ADAPTIVE COVARIANCE ESTIMATION OF LOCALLY STATIONARY PROCESSES

STÉPHANE MALLAT *, GEORGE PAPANICOLAOU † AND ZHIFENG ZHANG †

FEBRUARY 1998

Contents

1. Introduction. Second order moments characterize entirely Gaussian processes and are often sufficient to analyze stochastic models, even though the processes may not be Gaussian. When processes are wide-sense stationary, their covariance defines a convolution operator. Many spectral estimation algorithms allow one to estimate the covariance operator from a few realizations, because it is diagonalized with Fourier series or integrals. When processes are not stationary, in the wide-sense, covariance operators may have complicated time varying properties. Their estimation is much more delicate since we do not know a priori how to diagonalize them. We will be dealing with wide-sense properties of processes

* Ecole Polytechnique, Paris, France and Courant Institute, New York University, mallat@cmapx.polytechnique.fr, supported by AFOSR grant F49620-96-1-0455
† Department of Mathematics, Stanford University Stanford CA 94305, papanico@math.stanford.edu, zhangz@math.stanford.edu. Work supported by NSF grant DMS 9308471 and by AFOSR grant F49620-1-0028

in this paper so we will not mention this explicitly. The ideas and methods of Calderon and Zygmund [8] in harmonic analysis have shown that although we are not able to find the basis which diagonalizes complicated integral operators in general, it is nevertheless possible to find well structured bases which compress them. This means that the operator is well represented by a sparse matrix with such a basis. This approach allows characterization of large classes of operators by the family of bases which do the compression. We show here that the ability to represent covariance operators by sparse matrices in a suitable basis leads to its efficient estimation from a few realizations.

We concentrate attention on the class of locally stationary processes, that is, processes whose covariance operators are approximately convolutions. Since cosines and sines diagonalize the covariance of stationary processes, it is natural to expect that *local* cosine functions are "almost" eigenvectors of locally stationary processes. This property is formalized by postulating that the covariance operator is well approximated by a nearly diagonal one in an appropriate local cosine basis. We show that if the covariance operator is a pseudo-differential operator of a specified class, then the process is locally stationary.

To estimate the covariance operator of a locally stationary process we search for a local cosine basis which compresses it and estimate its matrix elements. The size of the windows of a suitable local cosine basis must be adapted to the size of the intervals where the process is approximately stationary. Since we do not know in advance the size of approximate stationarity intervals, we introduce an algorithm that searches within a class of bases for a "best" basis, to compress the covariance operator. This search is done using data provided by a few realizations of the process. For locally stationary processes, we have a fast implementation of the search for a best local cosine basis based on the local cosine trees of Meyer [9] and Coifman and Wickerhauser [6].

In section 2 we study the properties of locally stationary processes and in section 3 we analyze the estimation of covariance operators with a "best" basis search. Fast numerical algorithms and their application to examples of locally stationary processes are described in section 4.

2. Locally Stationary Processes. Locally stationary processes appear in many physical systems in which the mechanisms that produce random fluctuations change slowly in time or space. Over short time intervals, such processes can be approximated by a stationary one. This is the case for many components of speech signals. Over a sufficiently short time interval, the throat behaves like a steady resonator which is excited by a stationary noise source. The length of these stationary time intervals can however vary greatly depending on the type of sound that is generated. In the next section we describe qualitatively the basics of locally stationary processes and explain how to construct "almost" eigenvectors of the covariance operator with local Fourier analysis. The corresponding "almost" eigenvalues are given by the time-varying spectrum. This intuitive introduction is made precise in section 2.2 by defining locally stationary processes as those whose covariance operators are well compressed in some local cosine basis. In section 2.3 we prove that pseudo-differential covariance operators are locally stationary. Such processes may also be realized by filtering white noise with a time-varying filter whose properties are described in section 2.4.

2.1. Time-varying spectrum. Let $X(t)$ be a real valued zero-mean process with covariance

$$R(t,s) = E\left\{X(t)X(s)\right\}.$$

The covariance operator is defined for any $f \in \mathbf{L^2}(\mathbb{R})$ by

$$Tf(t) = \int_{-\infty}^{+\infty} R(t,s)f(s)ds. \tag{1}$$

The inner product

$$< f, X >= \int_{-\infty}^{+\infty} f(t)X(t)dt$$

is a random variable which is a linear combination of the process values at different times. For any $f, g \in \mathbf{L^2}(\mathbb{R})$, the covariance operator gives the cross-correlation

$$E\{< f, X) >< g, X >^*\} =< Tf, g > . \tag{2}$$

The covariance can be expressed in terms of the distance between t and s and the mid-point between them

$$R(t,s) = C_0(\frac{t+s}{2}, t-s). \tag{3}$$

When the process is stationary then

$$C_0(\frac{t+s}{2}, t-s) = C_0(t-s)$$

and the covariance operator is a convolution

$$Tf(t) = \int_{-\infty}^{+\infty} C_0(t-s)f(s)ds = (C_0 \star f)(t).$$

If the process is locally stationary, we expect that in the neighborhood of any $x \in \mathbb{R}$, there exists an interval of size $l(x)$ where the process can be approximated by a stationary one. The size $l(x)$ of intervals of approximate stationarity may vary with the location x. For $t \in [x - \frac{l(x)}{2}, x + \frac{l(x)}{2}]$, the covariance is well approximated by a function of $t - s$

$$E\{X(t)X(s)\} \approx C(x, t-s) \ \text{ if } |t-s| \leq \tfrac{l(x)}{2}. \tag{4}$$

The decorrelation length $d(x)$ gives the maximum distance between two correlated points. For $t \in [x - \frac{l(x)}{2}, x + \frac{l(x)}{2}]$

$$E\{X(t)X(s)\} = C(\frac{t+s}{2}, t-s) \quad \approx 0 \ \text{ if } |t-s| \geq d(x). \tag{5}$$

Locally stationary processes have a decorrelation length that is smaller than half the size $l(x)$ of the stationarity interval

$$d(x) < \frac{l(x)}{2}. \tag{6}$$

The conditions (4) and (5) imply that if $t \in [x - \frac{l(x)}{2}, x + \frac{l(x)}{2}]$ then

$$C(\frac{t+s}{2}, t-s) \approx C(x, t-s) \ \ \forall s \in \mathbb{R}. \tag{7}$$

With the change of variables (3), the covariance operator

$$Tf(t) = \int_{-\infty}^{+\infty} C_0(\frac{t+s}{2}, t-s) f(s) ds$$

can be interpreted as a time varying convolution. To analyze the properties of this operator when $C(u,v)$ is a smooth function of u, Martin and Flandrin [10] have introduced a real "time varying spectrum", which is the Fourier transform of $C_0(u,v)$ with respect to v

$$\Lambda_0(u,\omega) = \int_{-\infty}^{+\infty} C_0(u,v) e^{-i\omega v} dv \tag{8}$$

$$= \int_{-\infty}^{+\infty} R(u+\frac{v}{2}, u-\frac{v}{2}) e^{-i\omega v} dv \tag{9}$$

$$= \int_{-\infty}^{+\infty} E\{X(u+\frac{v}{2}) X(u-\frac{v}{2})\} e^{-i\omega v} dv. \tag{10}$$

This "time-varying" spectrum is the expected Wigner-Ville distribution of the process $X(t)$

$$\Lambda_0(u,\omega) = E\{WX(t)\}$$

where the Wigner-Ville distribution is defined by

$$Wf(u,\omega) = \int_{-\infty}^{+\infty} f(u+\frac{v}{2}) f(u-\frac{v}{2}) e^{-i\omega v} dv. \tag{11}$$

The terminology "spectrum" should be interpreted carefully because $\Lambda_0(u,\omega)$ is generally not equal to the eigenvalues of T. It may in fact take negative values whereas T is a symmetric, positive operator whose spectrum is therefore always positive.

The regularity of the time-varying spectrum is related to the size of stationarity intervals $l(x)$ and the decorrelation length $d(x)$. If $u \in [x - \frac{l(x)}{2}, x + \frac{l(x)}{2}]$ then (5) shows that the covariance $C(u,v)$ has a fast decay in v relatively to $d(x)$. Its Fourier transform $\Lambda_0(u,\omega)$ with respect to v thus remains approximately constant over intervals of size $\frac{2\pi}{d(x)}$. Since $C(u,v)$ has negligible time-variation in $[x - \frac{l(x)}{2}, x + \frac{l(x)}{2}]$ we derive that for any $\xi \in \mathbb{R}$ the spectrum $\Lambda_0(u,\omega)$ can be approximated by a constant $\Lambda_0(x,\xi)$ in the time-frequency rectangle

$$(u,\omega) \in [x - \frac{l(x)}{2}, x + \frac{l(x)}{2}] \times [\xi - \frac{\pi}{d(x)}, \xi + \frac{\pi}{d(x)}]. \tag{12}$$

If the process $X(t)$ is stationary, the covariance operator T is a convolution whose eigenvectors are therefore the complex exponentials $e^{-i\omega v}$. In this case, the eigenvalues are given by the spectrum

$$\Lambda_0(u,\omega) = \Lambda_0(\omega) = \int_{-\infty}^{+\infty} C_0(v) e^{-i\omega v} dv.$$

If the process $X(t)$ is locally stationary, we show that $\Lambda_0(x,\xi)$ is an approximate eigenvalue of the covariance operator T. Approximate eigenvectors are time-frequency atoms whose energy are concentrated in the time-frequency rectangle (12), where $\Lambda_0(u,\xi)$ is approximately constant. The uncertainty principle proves that it is possible to construct such a time-frequency atom only if $d(x)$ is smaller than $l(x)$, which corresponds to the local stationarity condition (6).

Let $g_x(t)$ be a smooth window whose support is equal to $[x - \frac{l(x)}{2}, x + \frac{l(x)}{2}]$, and $\phi_{x,\xi}(t) = g_x(t)\cos(\xi t + \theta)$. We show with non rigorous derivations that if $X(t)$ is locally stationary then

$$T\phi_{x,\xi}(t) \approx \Lambda_0(x,\xi)\phi_{x,\xi}(t). \tag{13}$$

Applying the covariance operator to $\phi_{x,\xi}(t)$ gives

$$T\phi_{x,\xi}(t) = \int_{-\infty}^{+\infty} C(\frac{t+s}{2}, t-s)\phi_{x,\xi}(s) ds.$$

The support of $\phi_{x,\xi}(s)$ is $[x - \frac{l(x)}{2}, x + \frac{l(x)}{2}]$. The local stationarity condition (7) thus implies that

$$T\phi_{x,\xi}(t) \approx \int_{-\infty}^{+\infty} C(x, t-s)\phi_{x,\xi}(s) ds.$$

Parseval's identity gives

$$T\phi_{x,\xi}(t) \approx \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Lambda_0(x,\omega)\hat{\phi}_{x,\xi}(\omega) e^{i\omega u} d\omega, \tag{14}$$

where $\hat{\phi}_{x,\xi}(\omega)$ is the Fourier transform of $\phi_{x,\xi}(t)$

$$\hat{\phi}_{x,\xi}(\omega) = \frac{1}{2}[e^{i\theta}\hat{g}_x(\omega - \xi) + e^{-i\theta}\hat{g}_x(\omega + \xi)].$$

If $g_x(t)$ is a smooth window function, the energy of its Fourier transform $\hat{g}_x(\omega)$ is mostly concentrated in $[-\frac{\pi}{l(x)}, \frac{\pi}{l(x)}]$. The energy of $\hat{\phi}_{x,\xi}(\omega)$ is therefore localized in $[-\xi - \frac{\pi}{l(x)}, -\xi + \frac{\pi}{l(x)}] \cup [\xi - \frac{\pi}{l(x)}, \xi + \frac{\pi}{l(x)}]$. Since $\Lambda_0(x,\omega) = \Lambda_0(x,-\omega)$ and $d(x) < \frac{l(x)}{2}$, (12) implies that

$$\Lambda_0(x,\omega) \approx \Lambda_0(x,\xi) \quad \text{for } |\omega| \in [\xi - \tfrac{\pi}{l(x)}, \xi + \tfrac{\pi}{l(x)}].$$

