter λ_T. If the slope λ_T is picked too large, then the criterion function moves up too quickly. As a result, the global minimum does not lie at the first downward spike around θ_0 but at $\theta = 1$. This situation is illustrated in panel (b). If λ_T is chosen too small on the other hand, then the criterion function decreases, taking its global minimum not at the first downward spike but at a subsequent one. This situation is depicted in panel (c).

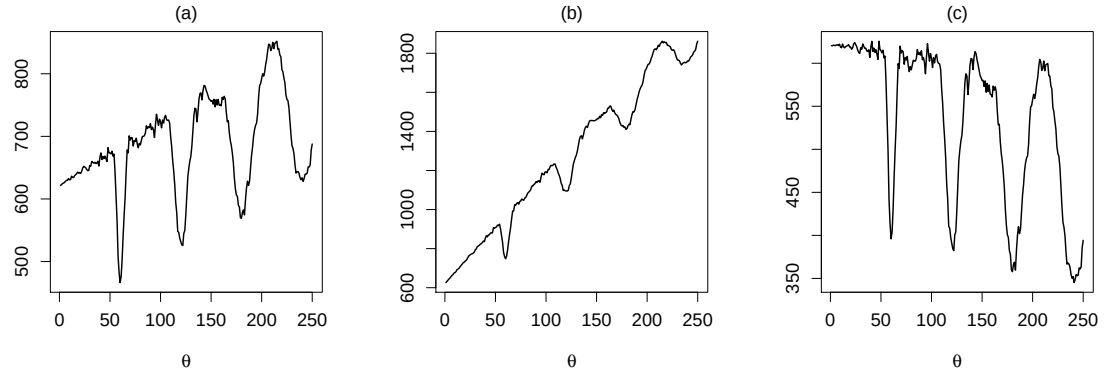

Figure 3: Plot of the criterion function for a typical simulation with $T = 500$, $\sigma^2 = 0.5$ and three different choices of λ_T. In particular, λ_T is given by $\lambda_T^{(a)} = \check{\sigma}^2 \log T$, $\lambda_T^{(b)} = 4\lambda_T^{(a)}$ and $\lambda_T^{(c)} = \lambda_T^{(a)}/4$ in the three different panels.

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 have a glance at the criterion function. If it exhibits large downward spikes at a certain 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 and keep in mind that 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 (cp. Bloomfield (1992) or Hansen et al. (2002) among others). This global warming trend is also visible in the time series presented in Figure 4. 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 Celsius degree).[3]

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 not only contain 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

[3]The data set is called HadCRUT3 and can be downloaded from `http://www.cru.uea.ac.uk/cru/data/temperature`. A detailed description of the data is given in Brohan et al. (2006).

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 which is able to accurately separate the periodic and the trend component. Otherwise, an inaccurate picture of the warming trend emerges. Moreover, a precise estimate of both components is needed to reliably predict future temperature changes.

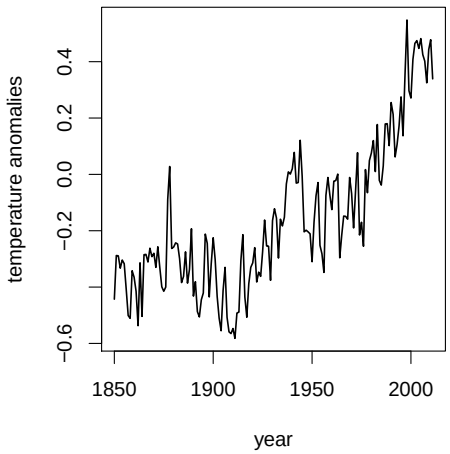

Figure 4: Yearly global temperature anomalies from 1850 to 2011 (measured in °C).

In what follows, we apply our three-step procedure to the temperature anomalies from Figure 4. We thus fit the model

$$Y_{t,T} = g\Big(\frac{t}{T}\Big) + m(t) + \varepsilon_{t,T}$$

with $\mathbb{E}[\varepsilon_{t,T}] = 0$ 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.

To estimate the period θ_0, we employ our penalized least squares method with the penalty term $\lambda_T = \check{\sigma}^2 \log T$. As in the simulations, $\check{\sigma}^2$ is an estimate of the error variance which is constructed as described in Section 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 Figure 5. 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.

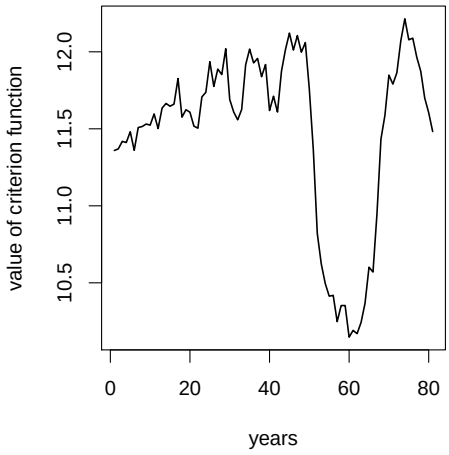

Figure 5: Plot of the criterion function $Q(\theta, \lambda_T)$.

We next turn to the estimation of the periodic component m. The estimator $\hat{m}$ is presented in the left-hand panel of Figure 6 over a full cycle of 60 years. A smoothed version of $\hat{m}$ is plotted as the solid black curve in the right-hand panel. The grey time series in the background displays the detrended anomaly data, i.e. the values $Y_{t,T} - \hat{g}(\frac{t}{T})$ with $\hat{g}$ being the estimator of the trend g.

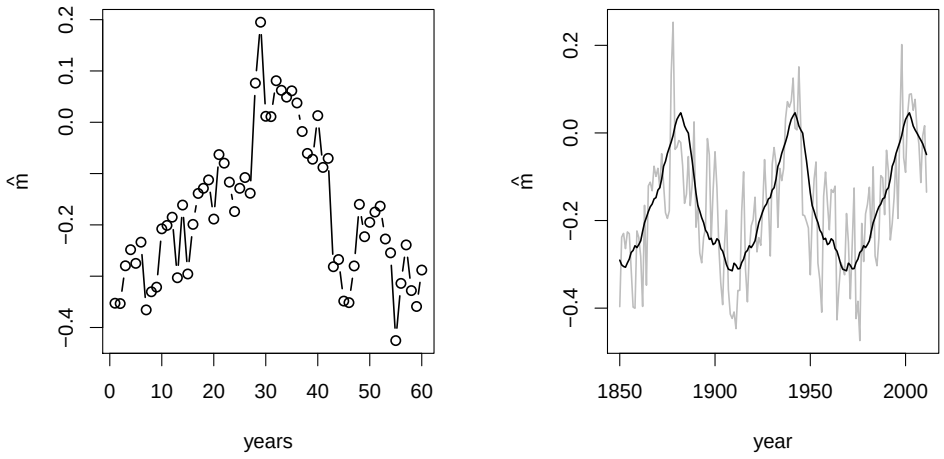

Figure 6: Estimation results for the periodic component m. The left-hand panel presents the estimator $\hat{m}$, the right-hand panel a smoothed version of it. The grey time series in the background are detrended temperature anomalies.

The estimation results concerning the trend function g are depicted in Figure 7. The solid curve in the left-hand panel shows the local linear smoother $\hat{g}$, the dashed lines are the corresponding 95% pointwise confidence bands. The right-hand panel once again displays the estimator $\hat{g}$, but this time against the background of the anomaly data from which the periodic component has been removed. For the estimation, we have used an Epanechnikov kernel and have chosen the bandwidth to equal $h = 0.15$. 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 only report the

findings for the bandwidth $h = 0.15$.

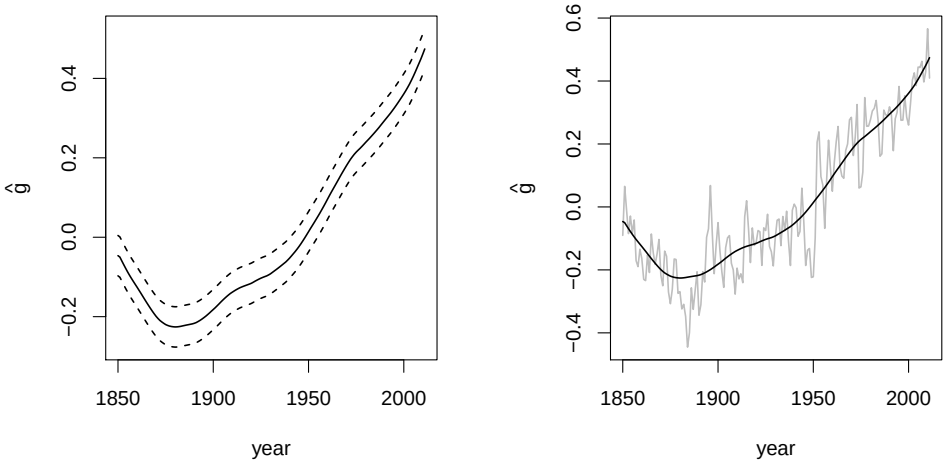

Figure 7: Estimation results for the trend function g. The solid line both in the left- and right-hand panel is the smoother $\hat{g}$. The dashed lines are pointwise 95% confidence bands, the grey time series in the background displays the data points $Y_{t,T} - \hat{m}(t)$.

