Transformation to approximate independence for locally stationary Gaussian processes

Joseph Guinness, Michael L. Stein

We provide new approximations for the likelihood of a time series under the locally stationary Gaussian process model. The likelihood approximations are valid even in cases when the evolutionary spectrum is not smooth in the rescaled time domain. We describe a broad class of models for the evolutionary spectrum for which the approximations can be computed particularly efficiently. In developing the approximations, we extend to the locally stationary case the idea that the discrete Fourier transform is a decorrelating transformation for stationary time series. The approximations are applied to fit nonstationary time series models to high frequency temperature data. For these data, we fit evolutionary spectra that are piecewise constant in time and use a genetic algorithm to search for the best partition of the time interval.

Keywords: Locally stationary; periodogram; spectral analysis.

1 Introduction

The class of nonstationary time series models is a broad one. Even if we restrict the class to univariate mean-zero Gaussian time series, we are still left with a class that may be indexed by the set of all positive (or non-negative) definite covariance functions defined on the observation domain. One popular approach to defining an interpretable subclass of nonstationary time series models has been through the use of evolutionary spectra, due to Priestley (1965, 1981). Obtaining rigorous asymptotic results under this model is difficult, and the asymptotic formulation of Dahlhaus (1996) has provided an important way forward in this regard. Specifically, let $\mathbb{T}$ be the unit circle and $\mathbb{N}$ the natural numbers. For $T \in \mathbb{N}$ and A a complex-valued transfer function on $[0,1] \times \mathbb{T}$ satisfying $A(u,-\nu) = A(u,\nu)^*$ for all $(u,\nu) \in [0,1] \times \mathbb{T}$, consider the class of real-valued stochastic processes Z_T on $1,\dots,T$ with

$$Z_T(t) = \int_{\mathbb{T}} A(t/T,\nu) e^{it\nu} \widehat{Z}(d\nu), \tag{1}$$

where $\widehat{Z}$ is a mean-zero complex-valued, orthogonal measure with $E|\widehat{Z}(d\nu)|^2 = d\nu$ and $\widehat{Z}(\nu) = \widehat{Z}(-\nu)^*$ to ensure Z_T is real. For T large, Z_T is close to stationary over time periods much shorter than T and, thus, the sequence of processes Z_T may be called a locally stationary sequence (Dahlhaus, 1996).

Several authors have provided useful methods for estimating the evolutionary spectrum. Much of the work has been based on expressing the evolutionary spectrum in terms of a collection of basis functions. Neumann and von Sachs (1997) use a wavelet basis. Adak (1998) uses a windowed Fourier Basis. Dohono, Mallat and von Sachs (1998) use the cosine packet transform. There are several papers, including Ombao et al. (2001), Ombao et al. (2002), and Guo et al. (2003) that use the smooth, localized complex exponential basis. Dahlhaus proposed estimation via the maximization of approximate Gaussian likelihoods, which depend on a local periodogram (Dahlhaus, 1997) and the preperiodogram (Dahlhaus, 2000) under some parametric model for the evolutionary spectrum. The preperiodogram is an estimate of the evolutionary spectrum introduced by Neumann and von Sachs (1997). Dahlhaus's (2000) approximate likelihood is a function of the preperiodogram and has the attractive property that it reduces to the Whittle likelihood (Whittle, 1962) when the evolutionary spectrum is constant as a function of time, i.e. when the process is stationary.

There may be some benefit, for both intuition and computation, to restricting the form of A. Priestley (1965) described the class of uniformly modulated processes, whose transfer functions may be expressed as

$$A(t/T,\nu) = m(t/T)\mu(\nu)$$

in the locally stationary framework. Here, the variance of the process is allowed to vary with time, but the underlying correlation structure is stationary. For many environmental time series, including the temperature data that we analyze in this paper, the uniformly modulated model is not sufficiently flexible to adequately capture the covariance structure exhibited in the data. We propose that a useful, interpretable, and flexible subclass of models for A is

$$A(t/T, \nu) = \sum_{k=1}^{K} m_k(t/T)\mu_k(\nu). \tag{2}$$

Of course, this subclass may be made quite general if K is allowed to be infinite, but when K is taken to be a small integer, (2) may be considered a low-rank approximation to a particular choice of transfer function coming from a more general class. The model in (2) possesses a close relationship to the approach taken in hidden state modeling, i.e. when the process can be described by a particular state or superposition of several states at any given time. Here, the states are represented by the μ_k's, and the contribution of state k at time t is given by $m_k(t/T)$. We assume A is real and positive, although Proposition 1 below holds for complex A. A phase that is constant in time is not identifiable, but a time-varying phase may be at least partially identifiable. We neglect the possibility of time-varying phase relationships, which is a loss of generality. Other authors have implicitly recognized the difficulty in estimating the phase by focusing on the estimation of the evolutionary spectrum. For example, both of Dahlhaus's (1997, 2000) likelihood approximations depend on A only through $|A|^2$.

The rest of the paper is organized as follows. Sections 2 and 3 develop the quadratic form and the log determinant approximations in the Gaussian likelihood, as well as describe the computational advantages that the model in (2) facilitates. Section 4 contains a numerical experiment demonstrating the accuracy and stability of the approximations. In Sections 5 and 6, we describe the temperature data and the piecewise constant in time model for A. Our optimization procedure, which employs a genetic algorithm to search for a partition of the time domain, is outlined in Section 7. The results of the model fitting for the temperature data are given in Section 8, and finally in Section 9, we discuss the accuracy of the approximations for the models that we fit to the temperature data, and we compare the accuracy to the likelihood approximations proposed by Dahlhaus.

2 Quadratic Form Approximation

The covariance function of the process described in (1) is given by

$$\begin{aligned} K_T(s,t;A) = K_T(s,t) &= \mathrm{cov}\{Z_T(s), Z_T(t)\} \\ &= \int_{\mathbb{T}} A(s/T,\nu)A(t/T,\nu)^* e^{i(s-t)\nu} d\nu. \end{aligned} \tag{3}$$

Defining the $T \times T$ matrix $\{C_T(A)\}_{t,j+1} = \sqrt{2\pi/T}\, A(t/T, \nu_j) e^{i\nu_j t}$, with $\nu_j = 2\pi j/T$ and H denoting the conjugate transpose operation, we might expect the covariance function defined by

$$\begin{aligned} D_T(s,t;A) = D_T(s,t) &= \{C_T(A)C_T(A)^H\}_{s,t} \\ &= \frac{2\pi}{T} \sum_{j=0}^{T-1} A\left(s/T, \nu_j\right) A\left(t/T, \nu_j\right)^* e^{i\nu_j(s-t)} \end{aligned} \tag{4}$$

to approximate (3). We study the convergence of the $T \times T$ matrix $\Delta_T(A) = D_T(\cdot,\cdot)$ to the matrix $\Sigma_T(A) = K_T(\cdot,\cdot)$ with respect to the Frobenius matrix norm, $\|\cdot\|_F$. The following proposition relaxes the results of Dahlhaus in that it does not require any smoothness of A in the rescaled time domain, although it requires uniform smoothness of A in the frequency domain.

Proposition 1: *If, for each u, $A(u,\nu)$ is a complex-valued twice continuously differentiable function on $\mathbb{T}$ with respect to ν, and the second partial derivative of $A(u,\nu)$ with respect to ν is uniformly bounded in u and ν, then*

$$\|\Sigma_T(A) - \Delta_T(A)\|_F = O(1) \text{ as } T \to \infty.$$

A proof is given in Appendix A.1. The assumption that A is real and positive is not needed for Proposition 1. If, in addition, the condition number of $\Sigma_T(A)$ is bounded, the inverse matrices converge as well (see Horn and Johnson, 2006, page 336), and thus the quadratic form term in the Gaussian likelihood can be approximated with

$$\frac{1}{2}\|C_T(A)^{-1}\boldsymbol{Z}_T\|^2, \tag{5}$$

where $\boldsymbol{Z}_T$ is the $T \times 1$ vector of observations. In one sense, the inverse transformation in (5) can be thought of as a generalization of the idea that the discrete Fourier transform (DFT) is a decorrelating transformation for stationary processes. In the stationary case, $A(t/T, \nu)$ is constant as a function of t, so we write $A(t/T, \nu) = S(\nu)$. A simple calculation confirms that $C_T(S)^{-1} = (2\pi)^{-1}C_T(1/S)^H$ (where $(1/S)(\nu) \equiv 1/S(\nu)$), so the inverse transformation is given by

$$\left[C_T(S)^{-1}\mathbf{Z}_T\right]_j = \frac{1}{S(\nu_j)}\frac{1}{\sqrt{2\pi T}}\sum_{t=0}^{T-1} e^{-i\nu_j t}Z_T(t),$$

which is simply the DFT of the observations scaled by the square root of the spectral density, so in the stationary case our transformation reduces to the transformation given in some versions of the Whittle likelihood.

There are several ways to compute the inverse matrix-vector product in (5). One could proceed by directly constructing and inverting the matrix $C_T(A)$, but this is clearly intractable for large T, when even storing a $T \times T$ matrix, let alone computing its inverse, may not be possible. We propose that the most attractive method is to arrive at the inverse matrix-vector product using iterative methods. Convergence within a small number of iterations depends on our ability to find a good preconditioning matrix M, for which $MC_T(A)$ has a small condition number relative to that of $C_T(A)$. We have found that preconditioning with $M = C_T(1/A)^H$ works quite well in all of our examples. The speed of each iteration depends on how fast $C_T(A)\boldsymbol{w}$ and $M\boldsymbol{w}$ can be computed for any vector $\boldsymbol{w}$. The model in (2) is well-suited to facilitate efficient computation of $C_T(A)\boldsymbol{w}$. If $\boldsymbol{w} = (w_0, \ldots, w_{T-1})'$ is a $T \times 1$ vector, then

$$\begin{aligned}
\left[C_T(A)\boldsymbol{w}\right]_t &= \sqrt{\frac{2\pi}{T}}\sum_{j=0}^{T-1}\left[\sum_{k=1}^{K} m_k(t/T)\mu_k(\nu_j)\right]e^{i\nu_j t}w_j \\
&= \sqrt{\frac{2\pi}{T}}\sum_{k=1}^{K} m_k(t/T)\sum_{j=0}^{T-1}\left[\mu_k(\nu_j)w_j\right]e^{i\nu_j t}.
\end{aligned}$$

Therefore, the matrix-vector product can be computed as a weighted sum of K inverse DFTs, each of which can be computed efficiently using the Fast Fourier Transform (FFT). Furthermore, there is no need to store the matrix $C_T(A)$. In Section 6, we explore a specific model for A for which the preconditioning multiplication can be computed using the FFT and also does not require $O(T^2)$ storage.

Although this estimate computes the exact inverse matrix-vector product, still present are the usual approximations that occur in the spectral analysis of stationary time series. Specifically, approximation (4) effectively assumes that the process is periodic with period T, so when the beginning and the end of the series have different properties, the quadratic form estimate tends to have larger errors. This is true in both the stationary and nonstationary cases. We do not expect our approximation to solve this problem, but it should perform no worse in the nonstationary case than the Whittle likelihood performs in the stationary case.