It results from (14) that

$$T\phi_{x,\xi}(t) \approx \frac{\Lambda_0(x,\xi)}{2\pi}\int_{-\infty}^{+\infty} \hat{\phi}_{x,\xi}(\omega) e^{i\omega t} d\omega = \Lambda_0(x,\xi)\phi_{x,\xi}(t).$$

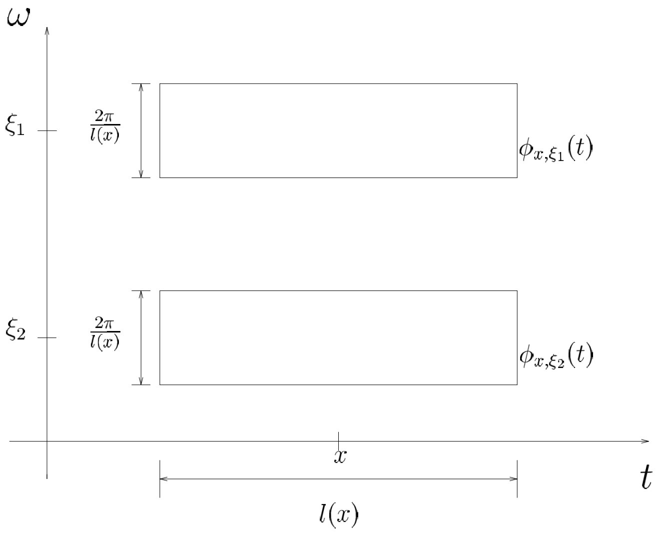

FIG. 1. A modulated window $\phi_{x,\xi}$ has a time support centered at x of size proportional to $l(x)$. Its Fourier transform is centered at $\omega = \xi$ and its energy is spread over an interval whose size is proportional to $\frac{2\pi}{l(x)}$. It is represented by a rectangle centered at (x, ξ) in the time-frequency plane (t, ω). Changing ξ translates the rectangle along the frequency axis.

In the time-frequency plane (t, ω), for $\omega > 0$ the approximate eigenfunction $\phi_{x,\xi}$ has an energy mostly concentrated in the rectangle

$$[x - \frac{l(x)}{2}, x + \frac{l(x)}{2}] \times [\xi - \frac{\pi}{l(x)}, \xi + \frac{\pi}{l(x)}].$$

Changing ξ modifies the location of the center of this rectangle as indicated in Figure 1. To show that $T\phi_{x,\xi}(t) \approx \Lambda_0(x, \xi)\phi_{x,\xi}(t)$ we used the fact that $\Lambda_0(t, \omega)$ is approximately constant over the time-frequency support of $\phi_{x,\xi}$. This is a crucial property for locally stationary processes.

2.2. Local Cosine Approximations. It could be tempting to look for the exact eigenvectors of the operator T. The characterization of covariance operators through eigenvectors is however unstable. Indeed, when the eigenvalues are close, slight changes of the operator may change completely the eigenvectors. In most cases, the eigenvectors are complicated functions that cannot be used easily to describe properties of T. Instead, we construct orthogonal bases of approximate eigenvectors. Let $\{\phi_n\}_{n\in\mathbb{N}}$ be an orthonormal basis of $\mathbf{L^2}(\mathbb{R})$. Any $f \in \mathbf{L^2}(\mathbb{R})$ can be expanded in this basis

$$f(t) = \sum_{n=0}^{+\infty} < f, \phi_n > \phi_n(t).$$

The covariance operator is represented by the matrix elements $\{< T\phi_n, \phi_m >\}_{(n,m)\in\mathbb{N}^2}$. The expansion coefficients of Tf in the basis $\{\phi_n\}_{n\in\mathbb{N}}$ are obtained by matrix multiplication

$$Tf(t) = \sum_n \left(\sum_m < T\phi_m, \phi_n >< f, \phi_m > \right) \phi_n(t). \tag{15}$$

In a basis of approximate eigenvectors the matrix coefficients $< T\phi_m, \phi_n >$ decay rapidly as $|n-m|$ increases.

For locally stationary processes, we saw that one can find window cosine functions $\phi_{x,\xi}(t) = g_x(t)\cos(\xi t + \theta)$ that are approximate eigenvectors of T. We explain the construction of Coifman, Malvar and Meyer [7, 9, 11] that yields an orthonormal basis with such local cosine vectors. The real line $\mathbb{R}$ is partitioned into intervals $[a_p, a_{p+1}]$ of size

$$l_p = a_{p+1} - a_p.$$

We suppose that the sequence a_p is increasing and that

$$\lim_{p\to-\infty} a_p = -\infty \quad , \quad \lim_{p\to+\infty} a_p = +\infty$$

so that the whole line is segmented by these intervals. Each interval $[a_p, a_{p+1}]$ is covered by a window function $g_p(t)$. Let $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$ be the support of $g_p(t)$. We construct $g_p(t)$ so that its support intersects only the support of $g_{p-1}(t)$ and the support of $g_{p+1}(t)$, which means that

$$l_p \geq \eta_p + \eta_{p+1}. \tag{16}$$

The supports of $g_p(t)$ and $g_{p-1}(t)$ intersect in $[a_p - \eta_p, a_p + \eta_p]$. Over this interval both windows must be symmetric with respect to a_p

$$g_p(t) = g_{p-1}(2a_p - t). \tag{17}$$

The windows $\{g_p(t)\}_{p\in\mathbf{Z}}$ are, moreover, constructed so as to cover uniformly the time axis

$$\forall t \in \mathbb{R} \quad , \quad \sum_{n=-\infty}^{+\infty} |g_p(t)|^2 = 1. \tag{18}$$

Such window functions are illustrated in Figure 2. The following theorem [9, 7] shows that the resulting local cosine family is an orthogonal basis.

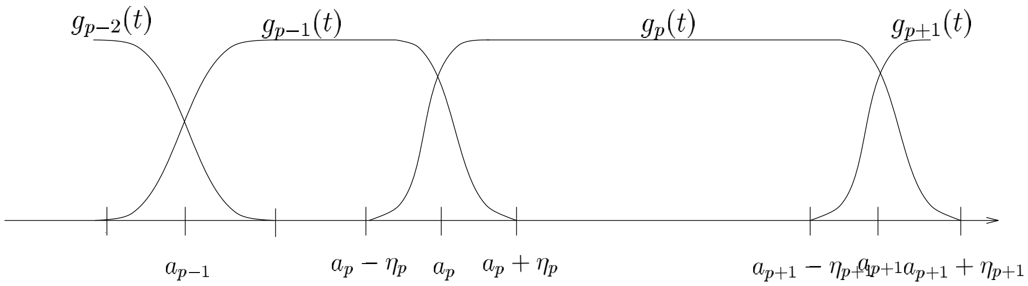

FIG. 2. Smooth cutoff window functions $g_p(t)$, $p \in N$, used in local cosine bases. The supports of adjacent windows $g_p(t)$ and $g_{p-1}(t)$ intersect over the interval $[a_p - \eta_p, a_p + \eta_p]$. Over this interval, both windows are symmetric with respect to a_p.

THEOREM 2.1 (COIFMAN, MALVAR, MEYER). *If (16,17,18) are satisfied then*

$$\left\{ \phi_{p,k}(t) = g_p(t)\sqrt{\frac{2}{l_p}} \cos\left[\frac{\pi(k+\frac{1}{2})}{l_p}(t-a_p)\right] \right\}_{k\in\mathbb{N}, p\in\mathbf{Z}} \tag{19}$$

is an orthonormal basis of $\mathbf{L^2}(\mathbb{R})$.

The support of $\phi_{p,k}(t)$ is $[a_p - \eta_p, a_{p+1} + \eta_{p+1}]$. The frequency of the cosine modulation is

$$\xi_{p,k} = \frac{\pi(k+\frac{1}{2})}{l_p}. \tag{20}$$

Let $\hat{g}_p(\omega)$ be the Fourier transform of $g_p(t)$. The Fourier transform of $\phi_{p,k}(t)$ is then

$$\hat{\phi}_{p,k}(\omega) = \frac{e^{ia_p\xi_{p,k}}}{\sqrt{2l_p}} \left(\hat{g}_p(\omega - \xi_{p,k}) + \hat{g}_p(\omega + \xi_{p,k})\right).$$

The bandwidth of $\hat{\phi}_{p,k}(\omega)$ around $\xi_{p,k}$ and $-\xi_{p,k}$ is equal to the bandwidth of $\hat{g}_p(\omega)$. If $g_p(t)$ is a smooth function, its frequency bandwidth is proportional to $\frac{2\pi}{l_p}$.

A local cosine basis can be attached to a partition (pavement) of the time-frequency plane by representing each $\phi_{p,k}(t)$ with a rectangle which approximates the time support by $[a_p, a_{p+1}]$ and the frequency support with $[\xi_{p,k} - \frac{\pi}{l_p}, \xi_{p,k} + \frac{\pi}{l_p}]$. The time and frequency spread of $\phi_{p,k}$ goes beyond the rectangle

$$[a_p, a_{p+1}] \times [\xi_{p,k} - \frac{\pi}{l_p}, \xi_{p,k} + \frac{\pi}{l_p}]$$

but this correspondence has the advantage of associating an exact partition of the time-frequency plane with any orthogonal local cosine basis $\{\phi_{p,k}(t)\}_{p\in\mathbf{Z},k\in\mathbb{N}}$, as shown in Figure 3.

Our qualitative analysis of locally stationary processes shows that there exists local cosine vectors that are almost eigenvectors of the covariance operator T. This property is used as a characterization of locally stationary processes by the following definition. It imposes the existence of an orthogonal basis of local cosine vectors that are almost eigenvectors of T. In a given time neighborhood, the size of the local cosine windows corresponds to the size of the interval where $X(t)$ is approximately stationary.

Definition 1. *A process* $X(t)$ *is locally stationary if there exists a local cosine basis*

$$\left\{\phi_{p,k}(t) = g_p(t)\sqrt{\frac{2}{l_p}}\cos\left[\frac{\pi(k+\frac{1}{2})}{l_p}(t-a_p)\right]\right\}_{k\in\mathbb{N},p\in\mathbf{Z}}$$

such that for some constants $\mu < 1$ *and* $A > 0$ *we have that for all* $p \neq q$

$$\frac{\max(l_p, l_q)}{\min(l_p, l_q)} \leq A|p-q|^\mu, \tag{21}$$

and for all $n > 1$ *we can find a constant* Q_n *such that for all* $(p,q,k,j) \in \mathbf{Z}^2 \times \mathbb{N}^2$ *the matrix elements of the covariance operator satisfy*

$$| < T\phi_{p,k}, \phi_{q,j} > | \leq \frac{Q_n}{(1+|p-q|^n)(1+|\max(l_p,l_q)(\xi_{p,k}-\xi_{q,j})|^n)}. \tag{22}$$