Figure 8 depicts the time series of the estimated residuals $\hat{\varepsilon}_{t,T} = Y_{t,T} - \hat{g}(\frac{t}{T}) - \hat{m}(t)$ 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.

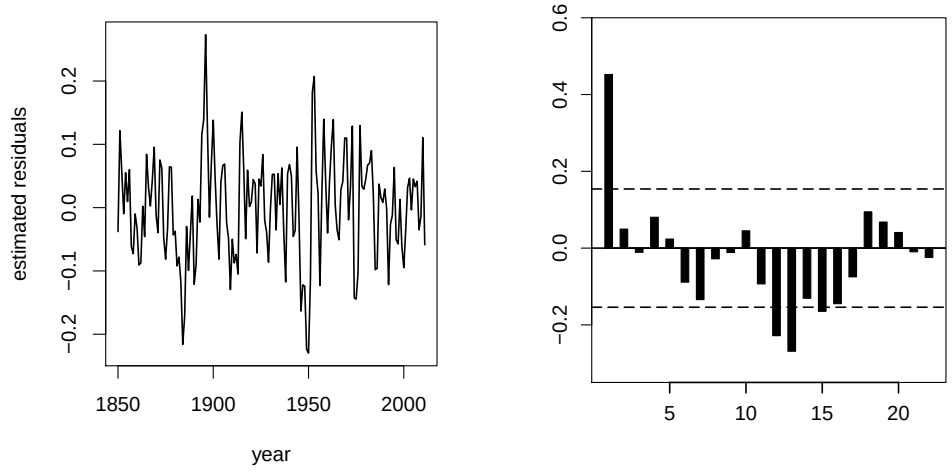

Figure 8: Time series of the estimated residuals (left panel) and its sample autocorrelation function (right panel). The dashed lines show the Bartlett bounds $\pm 1.96T^{-1/2}$.

Moreover, inspecting the sample autocorrelations, the residuals do not appear to be strongly dependent over time. Note that 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 has been used as a guideline in the design of the error terms in the simulations.

8 Variants and Extensions

In this paper, we have studied an additive regression setup featuring a cyclical component with an unknown period and a nonparametric trend function. We have provided a procedure to estimate the unknown period and the values of the periodic sequence as well as the trend function. Moreover, we have derived the asymptotic properties of our estimators. In addition, we have examined the small sample behaviour of our method by a simulation study and have illustrated it by an application to climate data. Our estimation method may be extended in various directions. We close the paper by outlining some of them.

8.1 Trend Estimation

When applying our procedure, we remove the estimated cyclical component from the data before estimating the trend. It is however also possible to set up a direct estimation method for the trend function. In particular, we may naively estimate g by a standard local linear smoother of the form

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

the weights $w_{t,T}(u)$ being defined in Subsection 3.3. The periodic component m enters the estimator $\hat{\hat{g}}(u)$ via the term $\sum_{t=1}^{T} w_{t,T}(u) m(t) / \sum_{t=1}^{T} w_{t,T}(u)$, which is a weighted average of the values $m(t)$. Renormalizing m and g to satisfy $\sum_{s=1}^{\theta_0} m(s) = 0$ for convenience, it is easily seen that $|\sum_{t=1}^{T} w_{t,T}(u) m(t) / \sum_{t=1}^{T} w_{t,T}(u)| \leq C/Th$. Hence, the periodic component gets smoothed out in a similar way as the trend function in Subsections 3.1 and 3.2. As a consequence, $\hat{\hat{g}}$ can be shown to have the same limiting behaviour as the oracle estimator which is based on the deseasonalized observations $Z_{t,T} = Y_{t,T} - m(t)$.

From an asymptotic perspective, it is thus possible to estimate the trend function g without taking into account the periodic model part at all. Nevertheless, this naive way of estimating the trend function should be treated with caution. The reason is that it may produce very poor estimates of g in small samples. In particular, when the period θ_0 is large relative to the sample size, then the estimator $\hat{\hat{g}}$ will tend to pick up the periodic component as part of the trend function. For example, if we estimate the warming trend in our application by $\hat{\hat{g}}$, we will wrongly incorporate the 60-year cyclical component into it. As a result, we obtain a totally distorted picture of the global warming trend.[4]

[4] We do not report the exact results of applying the estimator $\hat{\hat{g}}$ to our sample of temperature data. The details are however available upon request.

8.2 Reversing the Estimation Scheme

The previous subsection suggests that the steps of our estimation procedure may be reversed. Indeed, it is possible to start off with estimating the trend function and then proceed by estimating the periodic component. In what follows, we have a closer look at this modified estimation scheme. For convenience, we again normalize the components m and g to satisfy $\sum_{s=1}^{\theta_0} m(s) = 0$.

Step 1: Estimation of the trend function g. The trend function g can be estimated by the smoother $\hat{\hat{g}}$ defined in the previous subsection. When applying the estimator $\hat{\hat{g}}$ one should however keep in mind its potential pitfalls. In particular, one should avoid using it when the period of the cyclical part is expected to be large relative to the sample size.

Step 2: Estimation of the period θ_0. The period θ_0 may be estimated by applying our penalized least squares procedure to the approximately detrended data $\hat{\hat{W}}_{t,T} = Y_{t,T} - \hat{\hat{g}}(\frac{t}{T})$, where we undersmooth $\hat{\hat{g}}$ by picking the bandwidth h to be of the order $T^{-(\frac{1}{4}+\delta)}$ for some small $\delta > 0$. Let us denote the resulting estimator by $\hat{\hat{\theta}}$. Arguments similar to those for the proof of Theorem 1 show that $\hat{\hat{\theta}}$ consistently estimates the period θ_0.

An obvious drawback of the estimator $\hat{\hat{\theta}}$ is that it depends on the bandwidth h. This contrasts with the estimator $\hat{\theta}$ which is fully independent of h. Given a good choice of the bandwidth, intuition however suggests that the estimator $\hat{\hat{\theta}}$ should be more precise than $\hat{\theta}$. The reasoning is as follows: $\hat{\hat{\theta}}$ is based on preprocessed data from which the trend has been approximately removed. Since the trend plays the role of an additional noise component when it comes to estimating the periodic model part, $\hat{\hat{\theta}}$ should perform better than $\hat{\theta}$.

Having a closer look at the proof of Theorem 1, this intuition turns out to be misguided. As noted in Subsection 3.1, the trend function gets smoothed or integrated out in the proof. In particular, it shows up in sums of the form $S_T = \frac{1}{T}\sum_{t=1}^{T} g(\frac{t}{T})$ which are of the order $O(\frac{1}{T})$. If we estimate the trend in a first step by $\hat{\hat{g}}$, then S_T gets replaced by $\hat{\hat{S}}_T = \frac{1}{T}\sum_{t=1}^{T}[g(\frac{t}{T}) - \hat{\hat{g}}(\frac{t}{T})]$ in the proof. Since the error of estimating g by the smoother $\hat{\hat{g}}$ is of much larger order than $O(\frac{1}{T})$, the sum $\hat{\hat{S}}_T$ will in general be of larger order than S_T as well. Thus, approximately eliminating the trend in a first step tends to introduce additional "noise" in the estimation of θ_0 rather than to reduce it.

Step 3: Estimation of the periodic sequence values. The values $\{m(t)\}_{t\in\mathbb{Z}}$ can be estimated by applying the least squares procedure from Subsection 3.2 to the approximately detrended data $\hat{\hat{W}}_{t,T} = Y_{t,T} - \hat{\hat{g}}(\frac{t}{T})$, where as in the previous step we choose the bandwidth to be of the order $T^{-(\frac{1}{4}+\delta)}$. Going along the lines of the proof for Theorem 2, the resulting estimator $\hat{\hat{m}}(t)$ can be shown to be asymptotically normal for each fixed time point t. However, the limiting distribution will in general differ from that of the

oracle estimator which is based on the exactly detrended data $W_{t,T} = Y_{t,T} - g(\frac{t}{T})$. Thus, the error of estimating the trend function gets reflected in the asymptotic distribution of the periodic sequence values. This again indicates that approximately eliminating the trend in a first step tends to increase the "noise" in the subsequent estimation steps rather than to decrease it.