3 Log Determinant Approximation

Dahlhaus (2000, Proposition 2.5) shows that under suitable regularity conditions (including an assumption about some third order mixed partial derivatives of A), for any $\epsilon > 0$,

$$\log\det\Sigma_T(A) = \frac{T}{2\pi}\int_0^1\int_{-\pi}^{\pi}\log\{2\pi|A(u,\nu)|^2\}d\nu\,du + O(T^{\epsilon}). \tag{6}$$

Let us instead start with the closely related approximation,

$$\log|\det\{C_T(A)/\sqrt{2\pi}\}| \approx \frac{1}{T}\sum_{t,j=1}^{T}\log\{b_{tj}\}, \tag{7}$$

where $b_{tj} = A\left(t/T, \nu_{j-1}\right)$ is the tj'th entry defining the matrix $B_T(A)$. This approximation has the virtues of being directly applicable for numerical computation and of being exact when A has the form $A(u,\nu) = m(u)\mu(\nu)$. Figures 1 and 2 give some numerical results illustrating typical errors of the first order approximation given in (7) for four choices of A.

To study the approximation (7), it is helpful to consider the decomposition $A(u,\nu) = m(u)\mu(\nu)R(u,\nu)$, where m, μ and R are all positive-valued and bounded away from 0. This decomposition is not unique, and it should be chosen, roughly speaking, to minimize discontinuities in R. Suppose the following conditions on R hold:

(I) For every u, R is absolutely continuous in ν with a uniformly bounded almost everywhere derivative.

(II) Except for possibly some finite number of values of u (not depending on ν), for every ν, R is absolutely continuous in u with a uniformly bounded almost everywhere derivative. At these exceptional u values, R may not be continuous in u, but for definiteness, we will assume R is right continuous in u for all ν.

Then our numerical results suggest

$$\log|\det\{C_T(A)/\sqrt{2\pi}\}| = \frac{1}{T}\sum_{j,t=1}^{T}\log\{b_{tj}\} + O(1) \text{ as } T\to\infty. \tag{8}$$

That such a result might hold under sufficient smoothness conditions is perhaps not surprising in light of (6), which would give an $O(1)$ error for the log determinant if ϵ could be set to 0. However, the conjectures that no smoothness assumptions are needed on m and μ and that R can have discontinuities in u might be somewhat unexpected but is supported by the numerical results in Figures 1 and 2 and in other examples we have tried. Suppose the following stronger condition holds:

(III) R is absolutely continuous in both u and ν with uniformly bounded almost everywhere derivative in both arguments, and $R(0,\nu) = R(1,\nu)$ for all ν.

Then our numerical results suggest

$$\log|\det\{C_T(A)/\sqrt{2\pi}\}| = \frac{1}{T}\sum_{j,t}\log\{b_{tj}\} + O(T^{-1}) \text{ as } T\to\infty. \tag{9}$$

The assumption that $R(0,\nu) = R(1,\nu)$ for all ν may sometimes be plausible in cases for which there is a substantial periodic component to the nonstationarity such as for the high frequency temperature data considered in Section 5. In this example, the data encompass temperatures from a 24-hour period, so we might expect the beginning and the end of the series to have similar properties.

In a wide variety of cases, the following refinement of (7) works quite well: defining $b_{T+1,j} = b_{1j}$,

$$\beta_{tT}(A) = \frac{1}{T}\sum_{j}\log\frac{b_{t+1,j}}{b_{tj}}$$

and

$$\sigma_T(A) = \frac{1}{T}\sum_{j,t}\left\{\log\frac{b_{t+1,j}}{b_{tj}} - \beta_{tT}(A)\right\}^2,$$

then approximate $\log|\det\{C_T(A)/\sqrt{2\pi}\}|$ by

$$\log\widehat{\det}_T(A,\gamma) = \frac{1}{T}\sum_{j,t}\log\{b_{tj}\} + \gamma\sigma_T(A), \tag{10}$$

where γ appears to vary modestly with A with values between 0.25 and 0.35 working well. Note that if the approximation in (10) had an error of $o(T^{-1})$ (with γ possibly depending on A), then it would explain both (8) and (9), since under the conditions (I) and (II), Proposition 3 in Appendix A.3 shows that $\sigma_T(A) = O(1)$ and under the stronger condition (III), $\sigma_T(A) = O(T^{-1})$. When R is identically 1, then $\sigma_T(A) = 0$ and both (7) and (10) are exact. If R is identically 1, and m is constant, then (10) is equivalent to the approximation given in some versions of the Whittle likelihood. Another feature of the approximation (10) (and of (7) for that matter) is that it is invariant to moving the last column of $B_T(A)$ to the first column, an invariance that can be shown to hold also for the exact determinant. Proposition 2 in Appendix A.2 gives a theoretical justification for approximation (10) with $\gamma = 0.25$ in a limited setting.

The errors from approximations (7) and (10) are modest in all of the examples shown in Figures 1 and 2 and all others we have considered, ranging from about 0.001 to 1 and either decreasing with T or staying fairly constant as T varies. The corrected approximation (10) with $\gamma = 0.25$ always does better than (7) and sometimes much better. For $j = 1, 2$, $A_j(1, \nu)/A_j(0, \nu)$ is not constant in ν, so we only expect (8) and not (9) to hold for these functions. The function A_1 has another discontinuity at $t = 0.5$: $A_1(0.5^+, \nu)/A_1(0.5^-, \nu)$ is not constant in ν, but the numerical evidence suggests (8) still holds in this case and the refined approximation $\log \widehat{\det}_T(A, 0.25)$ has error more than an order of magnitude smaller than (7) for all T considered. Figure 2 shows two functions that are absolutely continuous in both arguments and satisfy $A(1, \nu)/A(0, \nu)$ constant; we see that (7) provides excellent approximations for these functions and $\log \widehat{\det}_T(A, 0.25)$ is much better still. Figure 3 plots the ratio of the error of (7) to $0.25\sigma_T(A)$ for the four functions considered in Figures 1 and 2 and it appears that the values are converging as T increases but to somewhat different values for the different functions. Because even the approximation (7) is very accurate when A is smooth and $A(1, \nu)/A(0, \nu)$ is constant, it makes more sense to set γ to a value that works well when these conditions do not hold (e.g., the functions A_1 and A_2 in Figure 1), so we recommend setting $\gamma = 0.25$.

These approximations do not and cannot be expected to work well for any function R. In particular, discontinuities in ν for R cause serious problems for the approximations, although numerical experiments suggest that the errors can grow more slowly than T for such functions, so that the approximations may still be of some use when there are such discontinuities. We consider discontinuities in frequency for this interaction term R (discontinuities in the "main effect" for frequency can be captured in the function μ and do not cause any problem) to be unlikely in practice, at least for natural processes.

4 A simple numerical experiment

To compare our approximations with those proposed by Dahlhaus (1997, 2000), we present two simple examples in which A depends on a single parameter, and we compute the expectations of the exact loglikelihood and each of the approximations over a range of parameter values. The purpose of this numerical experiment is to show an example in which not only are our approximations sharper than Dahlhaus's, but the errors are more stable with respect to a changing parameter, which is a desirable property when maximizing the likelihood with respect to unknown parameters, for which only differences in loglikelihood matter.

The exact negative loglikelihood for the locally stationary Gaussian time series is (minus a constant $T/2 \log(2\pi)$)

$$L(\theta) = \frac{1}{2} \log \det \Sigma_T(A_\theta) + \frac{1}{2} \boldsymbol{Z}_T' \Sigma_T(A_\theta)^{-1} \boldsymbol{Z}_T.$$

For evolutionary spectra that are piecewise constant on K blocks in time, i.e. $A_\theta(u, \nu) = \mu_{k,\theta}(\nu)$ if u is in block k, as will be the case in the examples to follow, we interpret the loglikelihood approximation in Dahlhaus (1997) to be

$$L_a(\theta) = \log \widehat{\det}_T(A_\theta, 0) + \frac{1}{4\pi} \sum_{k=1}^{K} |B_k| \int_{\mathbb{T}} \frac{J_k(\nu)}{|\mu_{k,\theta}(\nu)|^2} d\nu,$$

where B_k denotes the set of times in block k, $|B_k|$ is the length of block k, and J_k is the ordinary periodogram over the kth block of data. If we let K_T denote the covariance function of the time series with parameter θ equal to the true value θ_0, the periodograms have expectation

$$E(J_k(\nu)) = \frac{1}{2\pi|B_k|} \sum_{s,t \in B_k} K_T(s,t) e^{i\nu(s-t)}.$$

The loglikelihood approximation in Dahlhaus (2000) is

$$L_b(\theta) = \log \widehat{\det}_T(A_\theta, 0) + \frac{1}{4\pi} \sum_{t=1}^{T} \int_{\mathbb{T}} \frac{I_T(t/T, \nu)}{|A_\theta(t/T, \nu)|^2} d\nu,$$

where $I_T(t/T, \nu)$ is the preperiodogram, which has expectation

$$E(I_T(t/T, \nu)) = \frac{1}{2\pi} \sum_{1 \le \lfloor t+1/2 \pm k/2 \rfloor \le T} K_T(\lfloor t+1/2-k/2 \rfloor, \lfloor t+1/2+k/2 \rfloor) e^{-i\nu k},$$

where $\lfloor x \rfloor$ is the integer part of x. Discretizing the integrals in L_a and L_b into $2T$ terms gives sufficient accuracy in this example. Our likelihood approximation is

$$L_c(\theta) = \log \widehat{\det}_T(A_\theta, 0.25) + \frac{1}{2} \| C_T(A_\theta)^{-1} \boldsymbol{Z}_T \|^2,$$

whose quadratic form has expectation

$$\frac{1}{2} \mathrm{tr}(C_T(A_\theta)^{-1} \Sigma_T(A_{\theta_0}) C_T(A_\theta)^{-H}).$$

In this numerical experiment, we compare $E(L_{approx}(\theta))$ to $E(L(\theta))$ over a range of values for θ and for a few choices of θ_0. Defining $\mu_\theta(\nu) = 1/\sqrt{2\pi} \exp(\theta \cos(\nu))$, the two models we consider for this experiment are

$$A_\theta^{(1)}(t/T, \nu) = \begin{cases} 1/\sqrt{2\pi}, & t/T \in [0, 0.5] \\ \mu_\theta(\nu), & t/T \in (0.5, 1] \end{cases},$$

$$A_\theta^{(2)}(t/T, \nu) = \begin{cases} 1/\sqrt{2\pi}, & t/T \in [0, 0.25] \cup (0.75, 1] \\ \mu_\theta(\nu), & t/T \in (0.25, 0.75] \end{cases},$$

so that the dynamic range of the spectrum over one segment increases with θ. We take $T = 120$, and we plot in Figure 4 the covariance function corresponding to $A_\theta^{(1)}$, noting that when $\theta \neq 0$, there is nonzero dependence across blocks. Furthermore, the various approximations differ when $\theta \neq 0$ because the process is nonstationary in those cases. Otherwise, when $\theta = 0$ all of the approximations are equal to the Whittle likelihood approximation, which happens to be exact when $\theta = 0$ because the process is stationary white noise in that case.