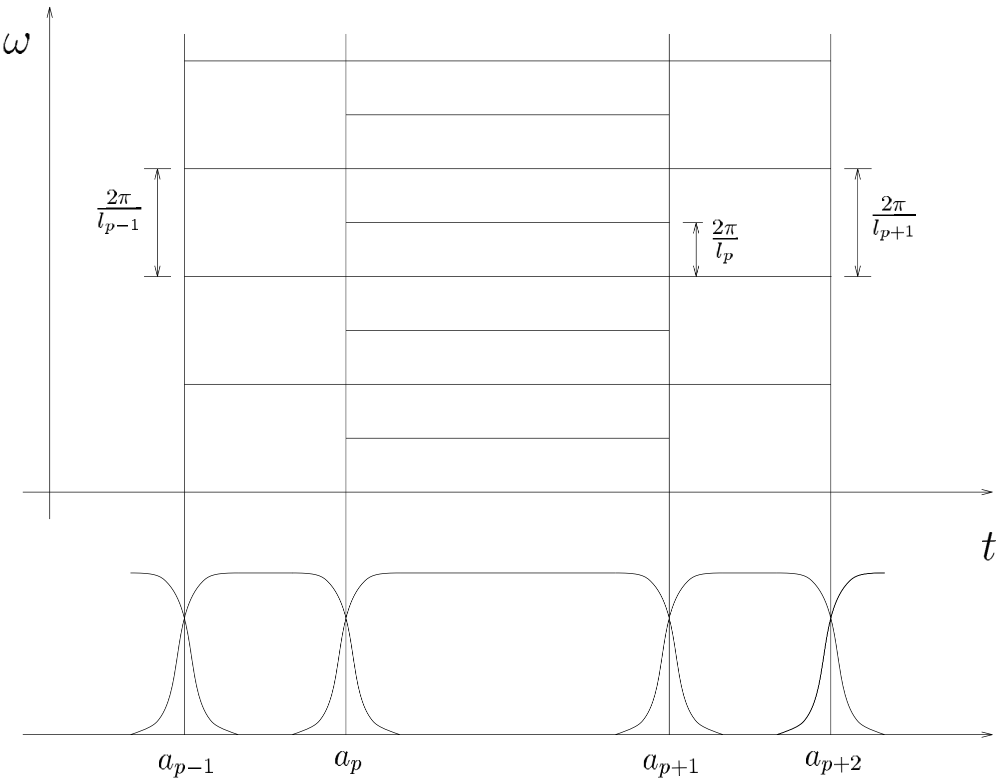

FIG. 3. Time-frequency tiling of a local cosine basis. Each box represents the time-frequency localization (approximate support) of a function in the basis. The collection of all the boxes forms a partition of the time-frequency plane.

The parameters $\{l_p\}$ specify the support of the windows $g_p(t)$. They indicate the size of the intervals where $X(t)$ is approximately stationary. Condition (21) demands that the size of these intervals should have a relatively slow variation in time. Condition (22) imposes that the matrix elements of the covariance $< T\phi_{p,k}, \phi_{q,j} >$ have a fast decay when we increase $|p-q|$ and $|\xi_{p,k} - \xi_{q,j}|$, which depend respectively upon the distance between the time and the frequency supports of $\phi_{p,k}$ and $\phi_{q,j}$. This means that $T\phi_{p,k}$ is a function that is mostly localized in the same time-frequency region as $\phi_{p,k}$. Each local cosine vector $\phi_{p,k}$ is therefore "almost" an eigenvector of T.

The covariance operator T is not diagonal in the local cosine basis but if it comes from a locally stationary process it can be approximated by a symmetric, sparse operator B_K constructed from T by keeping only the matrix elements $< T\phi_{p,k}, \phi_{q,j} >$ for which $\phi_{p,k}$ and $\phi_{q,j}$ are in the same time-frequency neighborhood. Inserting the expression (20) of $\xi_{p,k}$ and $\xi_{q,j}$ we define B_K by

$$< B_K\phi_{p,k}, \phi_{q,j} >= \begin{cases} < T\phi_{p,k}, \phi_{q,j} > & \text{if } |p-q| \le K \text{ and} \\ & |\max(l_p, l_q)(\frac{k+\frac{1}{2}}{l_p} - \frac{j+\frac{1}{2}}{l_q})| \le K \\ 0 & \text{otherwise} \end{cases}$$

For each (p,k), $< B_K\phi_{p,k}, \phi_{q,j} >\neq 0$ for at most $(2K+1)^2$ coefficients (q,j). When the window lengths l_p are not all the same, B_K does not have a band structure exactly. However, it has fewer non-zero coefficients than the band restriction of the T operator to elements for which $|k-j| \leq K$ and $|p-q| \leq K$.

The sup operator norm of T is denoted

$$\|T\|_s = \sup_{\|f\|=1} \|Tf\|,$$

where $\|f\|$ and $\|Tf\|$ are the $\mathbf{L^2}(\mathbb{R})$ norms. The following theorem shows that $\|T\|_s$ is bounded and that $\|T - B_K\|_s$ decays rapidly when K increases.

THEOREM 2.2. *If T is the covariance operator of a locally stationary process then*

$$\|T\|_s < +\infty. \tag{23}$$

Moreover, there exist for all integers $n > 1$ constant A_n such that for all $K > 0$

$$\|T - B_K\|_s \leq \frac{A_n}{1+K^n}. \tag{24}$$

The proof of this theorem is given in appendix A. The theorem guarantees that the covariance operator of a locally stationary process is arbitrarily well approximated by a sparse operator in an appropriate local cosine basis. Next section connects our definition of local stationarity to the properties of the covariance informally discussed in section 2.1.

2.3. Pseudo-differential Covariance Operators. The covariance operators of locally stationary processes introduced in section 2.1 were qualitatively described as time varying convolution operators. Such operators can be considered as pseudo-differential operators. We study necessary conditions which guaranty that the resulting process is locally stationary, in the sense of definition 1.

To study the properties of the covariance, we make a non orthogonal change of variables in the covariance $R(t,s)$, as opposed to the orthogonal change of variable (3), so that

$$R(t,s) = C_1(t, t-s). \tag{25}$$

The covariance operator can therefore be written as

$$Tf(t) = \int_{-\infty}^{+\infty} C_1(t, t-s) f(s) ds. \tag{26}$$

Let us define a new "time-varying spectrum" by

$$\Lambda_1(t,\omega) = \int_{-\infty}^{+\infty} C_1(t,v) e^{-iv\omega} dv.$$

The function $\Lambda_1(t,\omega)$ has complex values because in general $C_1(t,-v) \neq C_1(t,v)$. Applying Parseval's identity to (26) yields

$$Tf(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Lambda_1(t,\omega) \hat{f}(\omega) e^{i\omega t} d\omega.$$

In the theory of pseudo-differential operators $\Lambda_1(t,\omega)$ is called the symbol of T.

In section 2.1 we give two qualitative conditions for a process to be approximately stationary over an interval of size $l(t)$ in the neighborhood of t. One is that the covariance should vary slowly over $[t-\frac{l(t)}{2}, t+\frac{l(t)}{2}]$. This may be done by supposing that for all $k \geq 0$ there exists a constant A_k such that

$$|\partial_t^k \Lambda_1(t,\omega)| \leq \frac{A_k}{l^k(t)}.$$

The other is that the decorrelation or decay of $C_1(t,v)$ as a function of v should also be rapid compared to $l(t)$. This means that for any $j \geq 0$ there exists B_j such that

$$\int_{-\infty}^{+\infty} |v|^j |C_1(t,v)| dv \leq B_j l^j(t). \tag{27}$$

Since the Fourier transform of $(-iv)^j C_1(t,v)$ is $\partial_\omega^j \Lambda_1(t,\omega)$ and the integral (27) gives an upper bound on the Fourier transform, this condition implies that

$$|\partial_\omega^j \Lambda_1(t,\omega)| \leq B_j l^j(t).$$

We must now show that a process $X(t)$ satisfying these two conditions is locally stationary in the sense of Definition 1. This is the main theorem of this paper and it gives sufficient conditions on the covariance function so that there exists a basis of local cosine vectors that are almost eigenvectors of the covariance operator.

THEOREM 2.3. *Suppose that there exists a function $l(t)$ such that for all $k \geq 0$ and $j \geq 0$ we can find $A_{k,j}$ which satisfies*

$$|\partial_t^k \partial_\omega^j \Lambda_1(t,\omega)| \leq A_{k,j}\; l^{j-k}(t). \tag{28}$$

If for some $\alpha < \frac{1}{2}$ and a constant A

$$\forall (t,u) \in \mathbb{R}^2 \quad , \quad |l(t) - l(u)| \leq A|t-u|^\alpha, \tag{29}$$

and if

$$\inf_{t\in\mathbb{R}} l(t) > 0, \tag{30}$$

then T is the covariance operator of a locally stationary process in the sense of Definition 1.

The function $l(x)$ specifies the size of the neighborhood of x in which $X(t)$ is approximately stationary. When $l(t) = l$ is a constant, the covariance operator T whose symbol satisfies (28) is a classical pseudo-differential operator. It is well known [8] that such pseudo-differential operators are well compressed in a local cosine basis where all windows have a constant size $l_p = l$. When $l(t)$ varies and can potentially grow to $+\infty$, condition (28) on the symbol defines a larger and non-standard class of scaled pseudo-differential operators.

The proof in appendix B constructs an appropriate local cosine basis in which T satisfies the off-diagonal decay conditions (22)

$$|< T\phi_{p,k}, \phi_{q,j} >| \leq \frac{Q_n}{(1+|p-q|^n)(1+|\max(l_p,l_q)(\xi_{p,k}-\xi_{q,j})|^n)} \tag{31}$$

for all $(p, q, k, j) \in \mathbf{Z}^2 \times \mathbb{N}^2$. Each window $g_p(t)$ covers an interval $[a_p, a_{p+1}]$ of size $l_p = l(a_p)$. It corresponds to a time domain where $\Lambda_1(t, \omega)$ have small variations and where the underlined process $X(t)$ is approximately stationary. Conditions (29,30) guarantee that the windows length l_p satisfy the slow variation condition (21) imposed by the definition of local stationarity .

The stationarity length $l(t)$ is not uniquely specified by $\Lambda_1(t, \omega)$. When constructing the windows of the local cosine basis, we would like the matrix elements $| < T\phi_{p,k}, \phi_{q,j} > |$ to have the fastest possible decay away from the diagonal, so as to approximate as well as possible T with a sparse operator B_K. The constants Q_n that appear in (31) grow with the values of $A_{k,j}$ of (28). It is thus important to know when these constants $A_{k,j}$ are small and, if possible, remain uniformly bounded for all k and j. In many cases we can choose $l(x)$ to be proportional to

$$\frac{1}{\sup_{\omega \in \mathbb{R}} |\partial_t \Lambda_1(x, \omega)|},$$

which is a measure of the size of a neighborhood of x in which $\Lambda_1(t, \omega)$ has variations of order one, for all ω.

2.4. Time-Varying Filtering of White Noise. Stationary processes can be constructed by filtering white noise with a time invariant filter. We may therefore expect that a locally stationary process can be synthesized by filtering white noise with an appropriate time-varying filter. This approach to non stationary processes was followed by Priestley [3]. Here, by asking that the time-varying filter be a pseudo-differential operator, we show that the resulting process is locally stationary.

The Cramer representation gives a spectral decomposition of square integrable stationary processes $X(t)$

$$X(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(\omega) e^{i\omega t} d\hat{Z}(\omega),$$

where $\hat{Z}(\omega)$ has orthogonal increments

$$E\{d\hat{Z}(\omega) d\hat{Z}^*(\omega')\} = 2\pi\delta(\omega - \omega') d\omega d\omega'. \tag{32}$$

This can be interpreted as filtering of white noise with a time-invariant filter L defined for any $f \in \mathbf{L^2}(\mathbb{R})$ by

$$Lf(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(\omega) e^{i\omega t} \hat{f}(\omega) d\omega = \int_{-\infty}^{+\infty} K(t-s) f(s) ds,$$

where $\hat{f}(\omega)$ and $A(\omega)$ are respectively the Fourier transform of $f(v)$ and $K(v)$. The kernel $K(v)$ is the impulse response of L.

Priestley [3] studied a class of non-stationary processes obtained through a time varying filtering of white noise

$$X(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(t, \omega) e^{i\omega t} d\hat{Z}(\omega). \tag{33}$$

The process $\hat{Z}(\omega)$ has orthogonal increments that satisfy (32). The corresponding time-varying filter L is

$$Lf(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} A(t,\omega)e^{i\omega t}\hat{f}(\omega)d\omega = \int_{-\infty}^{+\infty} K(t,t-s)f(s)ds, \tag{34}$$

where $A(t,\omega)$ is the Fourier transform of $K(t,v)$ with respect to v. The kernel $K(t,v)$ can be interpreted as a time-varying impulse response.