The above remarks show that our estimation scheme can in principle be reversed. One should however keep in mind that setting up the procedure in this way comes along with some disadvantages and potential pitfalls.

8.3 Iterating the Estimation Scheme

It is also possible to iterate our procedure. In particular, we can set up a backfitting scheme of a similar type as described in Section 8.5 of Hastie & Tibshirani (1990):

(1) Perform the three steps of the estimation procedure described in Section 3. This yields initial estimates $\hat{\theta}^{(0)} = \hat{\theta}$, $\hat{m}^{(0)} = \hat{m}$ and $\hat{g}^{(0)} = \hat{g}$.

(2) Apply the first two estimation steps to the approximately detrended data $Y_{t,T} - \hat{g}^{(0)}(\frac{t}{T})$. This yields updated estimates $\hat{\theta}^{(1)}$ and $\hat{m}^{(1)}$.

(3) Apply the third estimation step to the data $Y_{t,T} - \hat{m}^{(1)}(t)$. This yields an updated estimator $\hat{g}^{(1)}$.

(4) Steps (2) and (3) may be repeated to get further updates of the estimators.

The motivation behind such a scheme is to improve the quality of the estimators. From an asymptotic point of view, there is however no gain at all from performing one or more backfitting steps. Moreover, backfitting comes along with the same disadvantages as reversing the estimation scheme. It is thus questionable whether backfitting pays off in any way when working with a specific sample of data.

8.4 Multiple Periods

In some applications, the periodic sequence m can be expected to be a superposition of multiple periodic components. Neglecting the trend function for simplicity, we may for example consider a model with two periods given by

$$Y_t = m_1(t) + m_2(t) + \varepsilon_t, \tag{8}$$

where m_i is a periodic sequence with unknown (smallest) period θ_i for $i = 1, 2$. The superposition $m = m_1 + m_2$ is periodic as well. As before, we denote its (smallest) period by θ_0. In many situations, θ_0 equals the least common multiple of θ_1 and θ_2. As shown in Restrepo & Chacón (1998), this is however not always the case. Applying

our penalized least squares method to model (8) yields a consistent estimator of θ_0. Hence, if we ignore the multiperiodic structure of the model, our procedure results in estimating the period θ_0 of the superposition m.

Sometimes, however, we are not primarily interested in estimating the period of the superposition but want to find out about the periods of the individual cyclical components. Tackling this problem is complicated by the fact that the periods θ_1 and θ_2 are not uniquely identified in general. Even though the superposition m and its period θ_0 are identified by Lemma A2, the superposition may be generated by different pairs of periodic sequences having different periods. More formally, let Θ be the set of pairs (θ_1, θ_2) such that there exist periodic sequences m_1 and m_2 with $m = m_1 + m_2$. In general, Θ contains more than one pair of periods.[5]

One possible way to estimate the elements of Θ is to construct a two-dimensional (or more generally a multi-dimensional) version of our penalized least squares method. Informally, the procedure looks as follows: For each pair of candidate periods, we fit a model with two cyclical components to the data and calculate the corresponding residual sum of squares. Our estimator is then defined by minimizing a penalized version of the latter. Which elements of Θ are approximated by this procedure will be determined by the structure of the penalty. For example, if we choose the penalty to have the form $\lambda_T(\theta_1 + \theta_2)$, then we will estimate the pair of periods in Θ with the smallest sum. As far as we can see, it is however not trivial at all to extend our theory to this multi-dimensional case. The main problem is that our proofs for the single-period case heavily draw on the rather simple structure of the design matrix X_θ. In the multiperiod case, this structure gets lost, making it hard to carry over some of the arguments. For the time being, we are thus content with estimating the period θ_0 of the superposition m.

[5] As an example, consider the pair of periodic sequences $\{m_1(1), \ldots, m_1(5)\} = \{-1, 0, 0, 0, 0\}$ and $\{m_2(1), \ldots, m_2(6)\} = \{1, 0, 2, -1, 2, 0\}$ having the periods 5 and 6. The sum of these two sequences generates a periodic sequence of period 30. The same sequence is generated by the sequences $\{m_1(1), \ldots, m_1(3)\} = \{-1, 0, 0\}$ and $\{m_2(1), \ldots, m_2(10)\} = \{1, 0, 2, 0, 2, -1, 2, 0, 2, 0\}$ with the periods 3 and 10.

Appendix

In this appendix, we prove Theorems 1–3. Throughout the appendix, the symbol C is used to denote a universal real constant which may take a different value on each occurrence.

Auxiliary Results

Before we come to the proofs of the main theorems, we state some auxiliary lemmas. The following result is needed at various points in the proofs later on.

Lemma A1. *Let θ be any natural number with $1 \leq \theta \leq \Theta_T$. Moreover, let $s \in \{1, \ldots, \theta\}$ and define $K_{s,T}^{[\theta]} = 1 + \lfloor (T-s)/\theta \rfloor$ to be the number of time points $t \in \{1, \ldots, T\}$ which can be written as $t = s + (k-1)\theta$ for some $k \in \mathbb{N}$. Then*

$$\Big| \frac{1}{K_{s,T}^{[\theta]}} \sum_{k=1}^{K_{s,T}^{[\theta]}} g\Big(\frac{s+(k-1)\theta}{T}\Big) - \int_0^1 g(u)du \Big| \leq \frac{C}{K_{s,T}^{[\theta]}}$$

with 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 $\int_0^1 g(u)du = 0$. More precisely, let $\bar{g}$ be a smooth trend function with $\int_0^1 \bar{g}(u)du = 0$ and $\bar{m}$ a periodic sequence with (smallest) period $\bar{\theta}_0$. If*

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

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

Proof. By assumption, for all $t = 1, \ldots, T$ and all $T = 1, 2, \ldots$,

$$\bar{g}\Big(\frac{t}{T}\Big) - g\Big(\frac{t}{T}\Big) = m(t) - \bar{m}(t).$$

Let $\theta^{\times}$ be the least common multiple of θ_0 and $\bar{\theta}_0$. As m and $\bar{m}$ are periodic with (smallest) period θ_0 and $\bar{\theta}_0$ respectively, they are both periodic with period $\theta^{\times}$. We thus obtain that

$$\bar{g}\Big(\frac{s+(k-1)\theta^{\times}}{T}\Big) - g\Big(\frac{s+(k-1)\theta^{\times}}{T}\Big) = m(s) - \bar{m}(s)$$

for all $s = 1, \ldots, \theta^{\times}$ and $k = 1, \ldots, K_{s,T}^{[\theta^{\times}]}$ with $K_{s,T}^{[\theta^{\times}]} = 1 + \lfloor (T-s)/\theta^{\times} \rfloor$. If $\bar{m} = m$, then clearly $\bar{g} = g$ follows since the points $(s+(k-1)\theta^{\times})/T$ become dense in $[0,1]$ as T increases and the functions $\bar{g}$ and g are smooth. We next assume that $\bar{m}(s) \neq m(s)$

for some $s \in \{1, \dots, \theta^\times\}$ and show that this leads to a contradiction: W.l.o.g. let $m(s) - \bar{m}(s) = d_s > 0$ for some $s \in \{1, \dots, \theta^\times\}$. Then

$$\bar{g}\Big(\frac{s + (k-1)\theta^\times}{T}\Big) - g\Big(\frac{s + (k-1)\theta^\times}{T}\Big) = d_s > 0$$

as well as

$$\frac{1}{K_{s,T}^{[\theta^\times]}} \sum_{k=1}^{K_{s,T}^{[\theta^\times]}} \Big[\bar{g}\Big(\frac{s + (k-1)\theta^\times}{T}\Big) - g\Big(\frac{s + (k-1)\theta^\times}{T}\Big)\Big] = d_s > 0.$$

However, by Lemma A1,

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{K_{s,T}^{[\theta^\times]}} \sum_{k=1}^{K_{s,T}^{[\theta^\times]}} \Big[\bar{g}\Big(\frac{s + (k-1)\theta^\times}{T}\Big) - g\Big(\frac{s + (k-1)\theta^\times}{T}\Big)\Big] \\ = \int_0^1 \bar{g}(u)du - \int_0^1 g(u)du = 0 \neq d_s,\end{aligned}$$

which is a contradiction. □

Proof of Theorem 1

We first introduce some notation. Let

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

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

$$X_\theta^\intercal X_\theta = (I_\theta, I_\theta, \dots) \begin{pmatrix} I_\theta \\ I_\theta \\ \vdots \end{pmatrix} = \begin{pmatrix} K_{1,T}^{[\theta]} & & 0 \\ & \ddots & \\ 0 & & K_{\theta,T}^{[\theta]} \end{pmatrix}$$

with $K_{s,T}^{[\theta]} = 1 + \lfloor (T-s)/\theta \rfloor$ for $s = 1, \dots, \theta$. $K_{s,T}^{[\theta]}$ is the number of time points t in the sample that satisfy $t = s + (k-1)\theta$ for some $k \in \mathbb{N}$. It is either equal to $\lfloor T/\theta \rfloor$ or to $\lfloor T/\theta \rfloor + 1$, in particular $K_{s,T}^{[\theta]} = O(T/\theta)$. The projection matrix Π_θ thus becomes

$$\Pi_\theta = X_\theta D_\theta X_\theta^\intercal = \begin{pmatrix} D_\theta & D_\theta & \dots \\ D_\theta & \ddots & \\ \vdots & & \end{pmatrix}$$

with

$$D_\theta = \begin{pmatrix} 1/K_{1,T}^{[\theta]} & & 0 \\ & \ddots & \\ 0 & & 1/K_{\theta,T}^{[\theta]} \end{pmatrix}.$$