In Figure 5, we plot the expected values of the various likelihood approximations over a range of values of θ for three different values of the true parameter, $\theta_0 = 0, 1, 2$. Under both models $A_\theta^{(1)}$ and $A_\theta^{(2)}$, all three approximations have a minimum at $\theta = 0$ when $\theta_0 = 0$ and closely track the exact loglikelihood over the range $\theta = -0.5$ to 1. This is not surprising because when θ is small, the model is close to stationary, and the three approximations are nearly the same. When $\theta_0 = 1$, all three approximations perform well under both models, but our new approximation appears to improve on the existing approximations when θ becomes large. Finally, when $\theta_0 = 2$, our approximation continues to track the exact loglikelihood, while the existing approximations fail. In these examples, when the likelihood approximations are not accurate, usually the error in approximating the quadratic form dominates. Under model $A_\theta^{(2)}$ our new approximation is nearly exact; the maximum absolute difference from the exact loglikelihood is less than 0.01 for all values of θ and θ_0 that we studied, and sometimes much smaller. We expect our approximations to perform well when the model has very short-range dependence at the beginning and the end of the series, as most of the error associated with replacing $\Sigma_T(A)$ with $\Delta_T(A)$ occurs in the bottom left and upper right corners of the matrix (see the proof of Proposition 1). However, the approximation does worsen with larger values of θ.

Table 1: Values of θ that minimize the expected value of the exact loglikelihood and each of the expected loglikelihood approximations when the true value is θ_0. "curv" gives the value of the second derivative at the minimum, as approximated by finite differences.

		$\theta_0 = 0.0$		$\theta_0 = 0.5$		$\theta_0 = 1.0$		$\theta_0 = 1.5$		$\theta_0 = 2.0$	
		min	curv	min	curv	min	curv	min	curv	min	curv
$A_\theta^{(1)}$	$E(L(\theta))$	0.000	59.25	0.500	59.28	1.000	59.40	1.500	59.66	2.000	60.24
	$E(L_a(\theta))$	0.000	60.00	0.489	60.58	0.958	63.79	1.344	77.91	1.553	130.09
	$E(L_b(\theta))$	0.000	60.00	0.489	60.63	0.955	64.22	1.331	79.99	1.529	136.11
	$E(L_c(\theta))$	0.000	59.50	0.498	59.60	0.995	59.96	1.489	60.74	1.979	62.98
$A_\theta^{(2)}$	$E(L(\theta))$	0.000	59.50	0.500	59.57	1.000	59.80	1.500	60.32	2.000	61.48
	$E(L_a(\theta))$	0.000	60.00	0.489	60.58	0.958	63.79	1.344	77.91	1.553	130.09
	$E(L_b(\theta))$	0.000	60.00	0.495	60.24	0.977	61.98	1.389	72.23	1.608	118.40
	$E(L_c(\theta))$	0.000	59.50	0.500	59.56	1.000	59.80	1.500	60.32	2.000	61.44

In Table 1 we display the value of θ that minimizes each of the expected likelihood approximations for five values of θ_0, as well as the second derivative of each expected approximation at the minimum (approximated by finite differences). For both models $A_\theta^{(1)}$ and $A_\theta^{(2)}$, the existing approximations begin to fail to as θ_0 increases, in that they underestimate θ_0 and overestimate the curvature at the minimum. Relative to the existing approximations, the new approximation is stable with respect to increasing θ_0, and under model $A_\theta^{(2)}$, the new approximation is nearly exact although it does worsen slightly as θ_0 increases. In this example, these results suggest that the new approximation is more stable than the existing approximations, and the new approximation is robust with respect to a model with a large dynamic range on the interior of the time domain, as is the case with $A_\theta^{(2)}$. We note that Dahlhaus's (2000) approximation was constructed for a smoothly-varying A, whereas the examples here have jumps. Nevertheless, it is interesting to note that this approximation performs similarly to the blockwise periodogram approximation.

5 Description of the data

We implement the techniques introduced in this paper to analyze a set of high-frequency temperature data. Specifically, we will consider two sets of 24 hours of temperature data from the Southern Great Plains region of the Atmospheric Radiation Measurement (ARM) program, which is recorded at regular one-minute intervals. All of the data can be accessed via the web at http://www.archive.arm.gov. Figure 6 below plots 24 hours ($T = 1440$) of temperature data from the first two days in October 2005 recorded at monitoring site EF-03, as well as the first differences of those data, which will be the focus of our analysis. We refer to the second day as the "normal" day because the temperature is low in the morning, warms up during the day and cools back down in the evening, and we refer to the first day as the "unusual" day due to its irregular temperature pattern.

The first differences, if considered to be a realization of a Gaussian process, show a high degree of nonstationarity. To see this more clearly, we can compute and plot local periodograms of the data (Figure 7). A local periodogram is simply the periodogram of a window of data around the time of interest. Here we use 60 minute windows. The vertical axis refers to frequency, and lighter colors on the image refer to higher power. The time series is plotted on top of the local periodogram. It is evident that not only does the variance of the processes change over time, but the shape of the spectrum changes as well. Dahlhaus (1997) uses a tapered version of a local periodogram in his approximation of the quadratic form term in the Gaussian likelihood.

The local periodogram seems to be a good diagnostic and exploratory tool for finding nonstationarities, but as

an estimate of the evolutionary spectrum, it is highly window-dependent, which is problematic when the spectral properties of the process undergo sharp changes in time. Neumann and von Sachs (1997) describe the preperiodogram, which is not window-dependent, as a starting point for their evolutionary spectrum estimator. The preperiodogram is the DFT of a local covariance estimate and is similar to the Wigner-Ville spectrum. Dahlhaus (2000) uses the preperiodogram instead of a local periodogram in his likelihood approximation.

6 Modeling

A simple nonstationary model for the first difference temperature process is that its evolutionary spectrum is piecewise constant in time. In this setting, the rescaled time interval is partitioned into a number of blocks, and a single function of frequency describes the transfer function within each block. This model is a special case of the model in (2). Indeed, if we have K blocks in the partition, and $I_k(t/T) = 1$ if t/T is a member of block k and zero otherwise, then

$$A(t/T, \nu) = \sum_{k=1}^{K} I_k(t/T)\mu_k(\nu)$$

is piecewise constant in time. In this case, I_k contains the information about the partition, and μ_k describes the transfer function within block k.

Furthermore, for a $T \times 1$ vector $\boldsymbol{u} = (u_1, \ldots, u_T)$ the preconditioning transformation is given by

$$\begin{aligned}\left[C_T(1/A)^H \boldsymbol{u}\right]_j &= \sum_{t=0}^{T-1}\sum_{k=1}^{K} \frac{I_k(t/T)}{\mu_k(\nu_j)} u_t e^{-i\nu_j t} \\ &= \sum_{k=1}^{K} \frac{1}{\mu_k(\nu_j)} \sum_{t=0}^{T-1} [I_k(t/T)u_t] e^{-i\nu_j t}.\end{aligned}$$

This can also be computed with K FFTs and does not require $O(T^2)$ storage. Not only is this model advantageous for the computation of the quadratic form, there is a simple and easily computed expression for the first order log determinant approximation and the second order correction. Specifically, $\log(b_{t+1,j}/b_{tj})$ is zero for all t except at the breakpoints and $t = T$, so we can ignore all other terms in $\sigma_T(A)$. Therefore, the log determinant approximation requires $O(TK)$ computations rather than $O(T^2)$ computations.

We will see that the piecewise constant spectrum assumption is reasonable for some periods, where it seems that the properties of the process undergo sharp changes at several time points. For other time periods, however, the properties of the process appear to change slowly and continuously over a long period of time. In those cases, a transfer function that evolves continuously in time may be more appropriate. The model in (2) includes such functions, and we describe one special case in Section 8.

We parametrize the log of the μ_k's as trigonometric polynomials,

$$\log \mu_k(\nu) = \sum_{j=0}^{N-1} c_{kj} \cos(j\nu) \tag{11}$$

with $N = 5$. As an alternative, one could parametrize with transfer functions that correspond to autoregressive processes, which would provide a way to conduct piecewise AR modeling without having to require independence among blocks, as is assumed in Davis et al. (2006).

7 Optimization algorithm

The optimization problem involves choosing the partition and set of c_{kj}'s to maximize the approximate likelihood. It is helpful to think of the problem as a continuous optimization (choosing the c_{kj}'s) nested inside of a discrete

optimization (choosing the partition). We perform the optimization by searching the space of partitions, and for each partition, we use continuous optimization techniques to choose the c_{kj}'s that maximize or approximately maximize the likelihood for that partition.

In this paper, we follow the example of Davis et al. (2006), who use a genetic algorithm to search for the breakpoints in piecewise stationary AR models. In their approach, the minimum description length is the criterion function used to select both the number of blocks in the partition and the breakpoints of the partition. We fix the number of blocks, and the criterion function we wish to maximize is the approximate likelihood described herein.

The canonical form of the genetic algorithm starts with an initial population of individuals, partitions in this case, and generates a new generation by crossing or mutating individuals from the current generation. The crossing operation involves selecting two parent individuals from the current generation and combining their attributes to bear a child with similar properties to its parents. The mutation operation involves selecting a single parent from the current generation and randomly perturbing some of its attributes to bear a child with similar properties to its parent. In both operations parents are randomly chosen from the current population, and parents with high values of the criterion function (approximate likelihood in our case) are more likely to be chosen than parents with low values of the criterion function. Once the population of the new generation is large enough, that generation becomes the parent population for the next generation. New generations are created in this way until some stopping criterion is met.

Each implementation of a genetic algorithm must therefore define the crossing and mutation operations. In our case, we can describe a partition with K blocks by its $K-1$ breakpoints. In the crossing operation, the child is generated by uniformly sampling from the collection of its parents' breakpoints, subject to the constraint that there must be a minimum distance between breakpoints, 30 minutes in this application. The constraint ensures that if both parents have a breakpoint at roughly the same location, at most one of those two breakpoints can appear in the child's partition. In the mutation operation, the child is generated by randomly moving one or several of its parent's breakpoints, subject to the same minimum distance constraint.

Our algorithm has population sizes of 40 partitions, and each child, independently of all the other children, is generated by a crossing operation with probability 0.9 and by a mutation operation with probability 0.1. The initial population of partitions is generated by sampling uniformly over all partitions of the time domain with the fixed number of blocks, subject to the minimum block size constraint described earlier. The algorithm is stopped after 1000 generations.

Of course, this is only half of the optimization procedure. We must still choose the parameters defining the transfer functions within each block to maximize the likelihood. For this parameterization, there is no simple form for obtaining the maximum likelihood parameter estimates, so we use an iterative optimization technique such as conjugate gradient. These techniques are quite slow compared to the time it takes to generate a new generation of partitions. We are helped by the fact that the gradient of the likelihood with respect to the parameters can be computed fast using techniques similar to those described for computing the likelihood and by the fact that the Hessian is relatively sparse; the mixed partial derivatives with respect to parameters in different blocks are zero. However, the time required is large enough to prohibit maximizing the likelihood for each partition in each generation.

To remedy this problem, we proceed as follows. In the initial generation of partitions, we choose the parameters for each partition by maximizing the approximate likelihood, and we store the partitions, along with the transfer function parameters, in a table. We then use this table to assign parameters to newly generated partitions in future generations. When a new partition is created, for each of its blocks, we find the partition in the table that has the most similar block, take the parameters describing the transfer function for that block, and assign them to the block of the current partition. This allows us to pick and choose parameters from several different partitions in the table and assign them to the current partition.