Priestley defines the evolutionary spectrum to be $|A(t,\omega)|^2$. The kernel $A(t,\omega)$ depends upon the covariance T of the process $X(t)$ since we only specify the second order properties of $d\hat{Z}(\omega)$. However, $A(t,\omega)$ and L are not determined uniquely by T. Since the increments $d\hat{Z}(\omega)$ are uncorrelated, use of (32) shows that

$$R(t,s) = E\{X(t)X^*(s)\} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} A(t,\omega)A^*(s,\omega)e^{i\omega(t-s)}d\omega.$$

The covariance operator is thus related to the time-varying filter by

$$T = LL^t, \tag{35}$$

where L^t is the adjoint operator. In other words, L is a "square root" of the positive symmetric operator T. There exists, however, an infinite number of such square roots. If L is any solution of (35) then for any U such that $UU^t = I$, LU is also a solution of (35). Note that the real time-varying spectrum $\Lambda_0(\frac{t+s}{2},\omega)$ defined by (8) also satisfies

$$R(t,s) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Lambda_0(\frac{t+s}{2},\omega)e^{i\omega(t-s)}d\omega.$$

However $A(t,\omega)A^*(s,\omega)$ is in general not equal to $\Lambda_0(\frac{t+s}{2},\omega)$. In particular, $|A(t,\omega)|^2$ is always positive whereas $\Lambda_0(t,\omega)$ is not. To define $A(t,\omega)$ in a unique way, Priestley imposes the condition that the inverse Fourier transform of $A(t,\omega)$ with respect to ω is maximally concentrated around zero [4]. This is equivalent to imposing a maximum smoothness conditions on $A(t,\omega)$ with respect to ω. When trying to estimate the evolutionary spectrum $|A(t,\omega)|^2$, there is, however, no guarantee that we do estimate the maximally smooth kernel. The non-uniqueness of the evolutionary spectrum has remained an issue in Priestey's approach, and we prefer to work directly with the covariance operator which is uniquely defined.

Benassi, Jaffard and Roux [1] have studied a class of non-stationary processes, obtained with elliptic pseudo-differential filters L, that have weak regularity conditions. They proved that the covariance operator of these processes is well compressed in a wavelet basis. These processes are not locally stationary but are used to construct multi-fractal models. The following theorem concentrates on locally stationary processes $X(t)$ and gives sufficient conditions on the symbol $A(t,\omega)$ of L.

THEOREM 2.4. *Suppose that there exists a function $l(t)$ such that for all $k \geq 0$ and $j \geq 0$ we can find $D_{k,j}$ which satisfies*

$$|\partial_t^k \partial_\omega^j A(t,\omega)| \leq D_{k,j}\, l^{j-k}(t). \tag{36}$$

If for some $\alpha < \frac{1}{2}$ and a constant A

$$\forall (t,u) \in \mathbb{R}^2 \quad , \quad |l(t) - l(u)| \leq A|t-u|^{\alpha}, \tag{37}$$

and if

$$\inf_{t \in \mathbb{R}} l(t) > 0 \quad , \tag{38}$$

then

$$X(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(t,\omega) e^{-i\omega t} d\hat{Z}(\omega).$$

is a locally stationary process.

The proof of this theorem is given in appendix C. A simple class of time-varying filters L is obtained by varying the scale, amplitude and frequency modulation of a linear filter. Let $h(v)$ be the impulse response of a time-invariant filter whose Fourier transform $\hat{h}(\omega)$ is concentrated at low frequencies. We construct a filter L whose time-varying impulse response is

$$K(t,v) = \frac{a(t)}{\sigma(t)} h(\frac{v}{\sigma(t)}) \cos(\xi(t)v). \tag{39}$$

The Fourier transform of $K(t,v)$ with respect to v is

$$A(t,\omega) = a(t) \left(\hat{h}[\sigma(t)(\omega - \xi(t))] + \hat{h}[\sigma(t)(\omega + \xi(t))] \right). \tag{40}$$

A Gaussian process obtained by filtering a Gaussian white noise can be written

$$X(t) = \int_{-\infty}^{+\infty} K(t, t-s) dZ(s) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} A(t,\omega) e^{i\omega t} d\hat{Z}(\omega), \tag{41}$$

where $Z(t)$ and $\hat{Z}(\omega)$ are Wiener processes.

To guarantee that $X(t)$ is locally stationary, we suppose that $h(t)$ is a Schwartz function but we must also impose some smoothness conditions on $a(t)$, $\sigma(t)$ and $\xi(t)$. If $a(t)$ and $\xi(t)$ are constant and if for all $k > 1$

$$|\partial_t^k \sigma(t)| \leq |\partial_t \sigma(t)| \leq 1$$

then it can be verified that the conditions (36) are satisfied with

$$l(t) = \frac{K_0}{\sup_{\omega \in \mathbb{R}} |\partial_t A(t,\omega)|} = \frac{K_1 \sigma(t)}{|\partial_t \sigma(t)|},$$

as long as $|\partial_t \sigma(t)| > \epsilon > 0$ and $|\sigma(t)| > \epsilon$ for some $\epsilon > 0$. The constants $B_{k,j}$ are then uniformly bounded for all k and j.

Figure 4 shows one realization of such a locally stationary Gaussian process $X(t)$. The amplitude $a(t)$ is a constant window inside $[0,1]$, with a smooth increasing profile beginning at $t = 0$, and a smooth decreasing profile ending at $t = 1$. The frequency shift $\xi(t) = \xi$ is constant. The filter impulse response is a Gaussian $h(t) = \frac{1}{\sqrt{\pi}} e^{-\frac{t^2}{2}}$. It is scaled by $\sigma(t)$ which increases on $[0, 1]$. As a result, the Fourier transform $\hat{h}(\omega)$ of $h(t)$ is scaled by a decreasing factor $\frac{1}{\sigma(t)}$. The integral (41) over time is discretized over $M = 1024$ samples for discrete calculations.

The bottom of Figure 4 gives the time-varying spectrum $\Lambda_0(t, \omega)$. Only positive frequencies $\omega \geq 0$ are shown. For fixed time t, along ω the function $\Lambda_0(t, \omega)$ is similar to a Gaussian centered at $\omega = \xi$ and scaled by $\frac{1}{\sigma(t)}$. At early times t, $\Lambda_0(t, \omega)$ is a wide because $\sigma(t)$ is close to zero. As $\sigma(t)$ increases, the bandwidth of $\Lambda_0(t, \omega)$ decreases. For t in the neighborhood of 0 and 1 $\Lambda_0(t, \omega)$ is nearly zero because the amplitude $a(t)$ is close to zero.

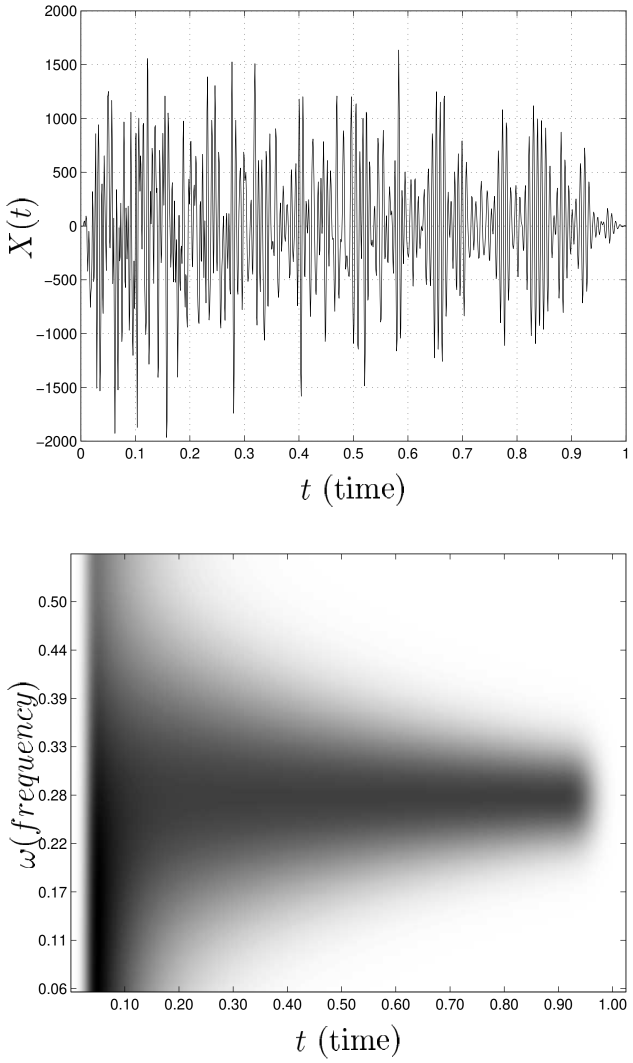

FIG. 4. The graph at the top shows one realization of a locally stationary process generated by filtering a Gaussian white noise. The image at the bottom displays the time-varying spectrum $\Lambda_0(t, \omega)$. The darker the image the larger $\Lambda_0(t, \omega)$.

3. Estimation of Covariance Operators. For general non-stationary processes the covariance matrix cannot be estimated reliably from a few realizations of the process. However, if we can find a basis in which the covariance operator is well approximated by a sparse matrix, it is possible to reduce substantially the variance by estimating only the (essentially) non-zero matrix elements. For example, locally stationary processes have covariances that are well approximated by a sparse matrix in an appropriate local cosine basis, whose windows depend on the size $l(t)$ of the intervals of stationarity. However, we do not know in advance $l(t)$, in general. It is thus necessary to estimate from the data the basis in which the covariance operator is well approximated by a sparse matrix as well as the non-zero matrix elements. We study this problem here in its full generality and present a best basis search algorithm which optimizes an additive measure of departure from being sparse. To simplify the explanations, we suppose that the sparse matrix is a band or near diagonal matrix, although this condition is not required in the best basis search.

3.1. Approximation of Covariance Operators. From N independent realizations $X^k(t)$, $k = 1, 2, ..., N$, of a zero mean process $X(t)$, we want to get an estimate $\tilde{T}$ of the covariance operator T with small bias and mean square error $E\{\|T - \tilde{T}\|_s^2\}$. By controlling the operator norm $\|T - \tilde{T}\|_s$, we also bound the maximum error between the eigenvalues of the estimated operator $\tilde{T}$ and the true covariance operator T. Let λ_n and $\tilde{\lambda}_n$ be the eigenvalues of T and $\tilde{T}$, respectively. From linear algebra we know that for all n

$$\inf_k |\lambda_n - \tilde{\lambda}_k| \leq \|T - \tilde{T}\|_s. \tag{42}$$

Let $\{\phi_n\}_{n\in\mathbb{N}}$ be an orthonormal basis of $\mathbf{L^2}(\mathbb{R})$. A simple but naive algorithm to compute $\tilde{T}$ is to estimate all the matrix elements

$$a_{n,m} =< T\phi_n, \phi_m >= E\{< X, \phi_n >< X, \phi_m >^*\}$$

with the sample means

$$\tilde{a}_{n,m} = \frac{1}{N}\sum_{k=1}^{N} < X^k, \phi_n >< X^k, \phi_m >^* . \tag{43}$$

The sample mean estimator is clearly unbiased

$$E\{\tilde{a}_{n,m}\} = a_{n,m}.$$

In the Gaussian case its variance is given by the following proposition.