Moreover, rewriting the residual sum of squares $\mathrm{RSS}(\theta)$ in terms of Π_θ yields

$$\begin{aligned}\mathrm{RSS}(\theta) &= (Y - X_\theta\hat{\beta}_\theta)^\intercal(Y - X_\theta\hat{\beta}_\theta)\\ &= ((I-\Pi_\theta)Y)^\intercal((I-\Pi_\theta)Y)\\ &= Y^\intercal(I-\Pi_\theta)Y.\end{aligned}$$

Finally, as already noted at the beginning of the appendix, the symbol C is used to denote a generic constant which 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.
With this notation at hand, we now turn to the proof. Our arguments are based on the inequality

$$\mathbb{P}(\hat{\theta} \neq \theta_0) \leq \sum_{\substack{1\leq\theta\leq\Theta_T\\ \theta\neq\theta_0}} \mathbb{P}\big(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T)\big).$$

In the sequel, we will show the following: If the sample size T is sufficiently large, then for all θ with $1 \leq \theta \leq \Theta_T$ it holds that

$$\mathbb{P}\big(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T)\big) \leq C(\kappa_T\Theta_T)^{-1}, \tag{9}$$

where $\{\kappa_T\}$ is a sequence of positive numbers that slowly diverges to infinity (e.g. $\kappa_T = \log\log T$). From this it immediately follows that

$$\mathbb{P}(\hat{\theta} \neq \theta_0) = o(1),$$

which in turn yields that $\hat{\theta} = \theta_0 + o_p(1)$, thus completing the proof. To show (9) we write for each fixed θ with $\theta \neq \theta_0$ and $1 \leq \theta \leq \Theta_T$,

$$\begin{aligned}&\mathbb{P}\big(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T)\big)\\ &\quad= \mathbb{P}\big(V_\theta \leq -B_\theta - 2S_\theta^\varepsilon - 2S_\theta^g + 2W_\theta^\varepsilon + W_\theta^g + \lambda_T(\theta_0-\theta)\big)\end{aligned}$$

with

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

In what follows, we proceed in two steps. In the first step, we analyze the terms $V_\theta, B_\theta, \ldots$ one after the other. In the second step, we combine the results on the various terms to derive the inequality (9).

To examine the properties of the terms $B_\theta, S_\theta^\varepsilon$ and S_θ^g, we first have a closer look at the expression $(I - \Pi_\theta)(X_{\theta_0}\beta)$ which is the common component of these terms. It holds that

$$\begin{aligned}(I - \Pi_\theta)X_{\theta_0}\beta &= (I - X_\theta D_\theta X_\theta^\intercal)X_{\theta_0}\beta \\ &= \left(I - \begin{pmatrix} D_\theta & D_\theta & \dots \\ D_\theta & \ddots & \\ \vdots & & \end{pmatrix}\right) \begin{pmatrix} m(1) \\ \vdots \\ m(\theta_0) \\ m(1) \\ \vdots \end{pmatrix} \\ &=: (\gamma_{1,T}, \dots, \gamma_{\theta^\times,T}, \gamma_{1,T}, \dots, \gamma_{\theta^\times,T}, \dots)^\intercal\end{aligned}$$

with

$$\gamma_{s,T} = m(s) - \frac{1}{K_{s_\theta,T}^{[\theta]}} \sum_{k=1}^{K_{s_\theta,T}^{[\theta]}} m((k-1)\theta + s_\theta)$$

for $s = 1, \dots, \theta^\times$, where $s_\theta = s - \theta\lfloor \frac{s}{\theta} \rfloor$ and $\theta^\times$ is the least common multiple of θ_0 and θ. A representation of $\gamma_{s,T}$ which will turn out to be useful in what follows is given by

$$\gamma_{s,T} = \zeta_s + R_{s,T} \tag{10}$$

with $R_{s,T} = R_{1,s,T} + R_{2,s,T}$ and

$$\begin{aligned}\zeta_s &= m(s) - \frac{1}{\theta_0} \sum_{k=1}^{\theta_0} m((k-1)\theta + s_\theta) \\ R_{1,s,T} &= \left(1 - \frac{\theta_0}{K_{s_\theta,T}^{[\theta]}} \left\lfloor \frac{K_{s_\theta,T}^{[\theta]}}{\theta_0} \right\rfloor\right) \frac{1}{\theta_0} \sum_{k=1}^{\theta_0} m((k-1)\theta + s_\theta) \\ R_{2,s,T} &= -\frac{1}{K_{s_\theta,T}^{[\theta]}} \sum_{k=\theta_0\lfloor K_{s_\theta,T}^{[\theta]}/\theta_0 \rfloor + 1}^{K_{s_\theta,T}^{[\theta]}} m((k-1)\theta + s_\theta).\end{aligned}$$

The components of the representation in (10) have the following properties: First of all, the remainder satisfies

$$|R_{s,T}| \leq \frac{C\theta_0}{K_{s_\theta,T}^{[\theta]}}, \tag{11}$$

since $|1 - \theta_0/K_{s_\theta,T}^{[\theta]} \cdot \lfloor K_{s_\theta,T}^{[\theta]}/\theta_0 \rfloor| \leq \theta_0/K_{s_\theta,T}^{[\theta]}$ and $|R_{2,s,T}| \leq C\theta_0/K_{s_\theta,T}^{[\theta]}$. To describe the properties of the expressions ζ_s, we distinguish between two cases:

Case A: $\theta \neq \theta_0$ and θ is no multiple of θ_0.

Case B: $\theta \neq \theta_0$ and θ is a multiple of θ_0.

The next lemma summarizes the properties of ζ_s in the above two cases.

Lemma A3. *(i) Assume that Case A holds. Then there exists an index $s \in \{1, \dots, \theta^\times\}$ with $\zeta_s \neq 0$. Moreover, there exists a small constant $\eta > 0$ such that $|\zeta_s| \geq \eta$ whenever $\zeta_s \neq 0$. (ii) If Case B holds, then $\zeta_s = 0$ for all s.*

Note that the constant η does not depend on any model parameters, in particular it is independent of θ and s. We postpone proving the above lemma as well as the subsequent ones until the arguments for Theorem 1 are completed.
Using Lemma A3, we can characterize the behaviour of the terms $B_\theta, S_\theta^\varepsilon$ and S_θ^g. To do so, define $\mathcal{S}$ to be the subset of indices $s \in \{1, \dots, \theta^\times\}$ for which $\zeta_s \neq 0$ and let $\#\mathcal{S} = n$. Moreover, write $\mathcal{S}^c = \{1, \dots, \theta^\times\} \setminus \mathcal{S}$.

Lemma A4. *There exists a natural number T_0 such that for all $T \geq T_0$, we have the following results:*

$$\begin{array}{llll} \textit{Case A:} & B_\theta \geq c\big(\frac{nT}{\theta}\big) & \mathbb{P}\big(|S_\theta^\varepsilon| > \nu_T\sqrt{\frac{nT}{\theta}}\big) \leq C\nu_T^{-2} & |S_\theta^g| \leq Cn \\ \textit{Case B:} & B_\theta = 0 & S_\theta^\varepsilon = 0 & S_\theta^g = 0. \end{array}$$

Here, $c > 0$ is a sufficiently small fixed constant and $\{\nu_T\}$ is an arbitrary sequence of positive numbers which diverges to infinity.

Note that in the above lemma, the constants c, C, and T_0 do neither depend on the period θ nor on the sample size T. Additionally to Lemma A4, the terms W_θ^g and W_θ^ε can be shown to have the following properties.