Every 20 generations, we choose the parameters again by maximizing the approximate likelihood, and we add the partitions from that generation, along with their parameters, to the table. Thus, as we move along in the

algorithm, the table is filled in more densely, and we are more likely to find partitions in the table with blocks closely matching the blocks of the current partition.

8 Optimization results

To choose an appropriate number of blocks for the partition, we ran our algorithm separately for varying numbers of blocks. For each number of blocks, we repeated the optimization 10 times and recorded the estimate of A giving the highest approximate likelihood, which means recording the partition and associated parameters. The analysis was repeated on the data from October 1 and 2, 2005. In Figures 8 and 9, we show the best partitions from each of the 10 runs for each number of blocks. Each best partition is represented as a row of circles plotted at the maximum loglikelihood for that partition. This allows us to see if the algorithm provided stable estimators of the partition and how much the likelihood increased when we added blocks to the partition.

For the "unusual" day, it seems that the algorithm returned a stable estimate of the partition for as many as five blocks, but when we increased the number of blocks, there tended to be some disagreement among the different runs, highlighting the inherent difficulty of searching a partition space with more than a few blocks. For the "normal" day, the estimates were not quite as stable, perhaps suggesting that a piecewise constant in time spectrum is not appropriate. As expected, we saw large increases in likelihood between the three- and four-block models but small differences in likelihood between the six- and seven-block models.

One may ask whether a model with a separate transfer function in each block like we have just considered gives a significantly better fit than a model with a single transfer function that is modulated by a different factor in each block. Recall that the models for A fit here contain K blocks corresponding to K different spectra:

$$A(t/T, \nu) = \sum_{k=1}^{K} I_k(t/T)\mu_k(\nu).$$

We refer to this as the full model. We can compare likelihoods and fitted values to those obtained from a uniformly modulated model for A, which we refer to as the multiplicative model:

$$A(t/T, \nu) = \sum_{k=1}^{K} c_k I_k(t/T)\mu(\nu).$$

This is a special case of the uniformly modulated model, in which the modulating function,

$$m(t/T) = \sum_{k=1}^{K} c_k I_k(t/T),$$

is piecewise constant in time. To compare the fits obtained for these two models, we plot the maximum likelihood estimates for A for both models, as well as the ratio of the two estimates of A in Figures 10 and 11. Here we consider only the best five-block partition found for the full model. We found that for the "unusual" day, the full model increased the loglikelihood by 39.00 over the uniformly modulated model, and for the "normal" day, we achieved an increase of 88.25 loglikelihood units.

As discussed earlier, for the "normal" day, the properties of the process appear to change slowly over several hours, rather than undergo sharp jumps. Perhaps a model that allows the spectrum to change continuously in time may be more appropriate. Consider the following model:

$$A(t/T, \nu) = \sum_{k=1}^{K} I_k(t/T) \exp\left[\alpha_k(t - t_k)/T\right] \mu_k(\nu), \tag{12}$$

where t_k/T is the time at which block k starts. This is again a special case of the model in (2). Our maximum likelihood estimate of A under the model in (12) with $K = 5$ blocks increased the loglikelihood by 27.37 over the full model. The estimate is plotted in Figure 12.

9 Numerical study of approximations

The time series considered here are intentionally chosen to be short enough that an exact Gaussian likelihood can be computed but long enough that our approximation gives a significant speed-up in computation. It would be too time-consuming to compute and maximize exact Gaussian likelihoods within the genetic algorithm, but after the partition is chosen and fixed, maximizing the exact Gaussian likelihood once with respect to the transfer function parameters can be completed within a few minutes on a personal computer. The exact negative loglikelihood is (minus a constant $T/2\log(2\pi)$)

$$L(\theta) = \frac{1}{2}\log\det\Sigma_T(A_\theta) + \frac{1}{2}\boldsymbol{Z}_T'\Sigma_T(A_\theta)^{-1}\boldsymbol{Z}_T.$$

The details of the computation of the exact likelihood under the model described by (1) are straightforward and left to the Appendix A.4. Again, as in Section 4, here are the approximations we consider:

$$L_a(\theta) = \log\widehat{\det}_T(A_\theta, 0) + \frac{1}{4\pi}\sum_{k=1}^{K}|B_k|\int_{\mathbb{T}}\frac{J_k(\nu)}{|\mu_k(\nu)|^2}d\nu,$$

$$L_b(\theta) = \log\widehat{\det}_T(A_\theta, 0) + \frac{1}{4\pi}\sum_{t=1}^{T}\int_{\mathbb{T}}\frac{I_T(t/T,\nu)}{|A_\theta(t/T,\nu)|^2}d\nu,$$

$$L_c(\theta) = \log\widehat{\det}_T(A_\theta, 0.25) + \frac{1}{2}\|C_T(A_\theta)^{-1}\boldsymbol{Z}_T\|^2,$$

$$\widehat{\theta} = \arg\min_\theta L(\theta), \quad \widehat{\theta}_a = \arg\min_\theta L_a(\theta), \quad \widehat{\theta}_b = \arg\min_\theta L_b(\theta), \quad \widehat{\theta}_c = \arg\min_\theta L_c(\theta).$$

We take the partitions of the time interval to be known, so in the following, θ is the parameter vector of coefficients c_{kj} in (11). We evaluated our approximations in two ways. The first is how well we approximated the exact log determinant and quadratic form at $\widehat{\theta}_c$ for each number of blocks. We compared our second order log determinant approximation to the first order log determinant approximation, which is essentially what Dahlhaus (1997, 2000) proposes, only the integral is replaced with a sum. We also compared our quadratic form approximation to that proposed in Dahlhaus (2000), which is based on the preperiodogram, and that proposed in Dahlhaus (1997), which in this application we take to be a sum of integrals of ordinary blockwise periodograms. The second evaluation we considered is how close the parameters chosen by maximizing the approximate likelihoods came to maximizing the exact likelihood. Specifically, we computed $L(\widehat{\theta}) - L(\widehat{\theta}_{approx})$ for each of the three approximations. This evaluates each approximation based on its ability to return an estimate that nearly maximizes the exact likelihood.

Figures 13 and 14 show that for both days, while the two approximations of the log determinant are equivalent in the stationary (one block) case, the second order approximation always gave an improvement over the first order approximation when there was more than one block in the partition. Furthermore, the second order approximation appears to retain a roughly equal level of accuracy regardless of the number of blocks, which is as good of a result as one could hope for with this approximation. Specifically, we would not expect the approximation to be more accurate in the nonstationary case than it is in the stationary case.

It appears in this example that our quadratic form approximation is more stable than those proposed by Dahlhaus. For the "normal" day the preperiodogram and blockwise periodogram estimates are inaccurate in the models with five or more blocks. In these cases, the issue is that the preperiodogram and the blockwise periodogram estimates find some low frequency information in the block that encompasses data roughly between hours 15 and 16, which is a similar situation to that encountered in the numerical experiment in Section 4. Tapering the observations within each block helps in these cases, but our approximation is still superior in the circumstances we have studied. The third plot in Figures 13 and 14 shows that for these data, the error in the likelihood approximation is dominated by the quadratic form.

Figure 15 shows that in all of our examples, maximizing our approximate likelihood always came closer to maximizing the exact likelihood, although the Dahlhaus (1997) approximation performs well on the "unusual" day.

For the "normal" day, we do not include the estimates for more than four blocks when using Dahlhaus's likelihoods because the quadratic form problem produced unreasonable estimates of A.

10 Concluding remarks

We have provided a numerical example and an application with temperature data that suggest that our Gaussian likelihood approximations are sharper and more stable than those proposed by Dahlhaus. Although we have no theorems proving that our approximations are uniformly more accurate, they have performed better in the specific examples we have studied. If the evolutionary transfer function can be written in the form in (2), the approximations can be computed efficiently. Otherwise, the approximations–particularly the quadratic form–are computationally intensive for long time series. However, the model in (2) is easily adaptable to diverse applications by adjusting the form of the component transfer functions, μ_k, and their respective modulating functions, m_k. The log determinant approximation may require alterations to facilitate fast computation in cases when the spectrum is not piecewise constant in time. For example, when we approximate the log determinant term for the model in (12), we can ignore all the terms in $\sigma_T(A)$ except those at the breakpoints and end of the series. This should provide reasonable approximations when the spectrum does not vary too quickly in time between the breakpoints. The quadratic form term does not require such an adjustment. However, more effort may be required to obtain a suitable preconditioner for the iterative algorithm, which should not affect the accuracy of the quadratic form (as long as the iterative solver converges) but will affect the speed of convergence.

Acknowledgements

This work was initially supported by the United States Department of Energy, Office of Science, Office of Biological and Environmental Research, Climate Change Research Division, under contract DE-AC02-06CH11357, as a part of the SciDAC program, and received subsequent support from US Department of Energy Grant DE-SC0002557.

Appendix

A.1

In this section, we prove Proposition 1:

Proposition 1: *If, for each u, $A(u,\nu)$ is a complex-valued twice continuously differentiable function on $\mathbb{T}$ with respect to ν, and the second partial derivative of $A(u,\nu)$ with respect to ν is uniformly bounded in u and ν, then*

$$\|\Sigma_T(A) - \Delta_T(A)\|_F = O(1) \text{ as } T \to \infty.$$

Proof: We define $f_{u,v}(\nu) = A(u,\nu)A^*(v,\nu)$. Because $A(u,\nu)$ is twice continuously differentiable with respect to ν for each u, $f_{u,v}$ is also twice continuously differentiable on $\mathbb{T}$ for every u and v. Therefore, the Fourier series of $f_{u,v}$ is uniformly convergent (Körner, 1988, Theorem 9.6), so we can write

$$f_{u,v}(\nu) = \sum_{k=-\infty}^{\infty} c_{u,v}(k)e^{ik\nu}.$$

Furthermore, the Fourier coefficients may be bounded by (Körner, 1988, Lemma 9.5)

$$|c_{u,v}(k)| \leq \frac{M_1}{|k|^2},$$

where $M_1 = \sup_{u,v,\nu} |\partial^2/\partial\nu^2 f_{u,v}(\nu)|$. We also have $\int f_{u,v}(\nu)d\nu = c_{u,v}(0) < \infty$ because $f_{u,v}$ is uniformly bounded on $\mathbb{T}$. Using the Fourier series representation, we may write the elementwise error as

$$
\begin{aligned}
&|D_T(s,t) - K_T(s,t)| = \\
&\left| \frac{2\pi}{T} \sum_{j=0}^{T-1} \sum_{k=-\infty}^{\infty} c_{u,v}(k) e^{ik\nu_j} e^{i\nu_j(s-t)} - \int_{-\pi}^{\pi} \sum_{k=-\infty}^{\infty} c_{u,v}(k) e^{ik\nu} e^{i\nu(s-t)} d\nu \right| \\
&= \left| \frac{2\pi}{T} \sum_{k=-\infty}^{\infty} \sum_{j=0}^{T-1} c_{u,v}(k) e^{ik\nu_j} e^{i\nu_j(s-t)} - \sum_{k=-\infty}^{\infty} \int_{-\pi}^{\pi} c_{u,v}(k) e^{ik\nu} e^{i\nu(s-t)} d\nu \right| \\
&= \left| \sum_{l=-\infty}^{\infty} 2\pi c_{u,v}((t-s) + lT) - 2\pi c_{u,v}(t-s) \right| \\
&= \left| \sum_{|l|=1}^{\infty} 2\pi c_{u,v}(t - s + lT) \right| \\
&\leq \sum_{|l|=1}^{\infty} \frac{2\pi M_1}{|t - s + lT|^2},
\end{aligned}
$$

where the order of summation and integration can be switched because of the absolute summability of the Fourier coefficients. Now we compute the square of the Frobenius norm.