PROPOSITION 3.1. *If $X(t)$ is a Gaussian process then*

$$E\{|\tilde{a}_{n,m}|^2\} = (1 + \frac{1}{N})|a_{n,m}|^2 + \frac{1}{N}a_{n,n}a_{m,m}, \tag{44}$$

and thus

$$E\{|\tilde{a}_{n,m} - a_{n,m}|^2\} = \frac{1}{N}(|a_{n,m}|^2 + a_{n,n}a_{m,m}). \tag{45}$$

Proof:

$$
\begin{aligned}
E\{|\tilde{a}_{n,m}|^2\} &= E\left|\frac{1}{N}\sum_{k=1}^{N} < X^k, \phi_n >< X^k, \phi_m >^*\right|^2 \\
&= \frac{1}{N^2}\sum_{k=1}^{N} E\{< X^k, \phi_n >< X^k, \phi_n >< X^k, \phi_m >^* < X^k, \phi_m >^*\} \\
&+ \frac{1}{N^2}\sum_{\substack{k,l=1 \\ k\neq l}}^{N} E\{< X^k, \phi_n >< X^k, \phi_m >^*\} E\{< X^l, \phi_n >< X^l, \phi_m >^*\}.
\end{aligned} \tag{46}
$$

Each $< X^k, \phi_n >$ are Gaussian random variables and for all k

$$E\{< X^k, \phi_n >< X^k, \phi_m >^*\} = a_{n,m}.$$

If A_1, A_2, A_3, A_4 are Gaussian random variables, one can verify that

$$E\{A_1A_2A_3A_4\} = E\{A_1A_2\}E\{A_3A_4\} + E\{A_1A_3\}E\{A_2A_4\} + E\{A_1A_4\}E\{A_2A_3\}.$$

Applying this to (46) yields

$$E\{|\tilde{a}_{n,m}|^2\} = \frac{1}{N^2}N(a_{n,n}a_{m,m} + 2a_{n,m}^2) + \frac{1}{N^2}(N^2 - N)a_{n,m}^2$$

which proves (44). Since $E\{|\tilde{a}_{n,m} - a_{n,m}|^2\} = E\{\tilde{a}_{n,m}^2\} - E\{a_{n,m}^2\}$, we get (45). □

Let $\tilde{T}$ be the covariance operator estimate whose matrix elements in $\{\phi_n\}_{n\in\mathbb{N}}$ are

$$< \tilde{T}\phi_n, \phi_m >= \tilde{a}_{n,m}.$$

The matrix elements of the error $\tilde{T} - T$ are $\tilde{a}_{n,m} - a_{n,m}$. The previous proposition shows that if $X(t)$ is Gaussian then $E\{(\tilde{a}_{n,m} - a_{n,m})^2\}$ does not depend only on $a_{n,m}$ but also on the diagonal elements $a_{n,n}$ and $a_{m,m}$. Thus, even though $a_{n,m}$ may decay quickly to zero when $|n - m|$ increases, since

$$E\{(\tilde{a}_{n,m} - a_{n,m})^2\} \geq \frac{a_{n,n}a_{m,m}}{N}, \tag{47}$$

the expected error remains large if the diagonal coefficients are large. The errors $\tilde{a}_{n,m} - a_{n,m}$ for the matrix elements accumulate and give a very large operator norm error $E\{\|T - \tilde{T}\|_s^2\}$.

To avoid this accumulation of error, we approximate T with the estimated coefficients in a band of size K around the diagonal. Let B_K be the band operator obtained by setting to zero all matrix elements $a_{n,m}$ of T with $|n - m| > K$

$$< B_K\phi_n, \phi_m >= \begin{cases} a_{n,m} & \text{if } |n - m| \leq K \\ 0 & \text{otherwise} \end{cases}$$

The estimated matrix elements in this band of width $2K + 1$ define an estimated band operator

$$< \tilde{B}_K\phi_n, \phi_m >= \begin{cases} \tilde{a}_{n,m} & \text{if } |n - m| \leq K \\ 0 & \text{otherwise} \end{cases}$$

Since $E\{\tilde{a}_{n,m}\} = a_{n,m}$, we derive

$$E\{\tilde{B}_K\} = B_K.$$

The error when estimating T with $\tilde{B}_K$ is the sum of the bias due to the difference between T and B_K and the variance of the estimator of B_K

$$E\{\|T - \tilde{B}_K\|_s^2\} = \|T - B_K\|_s^2 + E\{\|B_K - \tilde{B}_K\|_s^2\}.$$

The expected norm $E\{\|B_K - \tilde{B}_K\|_s^2\}$ varies typically like $\frac{(2K+1)^2}{N}$. Indeed, $E\{(a_{n,m} - \tilde{a}_{n,m})^2\}$ is proportional to N^{-1} and (47) shows that these coefficients do not decay away from the diagonal, within the band. The squared norm is thus proportional to the band width squared $(2K+1)^2$. This shows that the variance term increases when K increases. On the other hand, the bias $\|T - B_K\|_s^2$ decreases when K increases since the approximation band gets larger. Given the number of realizations N, an optimal choice for K is obtained by balancing the bias and variance terms. When N is very small, which is the case in many applications, the best choice is often $K = 0$ because the variance term dominates.

3.2. Best Basis Selection. The covariance operators of some processes may be well approximated by a band matrix in a particular basis that is chosen from a limited collection of bases, called a dictionary. For locally stationary processes, this dictionary is the collection of local cosine bases constructed with windows of varying sizes.

Let $\mathcal{D} = \{\mathcal{B}^\gamma\}_{\gamma\in\Gamma}$ be a dictionary of orthonormal bases $\mathcal{B}^\gamma = \{\phi_n^\gamma\}_{n\in\mathbb{N}}$ of $\mathbf{L^2}(\mathbb{R})$, indexed by $\gamma \in \Gamma$. We denote the matrix elements of T in $\mathcal{B}^\gamma$ by

$$a_{n,m}^\gamma =< T\phi_n^\gamma, \phi_m^\gamma > .$$

Let B_K^γ be the restriction of the operator T to a band of size $2K+1$ in the basis $\mathcal{B}^\gamma$

$$< B_K^\gamma \phi_n^\gamma, \phi_m^\gamma >= \begin{cases} a_{n,m}^\gamma & \text{if } |n-m| \le K \\ 0 & \text{otherwise} \end{cases}$$

Given a covariance operator, we would like to find the basis $\mathcal{B}^\alpha$ in the dictionary which minimizes the bias $\|T - B_K^\alpha\|_s$ so as to reduce the total estimation error

$$E\{\|T - \tilde{B}_K^\alpha\|_s^2\} = \|T - B_K^\alpha\|_s^2 + E\{\|\tilde{B}_K^\alpha - B_K^\alpha\|_s^2\}$$

However, the bias $\|T - B_K^\alpha\|_s$ cannot be computed directly since we do not know T. We must therefore try to control this bias from the band coefficients $\tilde{a}_{n,m}^\gamma$ of $\tilde{B}_K^\alpha$. This can be done by using a Hilbert-Schmidt norm.

The Hilbert-Schmidt norm of the operator T is the trace of TT^t and it is therefore equal to $\mathbf{L^2}(\mathbb{R}^2)$ norm of its kernel that we suppose to be finite

$$\|T\|_h^2 = \mathrm{tr}(TT^t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |R(t,s)|^2 dt ds < +\infty.$$

One can verify that the Hilbert-Schmidt norm of T can be also written as the sum of its matrix coefficients squared in any orthonormal basis $\mathcal{B}^\gamma$

$$\|T\|_h^2 = \sum_{n,m} |a_{n,m}^\gamma|^2. \tag{48}$$

Applying the Cauchy-Schwartz inequality on the expression (1) of Tf proves that the sup operator norm of T is bounded by its Hilbert-Schmidt norm

$$\|T\|_s \leq \|T\|_h. \tag{49}$$

Inequality (49) shows that we can control the bias $\|T - B_K^\gamma\|_s^2$ by a Hilbert-Schmidt norm

$$\|T - B_K^\gamma\|_s^2 \leq \|T - B_K^\gamma\|_h^2 = \sum_{\substack{n,m \\ |n-m|>K}} |a_{n,m}^\gamma|^2,$$

and hence

$$\|T - B_K^\gamma\|_s^2 \leq \|T\|_h^2 - \sum_{\substack{n,m \\ |n-m|\leq K}} |a_{n,m}^\gamma|^2. \tag{50}$$

To minimize this upper bound we choose among the dictionary the basis that maximizes

$$\|B_K^\gamma\|_h^2 = \sum_{\substack{n,m \\ |n-m|\leq K}} |a_{n,m}^\gamma|^2. \tag{51}$$

It is important to realize that the Hilbert-Schmidt norm $\|T - B_K^\gamma\|_h$ is often a crude upper bound for $\|T - B_K^\gamma\|_s$. In general, minimizing $\|T - B_K^\gamma\|_h$ is therefore not equivalent to minimize $\|T - B_K^\gamma\|_s$. However, Schur lemma A.1 shows that

$$\|T - B_K^\gamma\|_h^2 - \sum_{\substack{n,m \\ |n-m|>K}} |a_{n,m}^\gamma|^2$$

provides an effective control on $\|T - B_K^\gamma\|_s$ if we are also guaranteed that the coefficients $a_{n,m}^\gamma$ have a fast off-diagonal decay as $|n-m|$ increases. This will be the case when approximating locally stationary processes in local cosine bases. The maximization of $\|B_K^\gamma\|_h$ then selects a basis in which the operator norm $\|T - B_K^\gamma\|_s$ is small.

Given N realizations of the process $X(t)$, we compute sample mean estimates $\tilde{a}_{n,m}^\gamma$ (43) of the coefficients $a_{n,m}^\gamma$ in the band of B_K^γ. It defines an estimated band operator $\tilde{B}_K^\gamma$. The Hilbert-Schmidt norm $\|B_K^\gamma\|_h^2$ is then estimated by

$$\|\tilde{B}_K^\gamma\|_h^2 = \sum_{\substack{n,m \\ |n-m|\leq K}} |\tilde{a}_{n,m}^\gamma|^2.$$

If $X(t)$ is a Gaussian process then (44) shows that

$$E\{\|\tilde{B}_K^\gamma\|_h^2\} = \sum_{\substack{n,m \\ |n-m|\leq K}} \left\{(1 + \frac{1}{N})|a_{n,m}|^2 + \frac{a_{n,n}a_{m,m}}{N}\right\}$$

and hence

$$E\{\|\tilde{B}_K^\gamma\|_h^2\} = (1 + \frac{1}{N})\|B_K^\gamma\|_h^2 + \frac{1}{N} \sum_{\substack{n,m \\ |n-m|\leq K}} a_{n,n}a_{m,m}. \tag{52}$$

The estimate $\|\tilde{B}_K^\gamma\|_h^2$ is biased but its maximization is a reasonable procedure for maximizing $\|B_K^\gamma\|_h^2$. We will denote by $\mathcal{B}^{\tilde{\alpha}}$ the estimated "best" basis which maximizes the estimated sum of squares of matrix elements in the band of size K

$$\|\tilde{B}_K^{\tilde{\alpha}}\|_h = \sup_{\gamma\in\Gamma} \|\tilde{B}_K^\gamma\|_h.$$

The variable $\tilde{\alpha}$ labels the estimated best basis. It is a random variable since it is a functional of the observations.

3.3. Approximate Karhunen-Loeve Basis. As mentioned earlier, when the number of realizations N is small the variance term $\|\tilde{B}_K^\gamma - B_K^\gamma\|_s$ of the mean square error grows like $\frac{2K+1}{\sqrt{N}}$ and is often much larger than $\|T - B_K^\gamma\|_s$. To reduce the variance, therefore, we often choose $K = 0$. We let $D^\gamma = B_0^\gamma$ and $\tilde{D}^\gamma = \tilde{B}_0^\gamma$ be the resulting diagonal matrices. The basis $\mathcal{B}^\alpha$ which minimizes $\|T - D^\gamma\|_s$ can be interpreted as the best approximation, within the dictionary of bases, of a Karhunen-Loeve basis. A Karhunen-Loeve basis is indeed a basis in which the covariance operator T is diagonal. If the dictionary $\mathcal{D}$ includes a Karhunen-Loeve basis then $\|D^\gamma\|_h$ is maximized by this Karhunen-Loeve basis. The approximation of a Karhunen-Loeve basis with a "best" basis selected from a limited dictionary has already been studied by Coifman and Wickerhauser [6]. Their searching algorithm maximizes a different criterion, based on an entropy measure, which is not, however, directly related to the norm of the error $\|T - D^\gamma\|_s$.