Lemma A5. *For all $T \geq T_0$, it holds that $|W_\theta^g| \leq C$ and $\mathbb{P}(|W_\theta^\varepsilon| > \nu_T) \leq C\nu_T^{-2}$.*

Finally, note that the term V_θ can be written as

$$\begin{aligned} V_\theta &= \varepsilon^\intercal(\Pi_{\theta_0} - \Pi_\theta)\varepsilon \\ &= \varepsilon^\intercal \left(\begin{pmatrix} D_{\theta_0} & D_{\theta_0} & \dots \\ D_{\theta_0} & \ddots & \\ \vdots & & \end{pmatrix} - \begin{pmatrix} D_\theta & D_\theta & \dots \\ D_\theta & \ddots & \\ \vdots & & \end{pmatrix} \right) \varepsilon \\ &= \sum_{l=1}^{T} \frac{1}{K_{l_{\theta_0},T}^{[\theta_0]}} \sum_{k=1}^{K_{l_{\theta_0},T}^{[\theta_0]}} \varepsilon_{(k-1)\theta_0 + l_{\theta_0},T}\varepsilon_{l,T} - \sum_{l=1}^{T} \frac{1}{K_{l_\theta,T}^{[\theta]}} \sum_{k=1}^{K_{l_\theta,T}^{[\theta]}} \varepsilon_{(k-1)\theta + l_\theta,T}\varepsilon_{l,T} \\ &=: V_{\theta,1} + V_{\theta,2} \end{aligned}$$

with $l_\theta = l - \theta\lfloor l/\theta \rfloor$.

Using the results from Lemmas A4 and A5, we can now analyze the term $\mathbb{P}(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T))$. Set $\nu_T = (\kappa_T\Theta_T)^{1/2}$. In Case A, we obtain

$$\begin{aligned}
&\mathbb{P}\big(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T)\big) \\
&= \mathbb{P}\Big(V_\theta \leq -B_\theta - 2S_\theta^\varepsilon - 2S_\theta^g + 2W_\theta^\varepsilon + W_\theta^g + \lambda_T(\theta_0-\theta)\Big) \\
&\leq \mathbb{P}\Big(V_\theta \leq -B_\theta - 2S_\theta^\varepsilon - 2S_\theta^g + 2W_\theta^\varepsilon + W_\theta^g + \lambda_T(\theta_0-\theta), \\
&\qquad\qquad |S_\theta^\varepsilon| \leq \nu_T\sqrt{\tfrac{nT}{\theta}}, |W_\theta^\varepsilon| \leq \nu_T\Big) \\
&\quad + \mathbb{P}\big(|S_\theta^\varepsilon| > \nu_T\sqrt{\tfrac{nT}{\theta}}\big) + \mathbb{P}\big(|W_\theta^\varepsilon| > \nu_T\big) \\
&\leq \mathbb{P}\Big(V_\theta \leq -B_\theta + C\nu_T\sqrt{\tfrac{nT}{\theta}} + \lambda_T(\theta_0-\theta)\Big) + C\nu_T^{-2}.
\end{aligned}$$

Choosing λ_T to satisfy $\lambda_T/T \to 0$ and noting that the regularization term $\lambda_T(\theta_0-\theta)$ is negative for $\theta > \theta_0$, it can be seen that $C\nu_T\sqrt{\frac{nT}{\theta}} + \lambda_T(\theta_0-\theta) \leq \delta(\frac{nT}{\theta})$ for some arbitrarily small $\delta > 0$ and all $T \geq T_0$ with T_0 being sufficiently large. Hence,

$$-B_\theta + C\nu_T\sqrt{\frac{nT}{\theta}} + \lambda_T(\theta_0-\theta) \leq -(c-\delta)\frac{nT}{\theta} \leq -C_1\frac{nT}{\theta}$$

for some constant $C_1 > 0$. From this, it follows that

$$\begin{aligned}
\mathbb{P}\big(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T)\big) &\leq \mathbb{P}\Big(V_\theta \leq -C_1\frac{nT}{\theta}\Big) + C\nu_T^{-2} \\
&\leq \mathbb{P}\Big(V_\theta \leq -C_1\frac{T}{\Theta_T}\Big) + C\nu_T^{-2}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{P}\Big(V_\theta \leq -C_1\frac{T}{\Theta_T}\Big) &= \mathbb{P}\Big(V_{\theta,1} + V_{\theta,2} \leq -C_1\frac{T}{\Theta_T}\Big) \\
&\leq \mathbb{P}\Big(|V_{\theta,1}| + |V_{\theta,2}| \geq C_1\frac{T}{\Theta_T}\Big) \\
&\leq \mathbb{P}\Big(|V_{\theta,1}| \geq \frac{C_1T}{2\Theta_T}\Big) + \mathbb{P}\Big(|V_{\theta,2}| \geq \frac{C_1T}{2\Theta_T}\Big) \\
&=: P_{\theta,1} + P_{\theta,2}.
\end{aligned}$$

To deal with the probabilities $P_{\theta,1}$ and $P_{\theta,2}$, we introduce the following concept: We say that an index i_1 is separated from the indices $i_2,\ldots,i_d$ if $|i_1 - i_k| > C_2\log T$ for a sufficiently large constant C_2 (to be specified later on) and all $k = 2,\ldots,d$. With this definition at hand, we can use Chebychev's inequality to get

$$\begin{aligned}
P_{\theta,2} &= \mathbb{P}\Big(\Big|\sum_{l=1}^{T}\frac{1}{K_{l_\theta,T}^{[\theta]}}\sum_{k=1}^{K_{l_\theta,T}^{[\theta]}}\varepsilon_{(k-1)\theta+l_\theta,T}\varepsilon_{l,T}\Big| \geq \frac{C_1T}{2\Theta_T}\Big) \\
&\leq \frac{C\Theta_T^2}{T^2}\sum_{l,l'=1}^{T}\Big(\frac{1}{K_{l_\theta,T}^{[\theta]}K_{l'_\theta,T}^{[\theta]}}\Big)\sum_{k=1}^{K_{l_\theta,T}^{[\theta]}}\sum_{k'=1}^{K_{l'_\theta,T}^{[\theta]}}\mathbb{E}\big[\varepsilon_{(k-1)\theta+l_\theta,T}\varepsilon_{l,T}\varepsilon_{(k'-1)\theta+l'_\theta,T}\varepsilon_{l',T}\big]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{C\Theta_T^2}{T^2} \sum_{(l,l',k,k')\in\Gamma} \Big(\frac{1}{K_{l_\theta,T}^{[\theta]} K_{l'_\theta,T}^{[\theta]}}\Big) \mathbb{E}\big[\varepsilon_{(k-1)\theta+l_\theta,T}\varepsilon_{l,T}\varepsilon_{(k'-1)\theta+l'_\theta,T}\varepsilon_{l',T}\big] \\
&\quad + \frac{C\Theta_T^2}{T^2} \sum_{(l,l',k,k')\in\Gamma^c} \Big(\frac{1}{K_{l_\theta,T}^{[\theta]} K_{l'_\theta,T}^{[\theta]}}\Big) \mathbb{E}\big[\varepsilon_{(k-1)\theta+l_\theta,T}\varepsilon_{l,T}\varepsilon_{(k'-1)\theta+l'_\theta,T}\varepsilon_{l',T}\big] \\
&=: P_{\theta,2,a} + P_{\theta,2,b},
\end{aligned}
$$

where Γ is the set of tuples (l,l',k,k') such that none of the indices is separated from the others and Γ^c is its complement. Since $\mathbb{E}[\varepsilon_{t,T}^4] \leq C$ by assumption and the number of elements contained in Γ is smaller than $C(T\log T)^2$ for some sufficiently large constant C, it immediately follows that $P_{\theta,2,a} \leq C(\Theta_T^2 \log T/T)^2 \leq C(\kappa_T\Theta_T)^{-1}$, keeping in mind that $\Theta_T = o(T^{2/5})$. To cope with the term $P_{\theta,2,b}$, we exploit the mixing conditions on the error variables: For any tuple of indices $(l,l',k,k') \in \Gamma^c$, there exists an index, say l, which is separated from the others. We can thus apply Davydov's inequality to obtain

$$
\begin{aligned}
\big|\mathbb{E}\big[\varepsilon_{(k-1)\theta+l_\theta,T}\varepsilon_{l,T}\varepsilon_{(k'-1)\theta+l'_\theta,T}\varepsilon_{l',T}\big]\big| &= \big|\mathrm{Cov}\big(\varepsilon_{l,T}, \varepsilon_{(k-1)\theta+l_\theta,T}\varepsilon_{(k'-1)\theta+l'_\theta,T}\varepsilon_{l',T}\big)\big| \\
&\leq C\alpha(C_2\log T)^{1-\frac{1}{q}-\frac{1}{r}} \leq CT^{-C_3}
\end{aligned}
$$

with some $C_3 > 0$, where q and r are chosen slightly larger than $4/3$ and 4, respectively. Note that C_3 can be made arbitrarily large by choosing the constant C_2 large enough. Bounding the moments contained in the expression $P_{\theta,2,b}$ in this way, it is easily seen that $P_{\theta,2,b} \leq C(\kappa_T\Theta_T)^{-1}$. An analogous result holds for the term $V_{\theta,1}$. This shows that $\mathbb{P}(Q(\theta,\lambda_T) < Q(\theta_0,\lambda_T)) \leq C(\kappa_T\Theta_T)^{-1}$ in Case A.