$$
\begin{aligned}
\|\Delta_T(A) - \Sigma_T(A)\|_F^2 &= \sum_{0 \leq |t-s| \leq T-1} \left(\sum_{|l|=1}^{\infty} \frac{2\pi M_1}{|t - s + lT|^2} \right)^2 \\
&\leq 2 \sum_{k=0}^{T-1} \sum_{h=1}^{T-k} \left(\sum_{|l|=1}^{\infty} \frac{2\pi M_1}{|k + lT|^2} \right)^2
\end{aligned}
$$

Since k is between 0 and T, the largest term in the sum is the one with $|k - T|^2$ in the denominator, so we rewrite

the sum as

$$\begin{aligned}
\frac{1}{2}\|\Delta_T(A) - \Sigma_T(A)\|_F^2 &= \sum_{k=0}^{T-1}\sum_{h=1}^{T-k}\left(\frac{2\pi M_1}{|k-T|^2} + \sum_{l=1}^{\infty}\frac{2\pi M_1}{|k+lT|^2} + \sum_{l=-2}^{-\infty}\frac{2\pi M_1}{|k+lT|^2}\right)^2 \\
&\leq \sum_{k=0}^{T-1}\sum_{h=1}^{T-k}\left(\frac{2\pi M_1}{|k-T|^2} + \sum_{l=1}^{\infty}\frac{2\pi M_1}{|lT|^2} + \sum_{l=-1}^{-\infty}\frac{2\pi M_1}{|lT|^2}\right)^2 \\
&= \sum_{k=1}^{T-1}\sum_{h=1}^{T-k}\left(\frac{2\pi M_1}{|k-T|^2} + \frac{2}{T^2}\sum_{l=1}^{\infty}\frac{2\pi M_1}{l^2}\right)^2 \\
&= \sum_{k=1}^{T-1}\sum_{h=1}^{T-k}\left(\frac{2\pi M_1}{|k-T|^2} + \frac{2\pi^3 M_1}{3T^2}\right)^2 \\
&\leq \sum_{k=1}^{T-1}\sum_{h=1}^{T-k} 2\left(\frac{4\pi^2 M_1^2}{|k-T|^4} + \frac{4\pi^6 M_1^2}{9T^4}\right) \\
&\leq \sum_{k=1}^{T-1}\left(\frac{8\pi^2 M_1^2}{|k-T|^3}\right) + \frac{8\pi^6 M_1^2}{9T^2} \\
&= \sum_{j=1}^{T-1}\left(\frac{8\pi^2 M_1^2}{j^3}\right) + \frac{8\pi^6 M_1^2}{9T^2} \\
&\leq 8\pi^2 M_1^2 \sum_{j=1}^{\infty}\frac{1}{j^3} + \frac{8\pi^6 M_1^2}{9T^2} \\
&= M_1^2\left(8\pi^2\zeta(3) + \frac{8\pi^6}{9T^2}\right) < \infty. \ \square
\end{aligned}$$

The function ζ is the Riemann-Euler zeta function, and $\zeta(3) \approx 1.202$.

A.2

For the log determinant approximation (10) described in Section 3, a value for γ of 0.25 has some theoretical justification. Let $B_T(A) = \{b_{tj}\}$ be as described in Section 3 and $P_T = \{p_{tj}\}$ be the $T \times T$ matrix which performs the inverse DFT (scaled by $T^{-1/2}$), so that $C_T(A) = \sqrt{2\pi}P_T \circ B_T(A)$, where $\circ$ is the element-wise product. Consider the special case in which all entries of $B_T(A)$ are 1 except for those in the k'th row (i.e. $t = k$) so that $C_T(A) = \sqrt{2\pi}(P_T + \boldsymbol{e}_k\boldsymbol{v}^H)$, where $\boldsymbol{v}$ is the vector with j'th element $(b_{kj} - 1)\bar{p}_{kj}$ and $\boldsymbol{e}_k$ is the unit vector along the k'th coordinate.

Proposition 2: *If $b_{kj} = e^{\epsilon h_j}$, then with $\gamma = 0.25$,*

$$\log\left|\det C_T(A)/\sqrt{2\pi}\right| = \frac{1}{T}\sum_{j,t}\log\{b_{tj}\} + \gamma\sigma_T(A) + O(\epsilon^3)$$

as $\epsilon \to 0$.

Proof: By the matrix determinant lemma (Harville, 1997),

$$\det(C_T(A)/\sqrt{2\pi}) = \det(P_T)(1 + \boldsymbol{v}^H P_T^{-1}\boldsymbol{e}_k).$$

Because P_T is a unitary matrix, its determinant is ± 1, and its inverse is its conjugate tranpose. Therefore,

$$\left|\det\left(C_T(A)/\sqrt{2\pi}\right)\right| = 1 + \boldsymbol{v}^H P_T^{-1}\boldsymbol{e}_k = 1 + \frac{1}{T}\sum_{j=1}^{T}(b_{kj} - 1).$$

Now consider what happens as $\epsilon \to 0$. Expanding $\frac{1}{T}\sum_{j=1}^{T} b_{kj}$ in a Taylor series through order ϵ^2 and writing $\bar{h}$ for $\frac{1}{T}\sum_{j=1}^{T} h_j$, we have

$$\left|\det\left(C_T(A)/\sqrt{2\pi}\right)\right| = 1 + \left(\epsilon\bar{h} + \frac{\epsilon^2}{2T}\sum_{j=1}^{T} h_j^2 + O(\epsilon^3)\right).$$

Taking the log gives

$$\begin{aligned}\log\left|\det\left(C_T(A)/\sqrt{2\pi}\right)\right| &= \epsilon\bar{h} + \frac{\epsilon^2}{2T}\sum_{j=1}^{T} h_j^2 - \frac{1}{2}\epsilon^2\bar{h}^2 + O(\epsilon^3) \\ &= \epsilon\bar{h} + \frac{\epsilon^2}{2T}\sum_{j=1}^{T}(h_j - \bar{h})^2 + O(\epsilon^3)\end{aligned}$$

We have $\frac{1}{T}\sum_{j,k=1}^{T}\log\{b_{kj}\} = \epsilon\bar{h}$, so that the simple approximation (7) captures the $O(\epsilon)$ term in the log determinant. Furthermore, straightforward calculations yield $\sigma_T(A) = \frac{2\epsilon^2}{T}\sum_{j=1}^{T}(h_j - \bar{h})^2$, so that (10) has error $O(\epsilon^3)$ when $\gamma = 0.25$. □

A.3

In Section 3, we considered the decomposition $A(u,\nu) = m(u)\mu(\nu)R(u,\nu)$, where m, μ, and R are all positive-valued and bounded away from 0. In this appendix we explore the asymptotics of $\sigma_T(A)$ in each of two cases:
Case 1:

(I) For every u, R is absolutely continuous in ν with a uniformly bounded almost everywhere derivative.

(II) Except for possibly some finite number of values of u (not depending on ν), for every ν, R is absolutely continuous in u with a uniformly bounded almost everywhere derivative. At these exceptional u values, R may not be continuous in u, but for definiteness, we will assume R is right continuous in u for all ν.

Case 2:

(III) R is absolutely continuous in both u and ν with uniformly bounded almost everywhere derivative, and $R(0,\nu) = R(1,\nu)$ for all ν.

Proposition 3: *Under the conditions of Case 1,* $\sigma_T(A) = O(1)$*, and under the conditions of Case 2,* $\sigma_T(A) = O(T^{-1})$*.*

Proof: Recall that

$$\sigma_T(A) = \frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{T}\left\{\log\frac{b_{t+1,j}}{b_{tj}} - \frac{1}{T}\sum_{k=0}^{T-1}\log\frac{b_{t+1,k}}{b_{tk}}\right\}^2.$$

Writing $u_t = t/T$ and using the decomposition of A and $R(0,\nu) = R(1,\nu)$ for all ν, it is easily shown that

$$\sigma_T(A) = \frac{1}{T}\sum_{t=0}^{T-1}\sum_{j=0}^{T-1}\left\{\log\frac{R(u_{t+1},\nu_j)}{R(u_t,\nu_j)} - \frac{1}{T}\sum_{k=0}^{T-1}\log\frac{R(u_{t+1},\nu_k)}{R(u_t,\nu_k)}\right\}^2. \tag{13}$$

Expanding $\log\{R(u_{t+1},\nu_j)/R(u_t,\nu_j)\}$ in a Taylor series gives

$$\log\frac{R(u_{t+1},\nu_j)}{R(u_t,\nu_j)} = \frac{R(u_{t+1},\nu_j) - R(u_t,\nu_j)}{R(u_t,\nu_j)} + O\left(\left\{\frac{R(u_{t+1},\nu_j) - R(u_t,\nu_j)}{R(u_t,\nu_j)}\right\}^2\right).$$

Under the conditions of Case 2, using the uniformly bounded derivative condition, and the fact that R is bounded away from 0, we know that

$$\frac{R(u_{t+1},\nu_j) - R(u_t,\nu_j)}{R(u_t,\nu_j)} = O(T^{-1}).$$

Therefore, in (13) $\sigma_T(A)$ is T^{-1} multiplied by the sum of T^2 terms, each of which is $O(T^{-2})$, so $\sigma_T(A) = O(T^{-1})$.

Under the conditions of Case 1, R has finitely many discontinuities in u. Therefore,

$$\frac{R(u_{t+1},\nu_j) - R(u_t,\nu_j)}{R(u_t,\nu_j)} = \begin{cases} O(T^{-1}) & \text{no discontinuities between } u_t \text{ and } u_{t+1} \\ O(1) & \text{at least one discontinuity between } u_t \text{ and } u_{t+1}. \end{cases}$$

In (13), there are finitely many t for which the sum over j is $O(T)$, so $\sigma_T(A) = O(1)$. □

A.4

In order to compute the exact Gaussian likelihood, we must construct the exact covariance matrix, which results from the covariance function

$$K_T(s,t) = \int_{\mathbb{T}} A(s/T,\nu)A(t/T,\nu)^* e^{i\nu(s-t)}\, d\nu.$$

Here, we compute this integral numerically with the sum

$$\frac{2\pi}{J}\sum_{j=0}^{J-1} A(s/T,\nu_j)A(t/T,\nu_j)^* e^{i\nu_j(s-t)},$$

where $\nu_j = 2\pi j/J$, and J is a large integer ($J = 4000$ gives adequate accuracy in our examples). The computation is made efficient with the FFT, and the covariance matrix is filled quickly by taking advantage of its block Toeplitz structure under the piecewise constant spectrum model.