Let $d_n^\gamma = a_{n,n}^\gamma$ and $\tilde{d}_n^\gamma = \tilde{a}_{n,n}^\gamma$ be the diagonal coefficients of D^γ and $\tilde{D}^\gamma$. The Hilbert-Schmidt norm is the sum of the diagonal elements squared

$$\|\tilde{D}^\gamma\|_h^2 = \sum_n |\tilde{d}_n|^2.$$

When $K = 0$, (52) shows that the expected trace norm of the estimated diagonal coefficients (in a fixed basis) is

$$E\{\|\tilde{D}^\gamma\|_h^2\} = (1 + \frac{2}{N})\|D^\gamma\|_h^2. \tag{53}$$

The maximization of $\|\tilde{D}^\gamma\|_h^2$ is thus equivalent to the maximization of an unbiased estimator of $\|D^\gamma\|_h^2$.

Let $\mathcal{B}^{\tilde{\alpha}} = \{\phi_n^{\tilde{\alpha}}\}_{n\in\mathbb{N}}$ be the estimated best basis which maximizes $\|\tilde{D}^\gamma\|_h^2$. Since $\tilde{D}^{\tilde{\alpha}}$ is a diagonal matrix, its diagonal entries $\tilde{d}_n$ are the estimated eigenvalues of T. Note that for $K = 0$, we are guaranteed that the estimated covariance operator $\tilde{D}^{\tilde{\alpha}}$ is a positive operator, which is not always the case if $K > 0$.

In the diagonal case, the estimated time-varying spectrum is easily calculated from the Wigner-Ville distribution of each basis vector. Indeed, the estimated covariance is

$$\tilde{R}^{\tilde{\alpha}}(t,s) = \sum_n \tilde{d}_n \; \phi_n^{\tilde{\alpha}}(t)\phi_n^{\tilde{\alpha}}(s)$$

and the corresponding time-varying spectrum is

$$\tilde{\Lambda}_0(u,\omega) = \int_{-\infty}^{+\infty} \tilde{R}^{\tilde{\alpha}}(u+\frac{v}{2}, u-\frac{v}{2})e^{-i\omega v}dv.$$

Inserting the Wigner-Ville distribution $W\phi_n^{\tilde{\alpha}}(u,\omega)$ defined in (11) yields

$$\tilde{\Lambda}_0(u,\omega) = \sum_n \tilde{d}_n \; W\phi_n^{\tilde{\alpha}}(u,\omega). \tag{54}$$

4. Basis Selection and Estimation Algorithms. Theorem 2.2 proves that the covariance operators of locally stationary processes are well approximated by band matrices in a local cosine basis where the size of the windows is adapted to the size $l(t)$ of intervals over which the process is approximately stationary. We introduce a dictionary $\mathcal{D}$ of local cosine bases with windows of varying sizes. From a few realizations of the process we search in this dictionary for the best approximate Karhunen Loeve basis, as described in section 3.3. To implement this search with a fast algorithm we use the tree structured dictionary introduced by Coifman and Meyer [9]. In section 4.1 we describe this local cosine tree and in section 4.2 we give some numerical results for covariance estimation.

4.1. Local Cosine Binary Trees. To reduce the complexity of the search for a best local cosine basis with adapted window sizes, we limit the window sizes to powers of 2. We consider signals and processes with compact support included in $[0, M]$. Local cosine bases with dyadic window sizes are constructed along a binary tree. We consider separately the dictionaries of local cosine bases for continuous time and discrete time signals.

Coifman and Wickerhauser [6] construct a dyadic tree of local cosine bases by associating to each node of the tree a window that covers a sub-interval of $[0, M]$. The root of the tree corresponds to a window which covers the whole interval $[0, M]$. The left and right branch nodes are associated with the two half windows which cover $[0, \frac{M}{2}]$ and $[\frac{M}{2}, M]$, respectively. Each of these windows are divided further into a left and right window of half their size, and so on. Each node of the binary tree is characterized by the pair (j, p) which specifies its depth j and its position p from left to right, at depth j. Such a node corresponds to the window function $g_p^j(t)$ which covers the interval $[pM2^{-j}, (p+1)M2^{-j}]$, as illustrated in Figure 5. All window functions $g_p^j(t)$ have an increasing and decreasing profile constructed by translating a single, smooth function $\beta(t) \geq 0$ such that

$$\beta(t) = \begin{cases} 0 & \text{if } t < -\eta \\ 1 & \text{if } t > \eta, \end{cases}$$

and

$$\beta^2(t) + \beta^2(-t) = 1.$$

The window $g_p^j(t)$ is defined by

$$g_p^j(t) = \begin{cases} \beta(t - pM2^{-j}) & \text{if } t < pM2^{-j} + \eta \\ 1 & \text{if } pM2^{-j} + \eta \leq t < (p+1)M2^{-j} - \eta \\ \beta((p+1)M2^{-j} - t) & \text{if } t > (p+1)M2^{-j} - \eta \end{cases}$$

This is valid only if $M2^{-j} \geq 2\eta$, which limits the maximum depth of the tree to J

$$j \leq J = \log_2 \frac{M}{2\eta}.$$

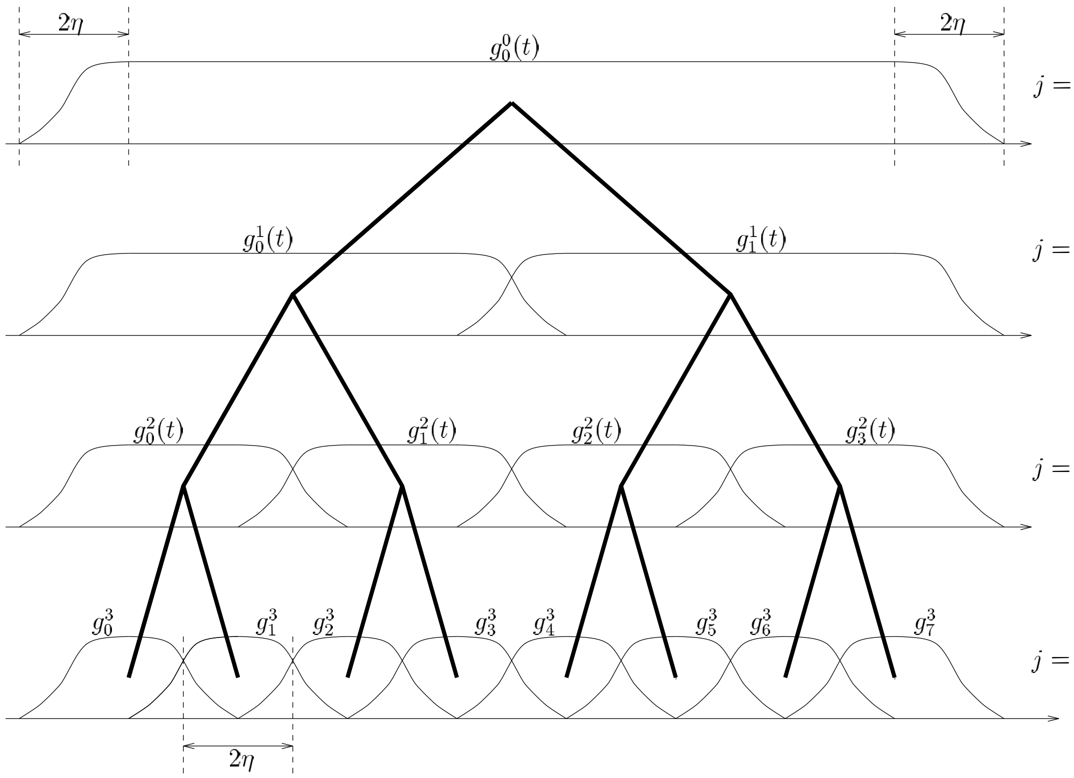

FIG. 5. Dyadic tree of local cosine bases. Each node is associated to a window modulated by cosine functions whose frequencies are inversely proportional to the window length. The leaves of any admissible subtree corresponds to a particular local cosine basis.

To each window we associate a local cosine family defined by

$$\left\{ \phi^j_{p,k}(t) = g^j_p(t)\sqrt{\frac{2}{M2^{-j}}}\cos\left[\pi(k+\frac{1}{2})\frac{t-M2^{-j}p}{M2^{-j}}\right]\right\}_{k\in\mathbb{N}}.$$

We call admissible binary tree any binary tree whose nodes have either 0 or 2 branches. We denote by γ the index set of the nodes (j,p) of a particular admissible binary tree. One can verify that the windows $\{g^j_p(t)\}_{(j,p)\in\gamma}$ define a partition of the interval $[0,M]$ into dyadic intervals of varying sizes. Figure 6 gives two examples of admissible binary trees and their corresponding window decomposition of the interval $[0,M]$. It can be shown from the local cosine theorem 2.1 that for any admissible binary tree indexed by γ

$$\mathcal{B}^\gamma = \{\phi^j_{p,k}(t)\}_{(j,p)\in\gamma,k\in\mathbb{N}}$$

is an orthogonal basis in a space $\mathbf{V}$ which includes $\mathbf{L^2}([\eta, M-\eta])$. The dictionary $\mathcal{D} = \{\mathcal{B}^\gamma\}_{\gamma\in\Gamma}$ of local cosine bases constructed with all admissible binary trees puts the local cosine bases in correspondence with all combinations of dyadic size windows that make an exact cover of $[0,M]$. There are more than $2^{J/2}$ different admissible binary trees of depth at most J and hence the dictionary $\mathcal{D} = \{\mathcal{B}^\gamma\}_{\gamma\in\Gamma}$ contains more than $2^{J/2}$ different local cosine basis.

Orthogonal bases for discrete time signals are obtained simply by discretizing the local cosine functions. It can be shown that for $m = 1, 2, ..., M$,

$$\left\{ \phi^j_{p,k}[m] = g^j_p(m)\sqrt{\frac{2}{M2^{-j}}}\cos\left[\pi(k+\frac{1}{2})\frac{m-2^{-j}p}{M2^{-j}}\right]\right\}_{0\le k<M2^{-j}}$$

is an orthogonal family of discrete cosine vectors. For any admissible binary tree whose branches have indices (j,p) in a set γ, one can also prove that

$$\mathcal{B}^\gamma = \{\phi^j_{p,k}[m]\}_{(j,p)\in\gamma,0\le k<M2^{-j}}$$

is an orthogonal family of M discrete vectors. It is an orthogonal basis in the space $\mathbf{V}$ which contains discrete signals having compact support in $[\eta, M-\eta]$. Since $\eta > 1$, the binary tree has depth

$$J = \log_2\frac{M}{2\eta} \le \log_2 M.$$

At depth j of the binary tree there are 2^j families of local cosine vectors $\{\phi^j_{p,k}[m]\}_{0\le k<M2^{-j}}$, which are a rearrangement of a total of M cosine vectors. By using a fast discrete cosine transform, for any discrete signal $f[m]$ whose support is in $[\eta, M-\eta]$ all inner products

$$\{< f, \phi^j_{p,k} >\}_{0\le p<2^j, 0\le k<M2^{-j}}$$

are calculated with $O(M\log_2 M)$ operations. To compute all discrete cosine products at all depth $0\le j\le J$ of the binary tree thus requires $O(JM\log_2 M)$ operations.