Let us now turn to Case B. The regularization term $\lambda_T(\theta_0-\theta)$ plays a crucial role in this case. In particular, it takes over the role of the term B_θ which is now equal to zero. Since S_θ^ε and S_θ^g are equal to zero as well, we have

$$
\begin{aligned}
&\mathbb{P}\big(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T)\big) \\
&= \mathbb{P}\Big(V_\theta \leq 2W_\theta^\varepsilon + W_\theta^g + \lambda_T(\theta_0-\theta)\Big) \\
&\leq \mathbb{P}\Big(V_\theta \leq 2W_\theta^\varepsilon + W_\theta^g + \lambda_T(\theta_0-\theta), |W_\theta^\varepsilon| \leq \nu_T\Big) + C\nu_T^{-2} \\
&\leq \mathbb{P}\Big(V_\theta \leq C\nu_T + \lambda_T(\theta_0-\theta)\Big) + C\nu_T^{-2}.
\end{aligned}
$$

Choosing λ_T such that $\nu_T/\lambda_T \to 0$ and noting that $\theta_0 - \theta < 0$ in Case B, we obtain that $C\nu_T + \lambda_T(\theta_0-\theta) \leq -C_4\lambda_T$ for some positive constant C_4 and T large enough. Hence,

$$
\mathbb{P}\big(Q(\theta,\lambda_T) \leq Q(\theta_0,\lambda_T)\big) \leq \mathbb{P}\big(V_\theta \leq -C_4\lambda_T\big) + C\nu_T^{-2}
$$

and by analogous arguments as for Case A,

$$
\mathbb{P}\big(V_\theta \leq -C_4\lambda_T\big) \leq C\Big(\frac{\Theta_T\log T}{\lambda_T}\Big)^2.
$$

Thus, choosing λ_T to satisfy $\lambda_T \geq \tau_T(\log T)\Theta_T^{3/2}$ with some sequence $\{\tau_T\}$ that slowly diverges to infinity (e.g. $\tau_T = \log\log T$), we get that $\mathbb{P}\big(Q(\theta,\lambda_T) < Q(\theta_0,\lambda_T)\big) \leq C(\kappa_T\Theta_T)^{-1}$ in Case B as well. □

Proof of Lemma A3. It is trivial to see that $\zeta_s = 0$ for all s in Case B. We thus only have to consider Case A.

We first show that there exists an index $s \in \{1,\dots,\theta^\times\}$ with $\zeta_s \neq 0$. The proof proceeds by contradiction: Suppose there exists some θ with $\zeta_s = 0$ for all $s \in \mathbb{N}$ (or equivalently for all $s \in \{1,\dots,\theta^\times\}$). As $s_\theta = (s + r\theta)_\theta$ for all natural numbers s and r, it holds that

$$\frac{1}{\theta_0}\sum_{k=1}^{\theta_0} m((k-1)\theta + s_\theta) = \frac{1}{\theta_0}\sum_{k=1}^{\theta_0} m((k-1)\theta + (s+r\theta)_\theta).$$

Moreover, as $\zeta_s = \zeta_{s+r\theta}$ $(= 0)$ by assumption, we obtain that $m(s) = m(s+r\theta)$ for all s and r, which means that m has the period θ. If $\theta < \theta_0$, this contradicts the assumption that θ_0 is the smallest period of m. If $\theta > \theta_0$, we run into the following contradiction: As θ is no multiple of θ_0, it holds that

$$m(s) = m(s+\theta) = m\Big(s + \Big\lfloor\frac{\theta}{\theta_0}\Big\rfloor\theta_0 + k\Big) = m(s+k)$$

for some k with $1 \leq k < \theta_0$. However, $m(s+k) \neq m(s)$ for at least one s, as otherwise $k < \theta_0$ would be a period of m.

It remains to show that $|\zeta_s| \geq \eta$ for some small constant $\eta > 0$ whenever $\zeta_s \neq 0$. To see this, first note that $\frac{1}{\theta_0}\sum_{k=1}^{\theta_0} m((k-1)\theta + s_\theta)$ is the average of θ_0 different elements of the sequence $\{m(t)\}_{t\in\mathbb{Z}}$. The sequence being periodic, this average can only take a finite number of different values. More precisely, there are at most $\binom{2\theta_0-1}{\theta_0}$ different values (independently of s and θ). From this, it immediately follows that $\zeta_s = m(s) - \frac{1}{\theta_0}\sum_{k=1}^{\theta_0} m((k-1)\theta + s_\theta)$ can only take a finite number of values as well. In particular, there is only a finite number of possible non-zero values. We can thus find a constant $\eta > 0$ with $|\zeta_s| \geq \eta$ whenever $\zeta_s \neq 0$. □

Proof of Lemma A4. To start with, we shortly comment on the results for Case B. Note that in this case, we do not only have that $\zeta_s = 0$ but even $\gamma_{s,T} = 0$ for all s. Hence, it holds that $(I - \Pi_\theta)X_{\theta_0}\beta = 0$, which immediately implies that the terms $B_\theta, S_\theta^\varepsilon$ and S_θ^g are all equal to zero.

Let us now turn to Case A: Using Lemma A3 together with (10) and (11), it is easily seen that $\gamma_{s,T} \to \zeta_s \neq 0$ with $|\gamma_{s,T} - \zeta_s| = |R_{s,T}| \leq C\Theta_T/T$ for all $s \in \mathcal{S}$ and $|\gamma_{s,T}| = |R_{s,T}| \leq C\Theta_T/T$ for all $s \in \mathcal{S}^c$. From this, it immediately follows that

$$\begin{aligned} B_\theta &= (X_{\theta_0}\beta)^\intercal (I - \Pi_\theta) X_{\theta_0}\beta \\ &= (X_{\theta_0}\beta)^\intercal (I - \Pi_\theta)^\intercal (I - \Pi_\theta) X_{\theta_0}\beta \\ &= (\gamma_{1,T},\dots,\gamma_{\theta^\times,T},\dots)(\gamma_{1,T},\dots,\gamma_{\theta^\times,T},\dots)^\intercal \geq c\frac{nT}{\theta} \end{aligned}$$

for some fixed constant $c > 0$ and all $T \geq T_0$ with T_0 being sufficiently large. Next write

$$S_\theta^\varepsilon = \sum_{t=1}^{T} \gamma_{t,T}\varepsilon_{t,T} = \sum_{t\in I_{\mathcal{S}}} \gamma_{t,T}\varepsilon_{t,T} + \sum_{t\in I_{\mathcal{S}^c}} \gamma_{t,T}\varepsilon_{t,T}$$

with $I_{\mathcal{S}} = \{t : t - \theta^\times \lfloor t/\theta^\times \rfloor \in \mathcal{S}\}$ and $I_{\mathcal{S}^c} = \{t : t - \theta^\times \lfloor t/\theta^\times \rfloor \in \mathcal{S}^c\}$. Then

$$\begin{aligned}\mathbb{P}\Big(|S_\theta^\varepsilon| > \nu_T\sqrt{\frac{nT}{\theta}}\Big) &\leq \mathbb{P}\Big(\Big|\sum_{t\in I_S} \gamma_{t,T}\varepsilon_{t,T}\Big| > \frac{\nu_T}{2}\sqrt{\frac{nT}{\theta}}\Big) \\ &\quad + \mathbb{P}\Big(\Big|\sum_{t\in I_{S^c}} \gamma_{t,T}\varepsilon_{t,T}\Big| > \frac{\nu_T}{2}\sqrt{\frac{nT}{\theta}}\Big) \\ &=: Q_{\theta,1} + Q_{\theta,2}.\end{aligned}$$

As $|\gamma_{s,T}| \leq C$ for all s and T for some sufficiently large constant C (which is evident from (10) and (11)), we can apply Chebychev's inequality and then exploit the mixing conditions on our model variables with the help of Davydov's inequality to get that $Q_{\theta,1} \leq C/\nu_T^2$. Using the same argument together with the fact that $|\gamma_{s,T}| \leq C\Theta_T/T$ for all $s \in \mathcal{S}^c$, we further obtain that $Q_{\theta,2} \leq C\Theta_T^3/(T\nu_T)^2 \leq C/\nu_T^2$. As a result,

$$\mathbb{P}\Big(|S_\theta^\varepsilon| > \nu_T\sqrt{\frac{nT}{\theta}}\Big) \leq \frac{C}{\nu_T^2}.$$