After the covariance matrix is constructed, we compute the Cholesky decomposition U. Then the log determinant term is simply the sum of the log of the diagonal entries of U, and the quadratic form term is $\boldsymbol{Z}_T'\Sigma_T(A)^{-1}\boldsymbol{Z}_T = \|U'^{-1}\boldsymbol{Z}_T\|^2$. Finally, we compute $U'^{-1}\boldsymbol{Z}_T$ with forward substitution.

References

Adak, S. (1998) Time-Dependent Spectral Analysis of Nonstationary Time Series. *Journal of the American Statistical Association* **93**, 1488–1501.

Dahlhaus, R. (1996) On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Processes and their Applications* **62**, 139–168.

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

Dahlhaus, R. (2000) A likelihood approximation for locally stationary processes. *The Annals of Statistics* **28**, 1762–1794.

Davis, R. A., Lee, T. C. M., Rodriguez-Yam, G. A. (2006) Structural Break Estimation for Nonstationary Time Series Models. *Journal of the American Statistical Association* **101**, 223-39.

Donoho, D., Mallat, S., and von Sachs, R. (1998) Estimating Covariances of Locally Stationary Processes: Rates of Convergence of Best Basis Methods, Technical Report 517, Stanford University, Dept. of Statistics.

Guo, W., Dai, M., Ombao, H. C., von Sachs, R. (2003) Smoothing Spline ANOVA for Time-Dependent Spectral Analysis. *Journal of the American Statistical Association* **98**, 643–52.

Harville, D. A. (1997) *Matrix Algebra From a Statistician's Perspective*. Springer-Verlag, New York.

Horn, R.A. and Johnson, C.R. (2006) *Matrix Analysis*. Cambridge University Press, Cambridge.

Körner, T. W. (1988) *Fourier Analysis*. Cambridge University Press, Cambridge.

Neumann, M. H., von Sachs, R. (1997) Wavelet thresholding in anisotropic function classes and application to

adaptive estimation of evolutionary spectra. *The Annals of Statistics* **25**, 38–76.

Ombao, H. C., Raz, J. A., von Sachs, R., Malow, B. A. (2001) Automatic Statistical Analysis of Bivariate Nonstationary Time Series. *Journal of the American Statistical Association* **96**, 543–60.

Ombao, H. C., Raz, J., von Sachs, R., Guo, W. (2002) The SLEX model of a non-stationary random process. *Annals of the Institute for Statistical Mathematics* **54**, 171–200.

Priestley, M. B. (1965) Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society: Series B* **27**, 204–237.

Priestley, M. B. (1981) *Spectral Analysis and Time Series*. Academic Press, London.

Whittle, P. (1962) Gaussian estimation in stationary time series. *Bulletin of the International Statistical Institute* **39**, 105–129.

Figure Captions:

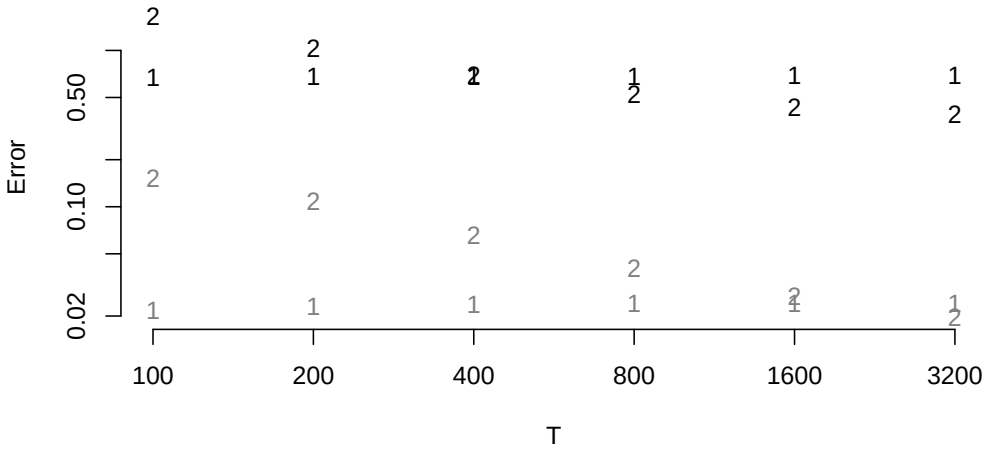

Figure 1: Absolute approximation errors of log determinants for $A_1(u,\nu) = \exp\{4\pi^{-1}\nu(2u - \lfloor 2u \rfloor)\}$ and $A_2(u,\nu) = \exp\{4\pi^{-1}\nu \sin(8.5\pi u)\}$. Symbol j corresponds to A_j, black to approximation (7) and gray to approximation (10) with $\gamma = 0.25$.

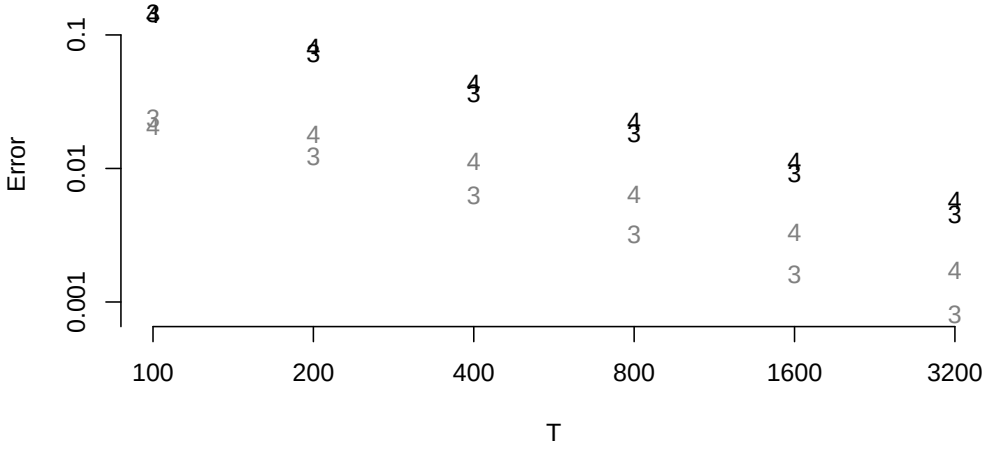

Figure 2: Absolute approximation errors of log determinants for $A_3(u,\nu) = \exp\{12\pi^{-1}\nu(1-2|u-0.5|)\}$ and $A_4(u,\nu) = \exp\{\pi^{-1}\nu\sin(12\pi u)\} + \cos^2(\pi^{-1}\nu\sin(16\pi u))$. Symbol j corresponds to A_j, black to approximation (7) and gray to approximation (10) with $\gamma = 0.25$.

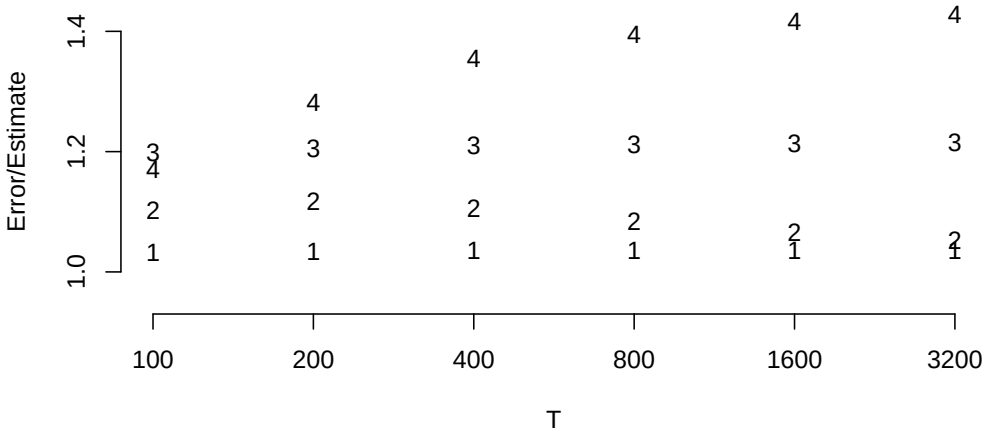

Figure 3: Ratio of the error of the first order approximation to $0.25\sigma_T(A_j)$ for the four functions considered in Figures 1 and 2.

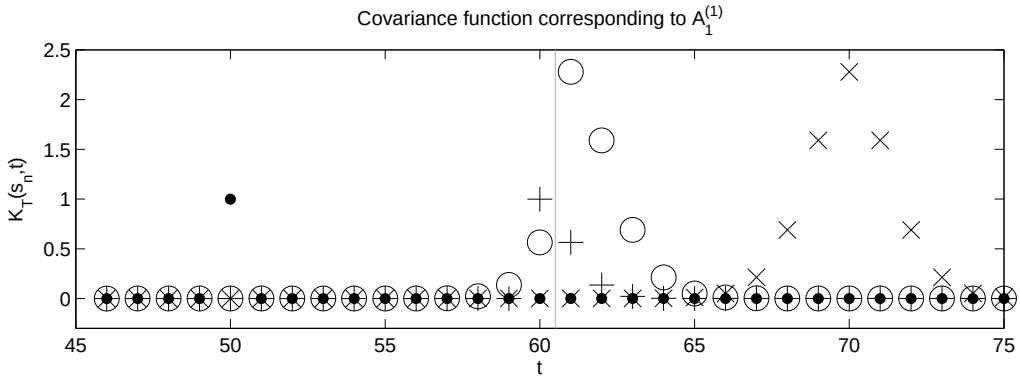

Figure 4: Covariance function corresponding to $A_\theta^{(1)}$ with $\theta = 1$. We plot $K_T(s_n, t)$ for $s_1 = 50$ (dots), $s_2 = 60$ (+), $s_3 = 61$ (circles), and $s_4 = 70$ ($\times$). The vertical line indicates the breakpoint in the process.

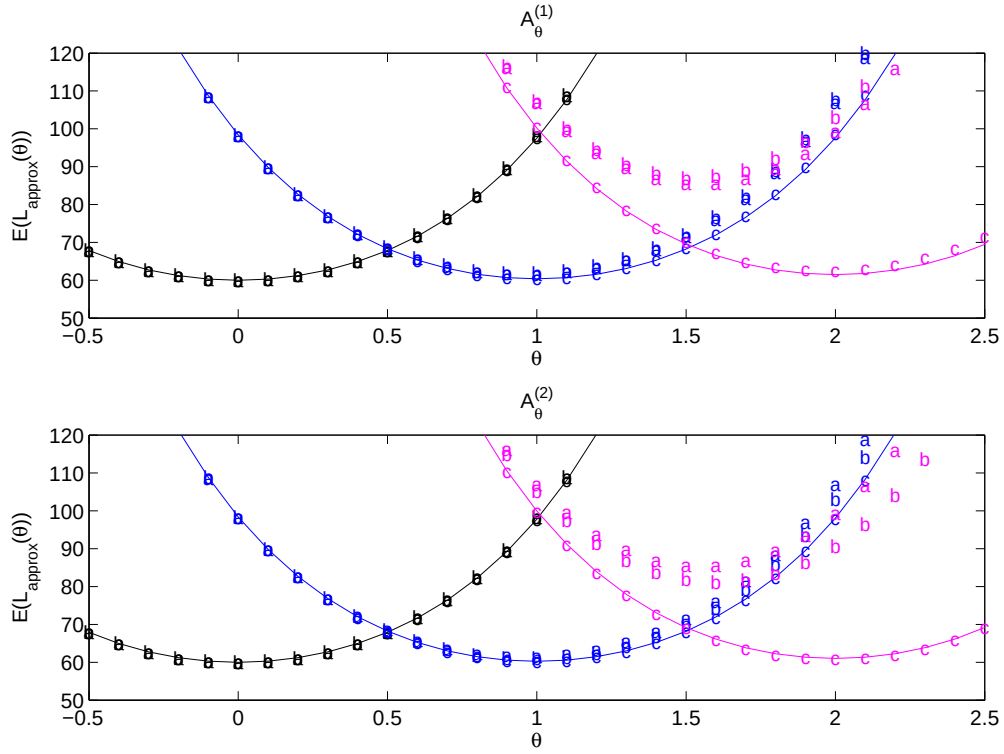

Figure 5: Expected values of the negative loglikelihood approximations $L_a(\theta)$, $L_b(\theta)$, and $L_c(\theta)$, with true parameter $\theta_0 = 0$ (black), $\theta_0 = 1$ (blue), and $\theta_0 = 2$ (magenta). The solid line indicates the expected value of the exact loglikelihood $L(\theta)$.