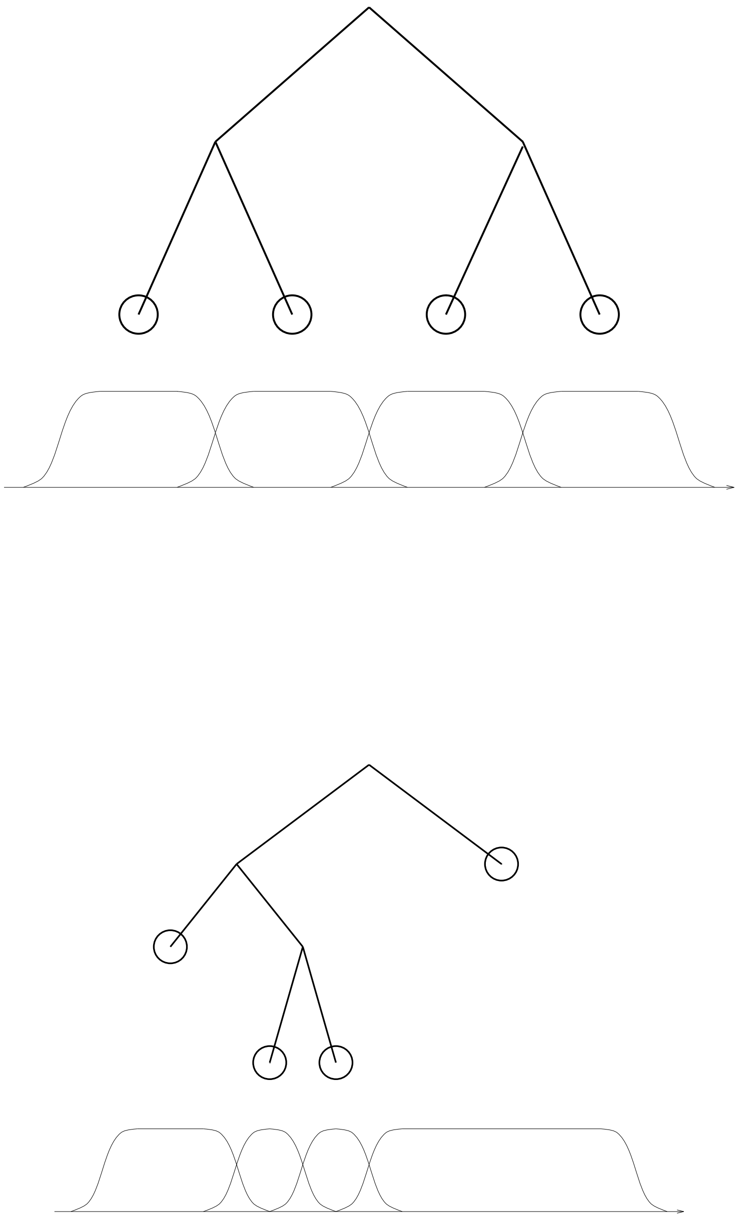

FIG. 6. Examples of admissible binary trees corresponding to two partitions of the interval with windows of varying sizes. The circles indicate the selected nodes. The resulting windows are drawn under the binary trees.

4.2. Best Local Cosine Basis Search. Let $X[m]$ be the samples of a locally stationary process whose support is included in $[\eta, M-\eta]$. Let us consider the dictionary $\mathcal{D}$ of discrete local cosine bases constructed with a binary tree of depth J. We search in the dictionary $\mathcal{D}$ for the best approximate Karhunen-Loeve basis as described in section 3.3.

For each realization $X^q[m]$, we compute all inner products with the JM cosine vectors stored in the binary tree

$$\{< X^q, \phi^j_{p,k} >\}_{0\leq j\leq J, 0\leq p<2^j, 0\leq k<M2^{-j}}$$

with $O(JM\log_2 M)$ operations. We estimate the diagonal covariance matrix elements for each cosine vector with the sample mean

$$\tilde{d}^j_{p,k} = \frac{1}{N}\sum_{q=1}^{N} | < X^q, \phi^j_{p,k} > |^2.$$

To each local cosine basis $\mathcal{B}^\gamma = \{\phi^j_{p,k}\}_{(j,p)\in\gamma, 0\leq k<M2^{-j}}$, corresponding to an admissible binary tree indexed by γ, we associate the diagonal matrix $\tilde{D}^\gamma$ whose diagonal elements are the estimated ones

$$\{\tilde{d}^j_{p,k}\}_{(j,p)\in\gamma, 0\leq k<M2^{-j}}.$$

The best basis maximizes the sum of the squares of these M diagonal coefficients

$$\|\tilde{D}^\gamma\|^2_h = \sum_{\substack{(j,p)\in\gamma \\ 0\leq k<M2^{-j}}} |\tilde{d}^j_{p,k}|^2.$$

Since $\|\tilde{D}^\gamma\|^2_h$ is an additive quantity over the local cosine coefficients of an admissible binary tree, we can use the fast dynamic programming algorithm of Coifman and Wickerhauser to find the best basis (admissible binary tree) which maximizes it. The dynamical programming algorithm uses a bottom up strategy which progressively constructs the best admissible tree by comparing the energy of the estimated local cosine coefficients of a tree node and its two branches. The best basis $\mathcal{B}^{\tilde{\alpha}}$ is found with $O(M\log_2 M)$ operations.

To guarantee that a local cosine basis compresses the covariance operator of a locally stationary process, the proof of theorem 2.3 indicates that one must also insure that the local cosine windows $g_p(t)$ have smooth rising and decaying profiles. These profiles should vary over intervals of size $2\eta_p$ and $2\eta_{p+1}$, comparable to the length of the interval $[a_p, a_{p+1}]$ covered by $g_p(t)$. This is a priori not satisfied by the windows included in the binary tree, which all have rising and decaying intervals of the same length, equal to 2η. This constraint is necessary in order to freely combine any window with any other one when constructing a local cosine basis. The parameter 2η is the minimum window size at the bottom of the binary tree. It is thus typically small compared to M. This means that the large windows at the top of the binary tree have rising and decaying intervals that are much smaller than the window size that they cover (see Figure 7). Clearly, these window functions are not as smooth as they could be. To by-pass this constraint, once the best basis $\mathcal{B}^{\tilde{\alpha}}$ is selected, we modify the rising and decaying profiles of the windows to increase their smoothness. The best basis choice decomposes the interval $[0, M]$ in dyadic size intervals that we denote by $[a_p, a_{p+1}]$. Over these best basis intervals, we construct a new local cosine basis with non-symmetric windows whose profiles rise and decay over the largest possible intervals compatible with the constraints imposed by the neighboring windows. The construction of these windows is specified at the beginning of appendix B by (67,68). It is illustrated in Figure 7. The estimated variance matrix elements are recomputed with this new basis by decomposing again the N realization of the process in this modified best basis. The diagonal operator in this new best basis is still denoted by $\tilde{D}^{\tilde{\alpha}}$.

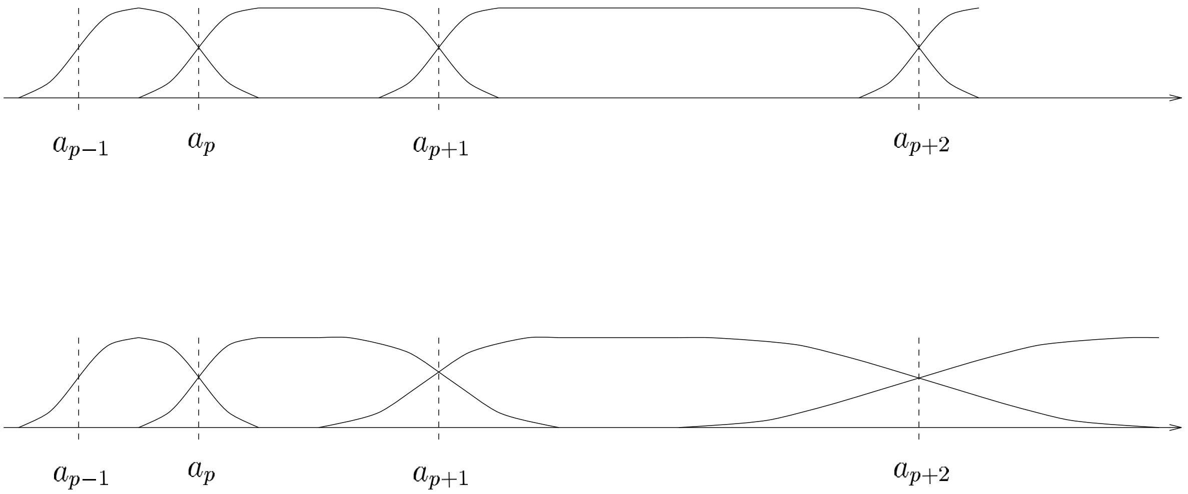

FIG. 7. The figure at the top gives an example of windows for a local cosine basis. The figure at the bottom shows how to dilate the rising and decaying profiles to obtain windows of maximum smoothness, while maintaining the necessary properties for local cosine bases.

4.3. Numerical Experiments. The algorithm is tested with a locally stationary process synthesized by filtering a Gaussian white noise through a time-varying filter specified by (39). Figure 4 shows one realization of this locally stationary process and its time-varying spectrum $\Lambda_0(t, \omega)$.

Equation (45) for $n = m$ proves that the error when estimating the diagonal covariance coefficients from N realizations of the process is

$$E\{|\tilde{d}^j_{p,k} - d^j_{p,k}|^2\} = \frac{2|d^j_{p,k}|^2}{N}. \tag{55}$$

A first experiment is performed with $N = 1000$ realizations in order to get accurate estimations of these coefficients, $\tilde{d}^j_{p,k} \approx d^j_{p,k}$. The time-frequency tiling of estimated the best basis is shown in Figure 8. Each rectangle indicates the time-frequency support of a local cosine window $\phi^j_{p,k}$ in the selected best basis $\mathcal{B}^{\tilde{\alpha}}$. The gray level of these rectangles gives the value of $\tilde{d}^j_{p,k}$. The darker the rectangle the larger $\tilde{d}^j_{p,k}$. The window sizes are adapted to the time and frequency variations of $\Lambda_0(t, \omega)$ that is shown at the bottom of Figure 9. The smoother the time variation of $\Lambda_0(t, \omega)$ the larger the time support of the local cosine windows. For t close to zero the frequency bandwidth of $\Lambda_0(t, \omega)$ decreases quickly which requires short time windows. As the rate of modification of this bandwidth decreases, the windows increase in size. For t close to 0 and 1 the amplitude of $\Lambda_0(t,\omega)$ has a rapid decay to zero which selects short time windows.

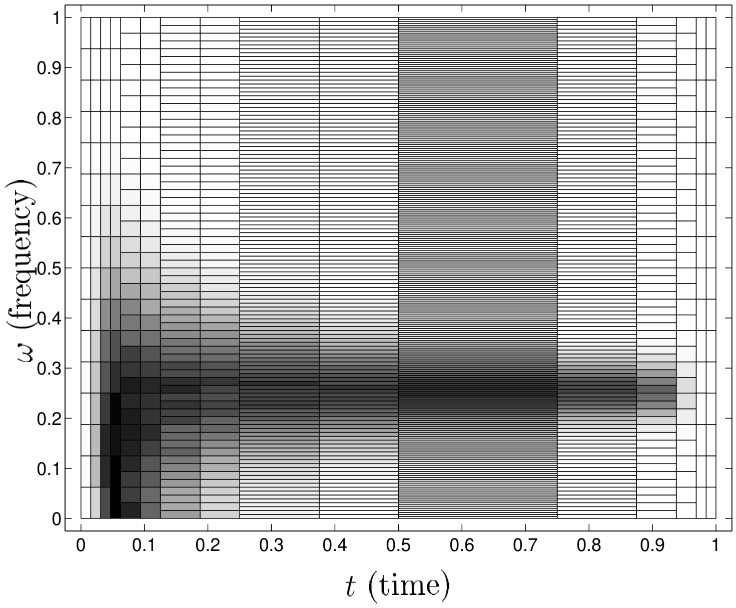

FIG. 8. Time-frequency tiling of the estimated best basis computed with 1000 realizations of the process. The width and height of each rectangle indicates the time and frequency spread of the a cosine window $\phi^j_{p,k}$. The darkness is proportional to estimated variance $\tilde{d}^j_{p,k}$. The distribution of is very similar to the time-varying spectrum $\Lambda_0(t,\omega)$ of the process displayed at the bottom of figure 9.