Finally,

$$\begin{aligned}|S_\theta^g| = |g^\intercal(I - \Pi_\theta)X_{\theta_0}\beta| &= \Big|\sum_{t=1}^{T} \gamma_{t,T} g\Big(\frac{t}{T}\Big)\Big| \\ &= \Big|\sum_{s=1}^{\theta^\times} \gamma_{s,T} K_{s,T}^{[\theta^\times]} \underbrace{\Big(\frac{1}{K_{s,T}^{[\theta^\times]}} \sum_{k=1}^{K_{s,T}^{[\theta^\times]}} g\Big(\frac{s + (k-1)\theta^\times}{T}\Big)\Big)}_{|\cdot| \leq C/K_{s,T}^{[\theta^\times]} \text{ by Lemma A1}}\Big| \leq Cn\end{aligned}$$

for some sufficiently large constant C. This completes the proof. □

Proof of Lemma A5. It holds that

$$(\Pi_\theta - \Pi_{\theta_0})g = \left(\begin{pmatrix} D_\theta & D_\theta & \dots \\ D_\theta & \ddots & \\ \vdots & & \end{pmatrix} - \begin{pmatrix} D_{\theta_0} & D_{\theta_0} & \dots \\ D_{\theta_0} & \ddots & \\ \vdots & & \end{pmatrix}\right) g$$

$$= \begin{pmatrix} \frac{1}{K_{1,T}^{[\theta]}} \sum_{k=1}^{K_{1,T}^{[\theta]}} g\Big(\frac{(k-1)\theta+1}{T}\Big) \\ \vdots \\ \frac{1}{K_{\theta,T}^{[\theta]}} \sum_{k=1}^{K_{\theta,T}^{[\theta]}} g\Big(\frac{(k-1)\theta+\theta}{T}\Big) \\ \frac{1}{K_{1,T}^{[\theta]}} \sum_{k=1}^{K_{1,T}^{[\theta]}} g\Big(\frac{(k-1)\theta+1}{T}\Big) \\ \vdots \end{pmatrix} - \begin{pmatrix} \frac{1}{K_{1,T}^{[\theta_0]}} \sum_{k=1}^{K_{1,T}^{[\theta_0]}} g\Big(\frac{(k-1)\theta_0+1}{T}\Big) \\ \vdots \\ \frac{1}{K_{\theta_0,T}^{[\theta_0]}} \sum_{k=1}^{K_{\theta_0,T}^{[\theta_0]}} g\Big(\frac{(k-1)\theta_0+\theta_0}{T}\Big) \\ \frac{1}{K_{1,T}^{[\theta_0]}} \sum_{k=1}^{K_{1,T}^{[\theta_0]}} g\Big(\frac{(k-1)\theta_0+1}{T}\Big) \\ \vdots \end{pmatrix}.$$

Hence,

$$\begin{aligned} W_\theta^g &= g^\intercal(\Pi_\theta - \Pi_{\theta_0})g \\ &= \sum_{l=1}^{T} \Big(\frac{1}{K_{l_\theta,T}^{[\theta]}} \sum_{k=1}^{K_{l_\theta,T}^{[\theta]}} g\Big(\frac{(k-1)\theta + l_\theta}{T}\Big)\Big) g\Big(\frac{l}{T}\Big) \\ &\quad - \sum_{l=1}^{T} \Big(\frac{1}{K_{l_{\theta_0},T}^{[\theta_0]}} \sum_{k=1}^{K_{l_{\theta_0},T}^{[\theta_0]}} g\Big(\frac{(k-1)\theta_0 + l_{\theta_0}}{T}\Big)\Big) g\Big(\frac{l}{T}\Big) \end{aligned}$$

with $l_\theta = l - \theta\lfloor l/\theta \rfloor$. Moreover,

$$\begin{aligned} &\Big| \sum_{l=1}^{T} \Big(\frac{1}{K_{l_\theta,T}^{[\theta]}} \sum_{k=1}^{K_{l_\theta,T}^{[\theta]}} g\Big(\frac{(k-1)\theta + l_\theta}{T}\Big)\Big) g\Big(\frac{l}{T}\Big) \Big| \\ &= \Big| \sum_{s=1}^{\theta} \frac{1}{K_{s,T}^{[\theta]}} \Big(\underbrace{\sum_{k=1}^{K_{s,T}^{[\theta]}} g\Big(\frac{(k-1)\theta + s}{T}\Big)}_{\leq C \text{ by Lemma A1}} \Big)^2 \Big| \leq \frac{C\theta}{K_{s,T}^{\theta}} \leq \frac{C\Theta_T^2}{T} \end{aligned}$$

and thus $|W_\theta^g| \leq C$. Similarly,

$$\begin{aligned} W_\theta^\varepsilon = \varepsilon^\intercal(\Pi_\theta - \Pi_{\theta_0})g = &\sum_{l=1}^{T} \Big(\frac{1}{K_{l_\theta,T}^{[\theta]}} \sum_{k=1}^{K_{l_\theta,T}^{[\theta]}} g\Big(\frac{(k-1)\theta + l_\theta}{T}\Big)\Big) \varepsilon_{l,T} \\ &- \sum_{l=1}^{T} \Big(\frac{1}{K_{l_{\theta_0},T}^{[\theta_0]}} \sum_{k=1}^{K_{l_{\theta_0},T}^{[\theta_0]}} g\Big(\frac{(k-1)\theta_0 + l_{\theta_0}}{T}\Big)\Big) \varepsilon_{l,T}. \end{aligned}$$

Rewriting W_θ^ε in this way, we can apply Chebychev's inequality and subsequently exploit our mixing assumptions by Davydov's inequality to obtain that

$$\mathbb{P}\big(|W_\theta^\varepsilon| > C\nu_T\big) \leq \frac{C}{\nu_T^2}$$

for any diverging sequence $\{\nu_T\}$. □

Proof of Theorem 2

We first prove the result on asymptotic normality: Let $\tilde{m}$ be the estimator of m in the oracle case where the true period θ_0 is known, i.e. $(\tilde{m}(1), \dots, \tilde{m}(\theta_0)) = \hat{\beta}_{\theta_0}$ and $\tilde{m}(s + k\theta_0) = \tilde{m}(s)$ for all $s = 1, \dots, \theta_0$ and all $k \in \mathbb{N}$. Then we can write

$$\sqrt{T}(\hat{m}(t) - m(t)) = \sqrt{T}(\hat{m}(t) - \tilde{m}(t)) + \sqrt{T}(\tilde{m}(t) - m(t)).$$

For any $\delta > 0$, it holds that

$$\begin{aligned}\mathbb{P}\Big(\big|\sqrt{T}(\hat{m}(t) - \tilde{m}(t))\big| > \delta\Big) \\ \leq \mathbb{P}\Big(\big|\sqrt{T}(\hat{m}(t) - \tilde{m}(t))\big| > \delta, \hat{\theta} = \theta_0\Big) + \mathbb{P}\big(\hat{\theta} \neq \theta_0\big).\end{aligned}$$

The right-hand side of the above inequality is $o(1)$, as the first term is equal to zero (note that $\hat{m}(t) = \tilde{m}(t)$ for $\hat{\theta} = \theta_0$) and $\mathbb{P}(\hat{\theta} \neq \theta_0) = o(1)$ by Theorem 1. Hence,

$$\sqrt{T}(\hat{m}(t) - m(t)) = \sqrt{T}(\tilde{m}(t) - m(t)) + o_p(1).$$

Next, note that we can write

$$\tilde{m}(t) = \frac{1}{K_{t_0,T}} \sum_{k=1}^{K_{t_0,T}} Y_{t_0+(k-1)\theta_0,T}$$

with $t_0 = t - \theta_0 \lfloor t/\theta_0 \rfloor$ and $K_{t_0,T} = 1 + \lfloor (T - t_0)/\theta \rfloor$, i.e. the estimate $\tilde{m}(t)$ can be expressed as the empirical mean of observations that are separated by a multiple of θ_0 periods. This can be seen by inspecting the formula for the least squares estimate $\hat{\beta}_{\theta_0}$. We thus obtain that

$$\begin{aligned}\sqrt{T}(\tilde{m}(t) - m(t)) &= \sqrt{T}\Big(\frac{1}{K_{t_0,T}} \sum_{k=1}^{K_{t_0,T}} g\Big(\frac{t_0 + (k-1)\theta_0}{T}\Big) + \frac{1}{K_{t_0,T}} \sum_{k=1}^{K_{t_0,T}} \varepsilon_{t_0+(k-1)\theta_0,T}\Big) \\ &=: Q_1 + Q_2.\end{aligned}$$

The term Q_1 approximates the integral $\int_0^1 g(u)du$. Using Lemma A1, the convergence rate is seen to be $O(\frac{1}{\sqrt{T}})$. As $\int_0^1 g(u)du = 0$ by our normalization, we obtain that Q_1 is of the order $O(\frac{1}{\sqrt{T}})$ and can thus be asymptotically neglected. Noting that $\{\varepsilon_{t,T}\}$ is mixing by (C1) and has mean zero, we can now apply a central limit theorem for mixing variables to Q_2 to get the normality result of Theorem 2.