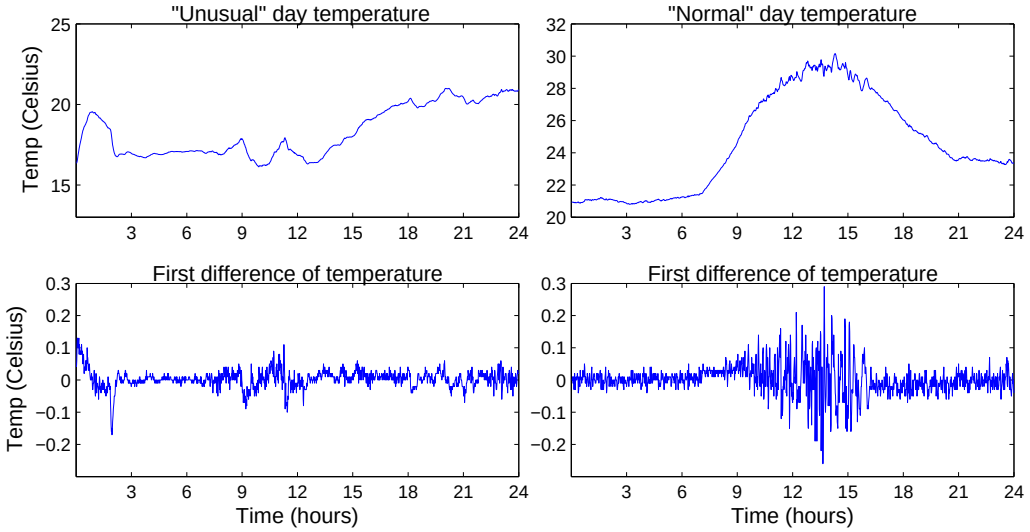

Figure 6: Temperature data

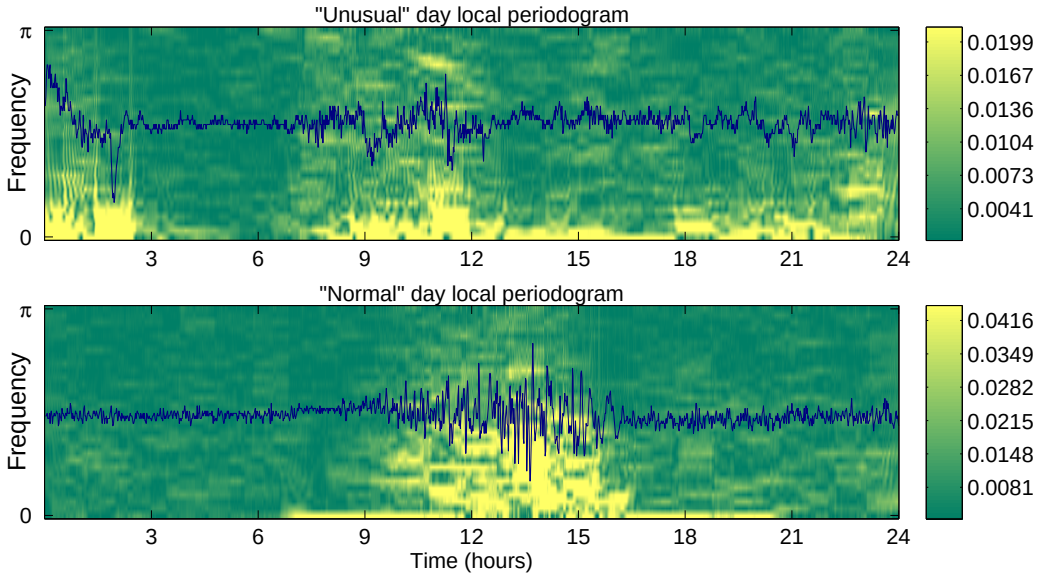

Figure 7: Local periodograms with 60 minute windows for both days

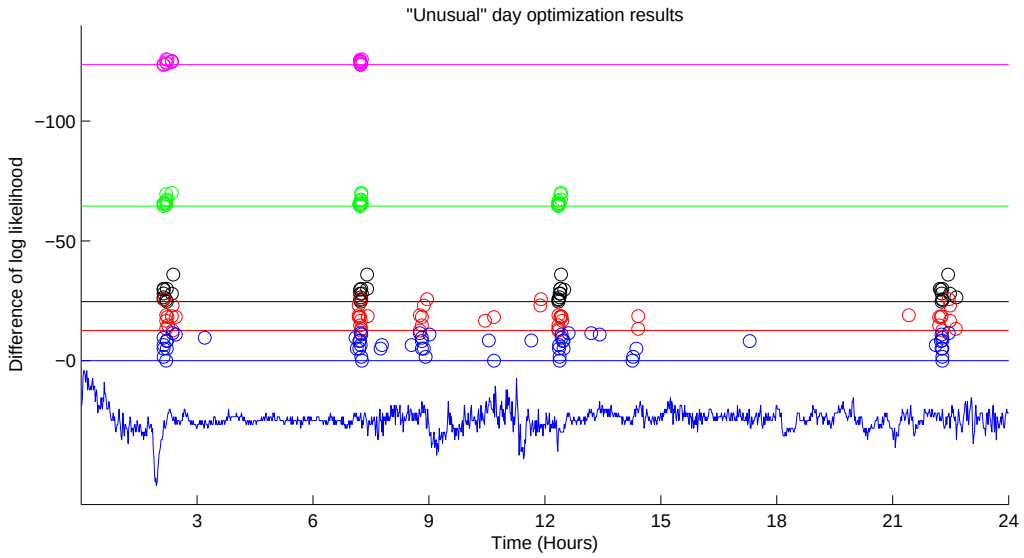

Figure 8: Optimization results for “unusual” day. Each row of circles represents the best partition found with the genetic algorithm.

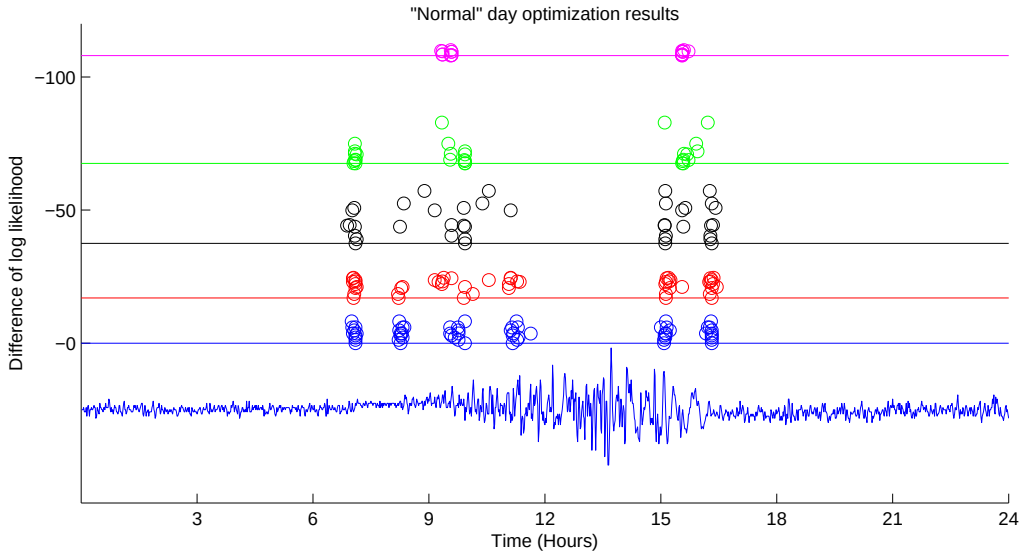

Figure 9: Optimization results for “normal” day. Each row of circles represents the best partition found with the genetic algorithm.

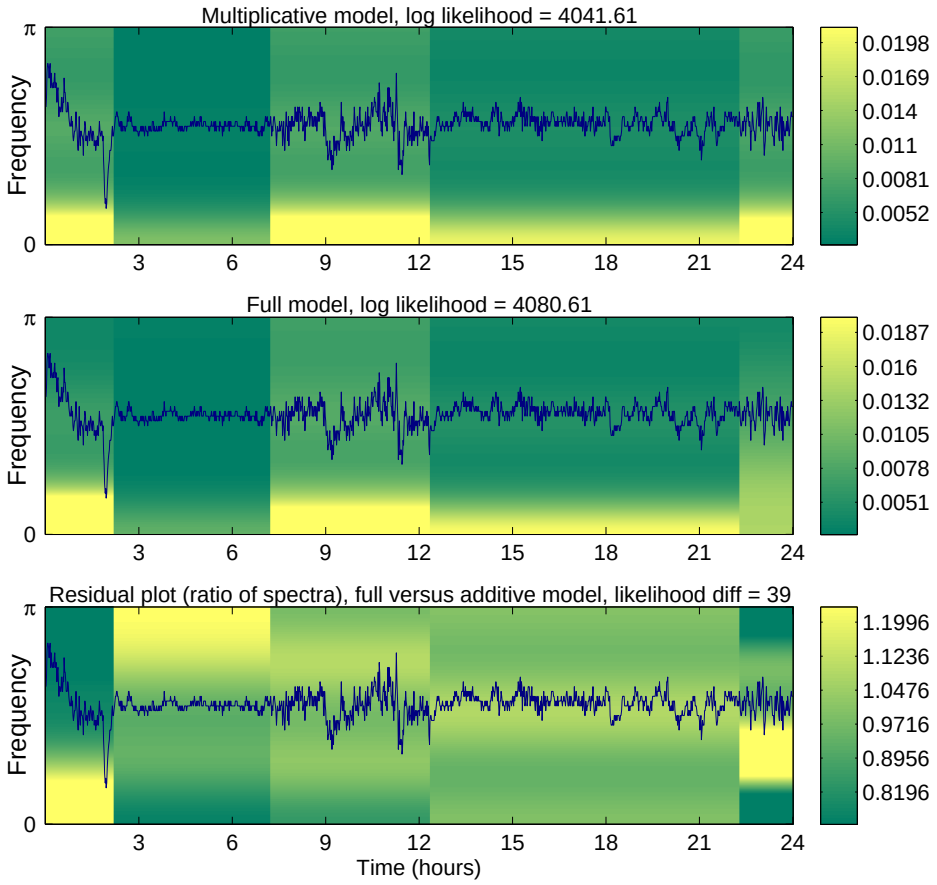

Figure 10: Maximum likelihood estimate of the time-varying transfer function of the “unusual” day under the multiplicative model (top), the full model (middle), and the ratio of the full model estimate to the multiplicative model estimate (bottom).