We next turn to the uniform convergence result. We have to show that for each $\delta > 0$ there exists a constant C such that

$$\mathbb{P}\Big(\max_{1 \leq t \leq T} \big|\hat{m}(t) - m(t)\big| > \frac{C}{\sqrt{T}}\Big) < \delta \tag{12}$$

for sufficiently large T. This can be seen as follows: For each fixed constant $C > 0$,

$$\begin{aligned}\mathbb{P}\Big(\max_{1 \leq t \leq T} \big|\hat{m}(t) - m(t)\big| > \frac{C}{\sqrt{T}}\Big) \\ \leq \mathbb{P}\Big(\max_{1 \leq t \leq T} \big|\hat{m}(t) - m(t)\big| > \frac{C}{\sqrt{T}}, \hat{\theta} = \theta_0\Big) + \mathbb{P}\big(\hat{\theta} \neq \theta_0\big).\end{aligned}$$

Moreover, $\mathbb{P}(\hat{\theta} \neq \theta_0) = o(1)$ by Theorem 1 and

$$\begin{aligned}
&\mathbb{P}\Big(\max_{1\leq t\leq T} \big|\hat{m}(t) - m(t)\big| > \frac{C}{\sqrt{T}}, \hat{\theta} = \theta_0 \Big) \\
&\qquad = \mathbb{P}\Big(\max_{1\leq t\leq \theta_0} \big|\hat{m}(t) - m(t)\big| > \frac{C}{\sqrt{T}}, \hat{\theta} = \theta_0 \Big) \\
&\qquad \leq \sum_{t=1}^{\theta_0} \mathbb{P}\Big(\big|\hat{m}(t) - m(t)\big| > \frac{C}{\sqrt{T}} \Big).
\end{aligned}$$

By the above arguments for the asymptotic normality result, $\hat{m}(t) - m(t) = O_p(\frac{1}{\sqrt{T}})$ for each fixed time point t. Hence, we can make the probabilities $\mathbb{P}(\big|\hat{m}(t) - m(t)\big| > \frac{C}{\sqrt{T}})$ for $t = 1, \ldots, \theta_0$ arbitrarily small by choosing the constant C large enough. From this, (12) immediately follows. □

Proof of Theorem 3

We start with the proof of the uniform convergence result. Letting $\tilde{g}$ be the infeasible estimator defined in (5), we can write

$$\sup_{u\in[0,1]} \big|\hat{g}(u) - g(u)\big| \leq \sup_{u\in[0,1]} \big|\hat{g}(u) - \tilde{g}(u)\big| + \sup_{u\in[0,1]} \big|\tilde{g}(u) - g(u)\big|.$$

Since $\max_{1\leq t\leq T} |m(t) - \hat{m}(t)| = O_p(1/\sqrt{T})$, it holds that

$$\sup_{u\in[0,1]} \big|\hat{g}(u) - \tilde{g}(u)\big| = \sup_{u\in[0,1]} \Big| \frac{\sum_{t=1}^{T} w_{t,T}(u)(m(t) - \hat{m}(t))}{\sum_{t=1}^{T} w_{t,T}(u)} \Big| = O_p\Big(\frac{1}{\sqrt{T}}\Big).$$

It thus remains to show that

$$\sup_{u\in[0,1]} \big|\tilde{g}(u) - g(u)\big| = O_p\Big(\sqrt{\frac{\log T}{Th}} + h^2\Big).$$

To do so, we decompose the local linear smoother $\tilde{g}$ into the variance component $\tilde{g}^V(u) = \tilde{g}(u) - \mathbb{E}[\tilde{g}(u)]$ and the bias component $\tilde{g}^B(u) = \mathbb{E}[\tilde{g}(u)] - g(u)$. Using a simplified version of the proof for Theorem 4.1 in Vogt (2012) or alternatively applying Theorem 1 of Kristensen (2009), it can be shown that $\sup_{u\in[0,1]} |\tilde{g}^V(u)| = O_p(\sqrt{\log T/Th})$. In addition, straightforward calculations yield that $\sup_{u\in[0,1]} |\tilde{g}^B(u)| \leq Ch^2$ for some sufficiently large constant C. This completes the proof of the uniform convergence result.

The result on asymptotic normality can be derived in an analogous way by first replacing $\hat{g}$ with the smoother $\tilde{g}$ and then using the decomposition $\tilde{g} = \tilde{g}^V + \tilde{g}^B$. Standard arguments show that the bias component $\tilde{g}^B(u)$ has the expansion $\tilde{g}^B(u) = h^2 B_u + o(h^2)$ for any fixed $u \in (0, 1)$. Moreover, applying a central limit theorem for mixing arrays yields that the term $\sqrt{Th}\tilde{g}^V(u)$ is asymptotically normal with mean zero and variance V_u. □

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85** 749-759.

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

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. *Journal of Geophysical Research* **111**.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for high-dimensional data.* Springer.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics* **25** 1-37.

Dahlhaus, R. & Subba Rao, S. (2006). Statistical inference for time-varying ARCH processes. *Annals of Statistics* **34** 1075-1114.

Delworth, T. L. & Mann, M. E. (2000). Observed and simulated multidecadal variability in the northern hemisphere. *Climate Dynamics* **16** 661-676.

Gassiat, E. & Lévy-Leduc, C. (2006). Efficient semiparametric estimation of the periods in a superposition of periodic functions with unknown shape. *Journal of Time Series Analysis* **27** 877-910.

Genton, M. G. & Hall, P. (2007). Statistical inference for evolving periodic functions. *Journal of the Royal Statistical Society B* **69** 643-657.

Hall, P., Reimann, J. & Rice, J. (2000). Nonparametric estimation of a periodic function. *Biometrika* **87** 545-557.

Hall, P. & Yin, J. (2003). Nonparametric methods for deconvolving multiperiodic functions. *Journal of the Royal Statistical Society B* **65** 869-886.

Hall, P. & Li, M. (2006). Using the periodogram to estimate period in nonparametric regression. *Biometrika* **93** 411-424.

Hall, P. (2008). Nonparametric methods for estimating periodic functions, with applications in astronomy. In *COMPSTAT 2008: Proceedings in Computational Statistics* (Ed. P. Brito) 3-18. Physika-Verlag, Heidelberg.

Hannan, E. J. (1957). The variance of the mean of a stationary process. *Journal of the Royal Statistical Society B* **19** 282-285.

Hannan, E. J. & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* **41** 190-195.

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. *Journal of the Statistical Royal Society B* **53** 173-187.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models.* Chapman & Hall, London.

de Jong, R. M. & Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica* **68** 407-423.

Kristensen, D. (2009). Uniform convergence rates of kernel estimators with heterogeneous, dependent data. *Econometric Theory* **25** 1433-1445.

Mazzarella, A. (2007). The 60-year solar modulation of global air temperature: the earths rotation and atmospheric circulation connection. *Theoretical and Applied Climatology* **88** 193-199.

Newey, W. K. & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* **55** 703-708.

Quinn, B. G. & Thomson, P. J. (1991). Estimating the frequency of a periodic function. *Biometrika* **78** 65-74.

Restrepo, A. & Chacón, L. P. (1998). On the period of sums of discrete periodic signals. *IEEE Signal Processing Letters* **5** 164-166.

Rice, J. A. & Rosenblatt, M. (1988). On frequency estimation. *Biometrika* **75** 477-484.

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. In *Statistical Analysis and Forecasting of Economic Structural Change* (Ed. P. Hackl) 253-264. Springer, Berlin.

Schlesinger, M. E. & Ramankutty, N. (1994). An oscillation in the global climate system of period 65–70 years. *Nature* **367** 723-726.

Sun, Y., Hart, J. D. & Genton, M. G. (2012). Nonparametric inference for periodic sequences. *Technometrics* **54** 83-96.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* **58** 267-288.

Vogt, M. (2012). Nonparametric regression for locally stationary time series. Forthcoming in *Annals of Statistics*.

Walker, A. M. (1971). On the estimation of a harmonic component in a time series with stationary independent residuals. *Biometrika* **58** 21-36.

Zhou, Z. & Wu, W. B. (2009). Local linear quantile estimation for nonstationary time series. *Annals of Statistics* **37** 2696-2729.