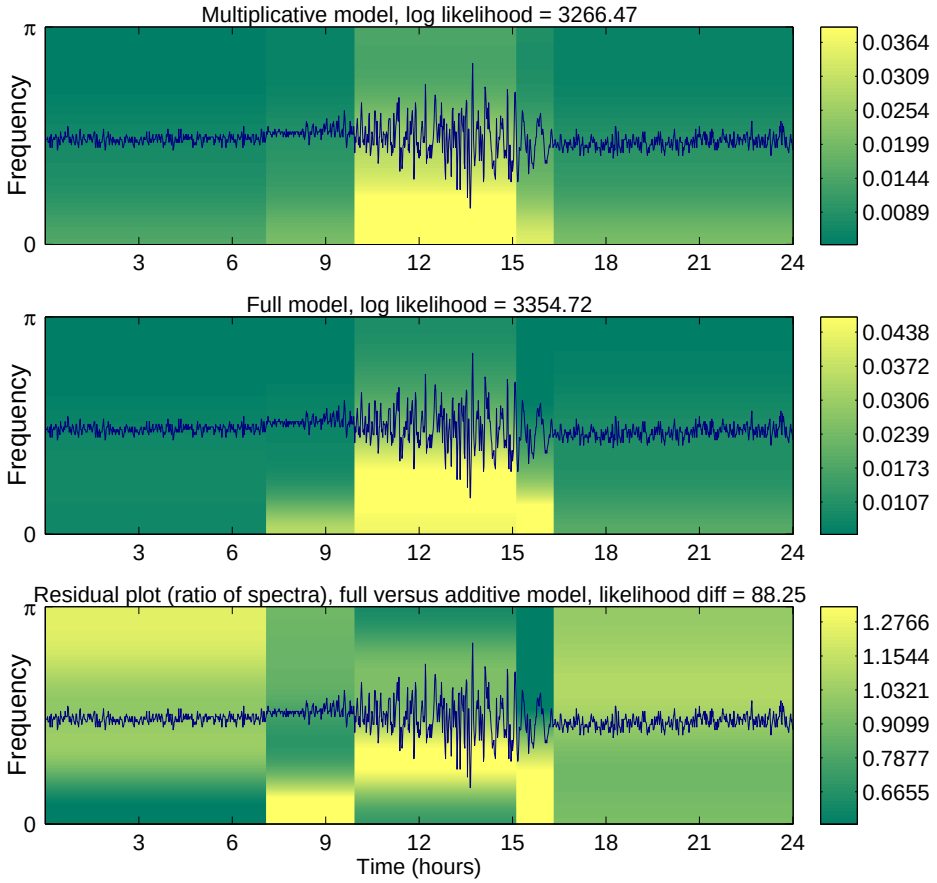

Figure 11: Maximum likelihood estimate of the time-varying transfer function of the "normal" day under the multiplicative model (top), the full model (middle), and the ratio of the full model estimate to the multiplicative model estimate (bottom).

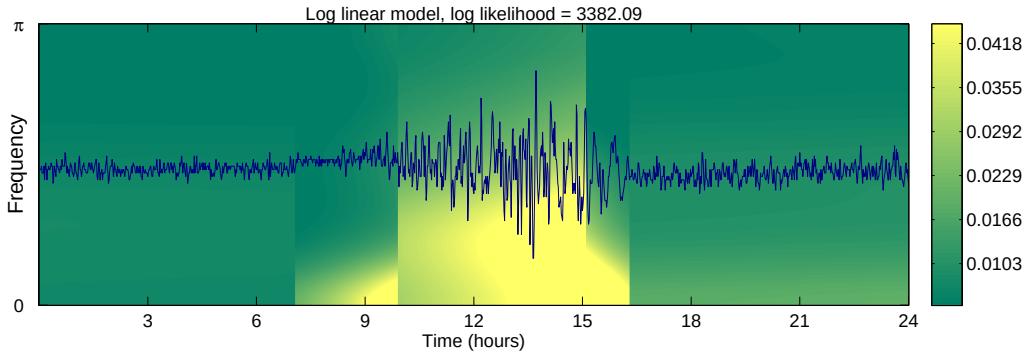

Figure 12: Maximum approximate likelihood estimate of the model in (12) for the "normal" day.

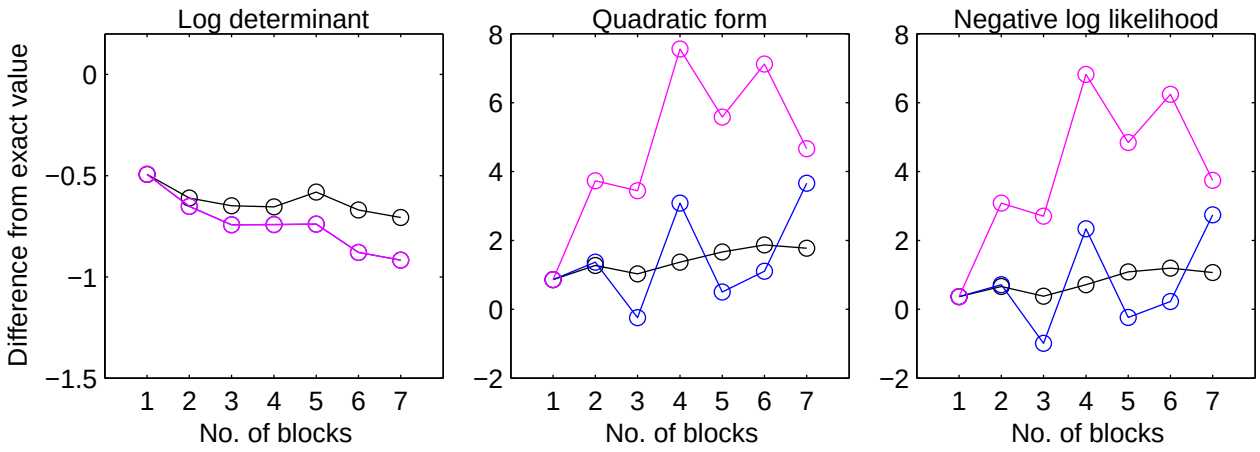

Figure 13: For the “unusual” day, the difference from the approximation to the exact value of the log determinant, the quadratic form and the loglikelihood for our approximations (black) and Dahlhaus’s approximations (1997, magenta), (2000, blue).

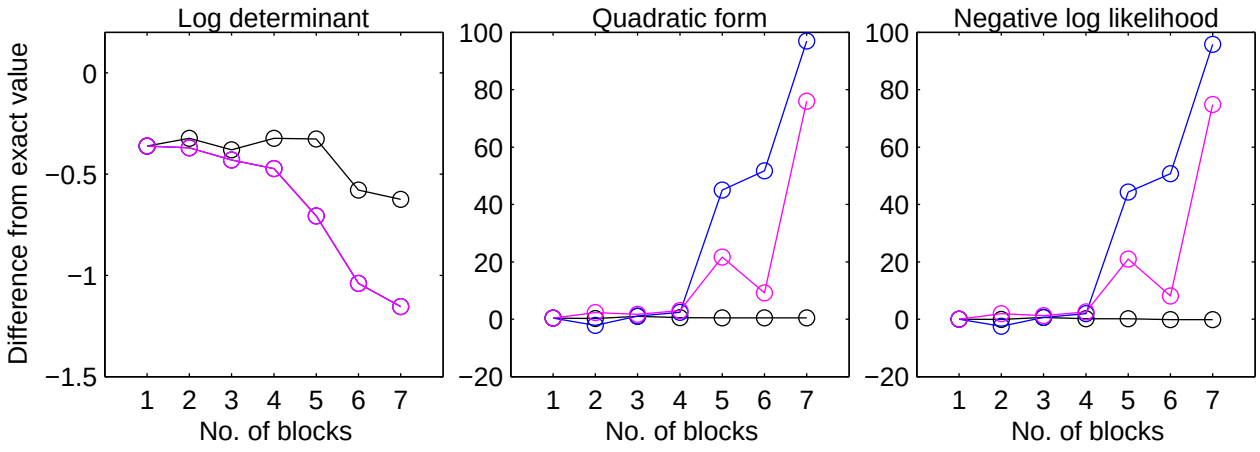

Figure 14: For the “normal” day, the difference from the approximation to the exact value of the log determinant, the quadratic form and the loglikelihood for our approximations (black) and Dahlhaus’s approximations (1997, magenta), (2000, blue).

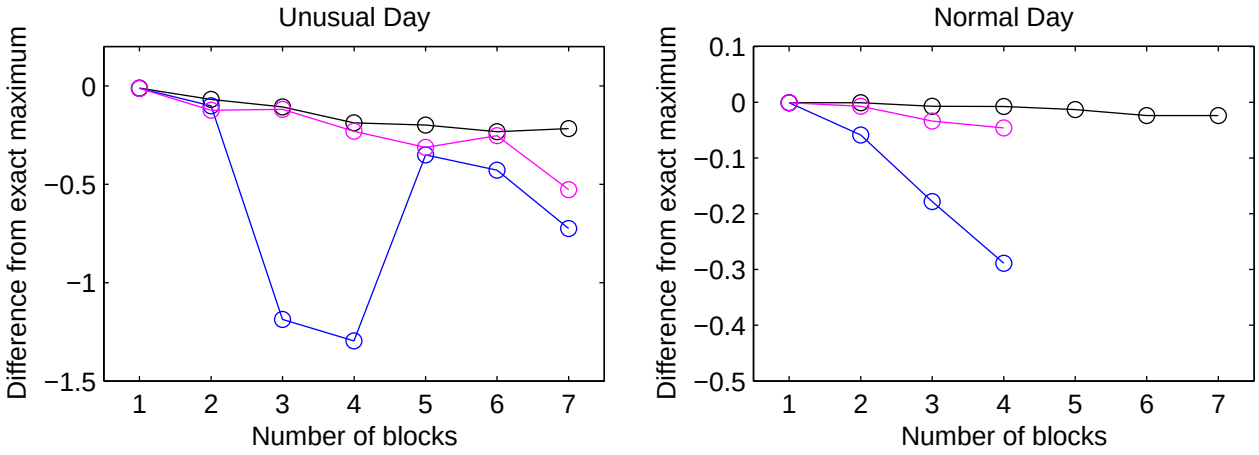

Figure 15: Exact loglikelihood on both days evaluated at the maximum likelihood estimate found using our approximation (black) and Dahlhaus's approximations (1997, magenta), (2000, blue). Difference from exact maximum loglikelihood, $L(\widehat{\theta}) - L(\widehat{\theta}_{approx})$ is plotted.