From the estimated diagonal covariance operator $\tilde{D}^{\tilde{\alpha}}$ we compute an estimated time-varying spectrum $\tilde{\Lambda}_0(t,\omega)$ with (54). The left image of Figure 9 is the estimated spectrum $\tilde{\Lambda}_0(t,\omega)$ obtained with the original local cosine windows having the same rising and decaying profiles, as illustrated at the top of Figure 7. The right image of Figure 9 is the estimated spectrum $\tilde{\Lambda}_0(t,\omega)$ computed after modifying the local cosine windows, as indicated at the bottom of Figure 7. Both spectra have the same qualitative behavior as the original time-varying spectrum $\Lambda_0(t,\omega)$ given in Figure 8. The errors are mostly concentrated in the time regions where the rising and decreasing profiles of the windows are located. The modified windows that are smoother reduce this error.

In most applications we must estimate the covariance from very few realizations. In speech processing, we only have 1 realization. The top of figure 10 shows the time-frequency tiling of the best basis computed with only $N = 1$ realization of the process $X(t)$. The gray level of the rectangles indicate the value of the estimated diagonal covariance coefficients $\tilde{d}^j_{p,k}$. In this case (55) proves that the expected estimation error is

$$E\{|\tilde{d}^j_{p,k} - d^j_{p,k}|^2\} = 2|d^j_{p,k}|^2.$$

This explains the considerable variation of $\tilde{d}^j_{p,k}$ in time-frequency regions where Figure 8 shows that $d^j_{p,k}$ is approximately constant. Next section explains how to reduce this variations with a time-frequency smoothing. The considerable variance on the covariance coefficient estimators also induces a large variance on the estimator $\mathcal{B}^{\tilde{\alpha}}$ of the best local cosine basis. Compared to Figure 8 we see that the selected window sizes is not optimal.

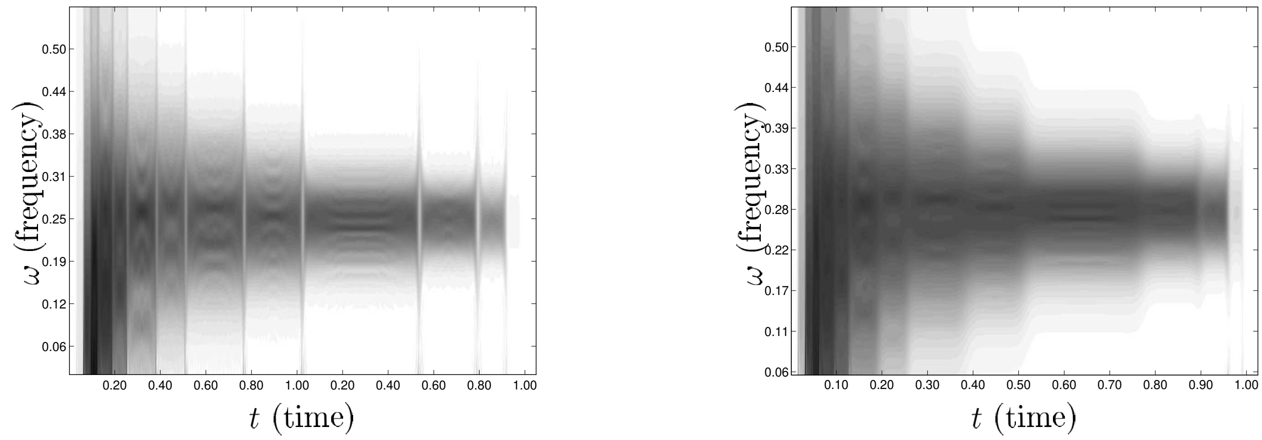

FIG. 9. The left image is the estimated time-varying power spectrum $\tilde{\Lambda}_0(t,\omega)$ in the best local cosine basis. The right image displays $\tilde{\Lambda}_0(t,\omega)$ computed in the same best basis, with modified maximally smooth windows.

Table 1 gives the expected estimation errors of the covariance operator for different numbers of realizations. Observe that

$$E\{\|T - \tilde{D}^{\tilde{\alpha}}\|_s^2\} \approx E\{\|T - D^{\tilde{\alpha}}\|_s^2\} + E\{\|D^{\tilde{\alpha}} - \tilde{D}^{\tilde{\alpha}}\|_s^2\}. \tag{56}$$

This indicates that the error $T - D^{\tilde{\alpha}}$ when approximating T by its diagonal restriction in the estimated best basis $\mathcal{B}^{\tilde{\alpha}}$ is uncorrelated with the error $D^{\tilde{\alpha}} - \tilde{D}^{\tilde{\alpha}}$ produces by the estimation of the diagonal coefficients in the estimated best basis. As expected, $E\{\|D^{\tilde{\alpha}} - \tilde{D}^{\tilde{\alpha}}\|_s^2\}$ is inversely proportional to the number of realizations N. The best basis diagonal approximation $E\{\|T - D^{\tilde{\alpha}}\|_s^2\}$ also decreases with N, which means that we do get more reliable estimates of the true best basis when the number of realizations increases. This value tends to $\|T - D^{\alpha}\|_s^2$ which is the error in the true best basis $\mathcal{B}^{\alpha}$. However, beyond these numerical results, we have no theoretical control on the convergence of the error in the estimated best basis compared to the error in the true best basis, when the number of realizations increases. For a number of realizations $N \leq 20$, $E\{\|T - D^{\tilde{\alpha}}\|_s^2\}$ is negligible compared to $E\{\|D^{\tilde{\alpha}} - \tilde{D}^{\tilde{\alpha}}\|_s^2\}$. This means that the error introduced by approximating the Karhunen-Loeve basis with the best local cosine basis is negligible compared to the error due to the estimation of the diagonal coefficients.

We mentioned that a naive estimation $\tilde{T}$ of T may be obtained by estimating all the matrix coefficients in a basis arbitrarily chosen, say a discrete Dirac basis. This is equivalent to compute the covariance function $R(t,s)$ directly with the sample mean

$$\tilde{R}(t,s) = \frac{1}{N}\sum_{k=1}^{N} X^k(t)X^k(s). \tag{57}$$

The resulting error $E\{\|T - \tilde{T}\|^2\}$ is proportional to $\frac{1}{N}$ multiplied by the full covariance matrix size M^2, which is huge. The first column of table 1 gives $E\{\|T - \tilde{D}^{\tilde{\alpha}}\|_s^2\}$ for N realizations. As expected, this error is much larger than the error $E\{\|T - D^{\tilde{\alpha}}\|_s^2\}$ obtained in an estimated local cosine basis. Next sections explains how to further reduce this error with an appropriate time-frequency smoothing of the estimated covariance coefficients.

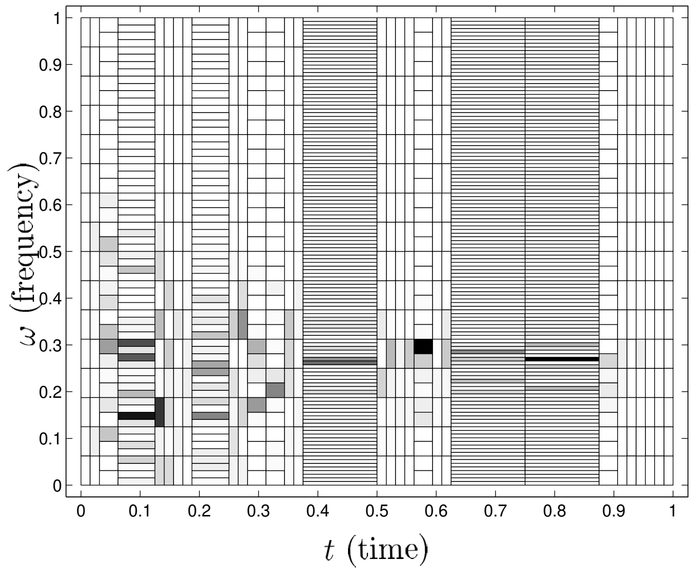

FIG. 10. At the top is the time-frequency tiling of the best basis computed with $N = 1$ realization. The darkness of each rectangle is proportional to the estimated variance $\tilde{d}^j_{p,k}$. The bottom displays the values of the smoothed coefficients $\hat{d}^j_{p,k}$ computed with a time-frequency averaging of $\tilde{d}^j_{p,k}$.

N	$E\{\|T - \tilde{D}^{\tilde{\alpha}}\|_s^2\}/\|T\|_s^2$	$E\{\|D^{\tilde{\alpha}} - \tilde{D}^{\tilde{\alpha}}\|_s^2\}/\|T\|_s^2$	$E\{\|T - D^{\tilde{\alpha}}\|_s^2\}/\|T\|_s^2$	$E\{\|T - \hat{D}^{\tilde{\alpha}}\|_s^2\}/\|T\|_s^2$	$E\{\|T - \tilde{T}\|_s^2\}/\|T\|_s^2$
1	50	49	0.14	0.6	139
5	4.9	4.8	0.08	0.40	33
10	1.9	1.8	0.08	0.4	18
20	0.81	0.75	0.09	0.37	10
40	0.36	0.32	0.09	0.41	6.2
80	0.15	0.12	0.09	0.35	3.7
160	0.11	0.06	0.08	0.38	2.3
320	0.08	0.03	0.07	0.35	1.5
640	0.07	0.02	0.06	0.31	0.94

TABLE 1

Estimation errors of the covariance operator for different numbers of realizations N. The first column gives the total error in the estimated best local cosine basis. The second column is the error induced by the diagonal coefficient estimation in the best basis. The third colum is the error when approximating T by its diagonal restriction in the estimated best basis. The fourth column is the total error in the estimated best basis with a time-frequency smoothing of the diagonal coefficients. This error is dramatically reduced. The last column gives the error when estimating T with a full matrix in a discrete Dirac basis.

4.4. Time-Frequency Smoothing. The variance error $E\{\|D^{\tilde{\alpha}} - \tilde{D}^{\tilde{\alpha}}\|^2\}$ is the main source of error and can often be reduced with a local averaging of the estimated diagonal coefficients of $\tilde{D}^{\tilde{\alpha}}$. This relies on an a priori assumption of smoothness of the diagonal coefficients of $D^{\tilde{\alpha}}$, which is not always true for all locally stationary processes. We defined locally stationary processes as those whose covariance operators have a fast off-diagonal decay in an appropriate local cosine basis. However, we do not we do not impose a priori any smoothness condition on the matrix coefficients along the diagonal. The same issue appears when estimating the spectrum of stationary processes. These processes are diagonalized in the Fourier basis. To reduce the variance of the spectrum estimation, most spectral estimation algorithms perform some type of averaging of the Fourier coefficients along the frequency axis. This averaging is justified only if the spectrum if smooth, which is not always the case.

The frequency axis gives a natural topology for the spectrum of stationary processes. For locally stationary processes, the natural topology is provided by the time-frequency plane. Local cosine functions are neighbors either in time or in frequency. Time-frequency smoothing kernels for the estimated "time-varying" spectrum $\tilde{\Lambda}_0(t,\omega)$ of non-stationary pro-cesses have been studied by several researchers [5, 13, 14]. In our numerical experiments, we perform a direct averaging of the estimated local cosine coefficients $\tilde{d}^j_{p,k}$. This short study illustrates the result of such an averaging, without any theoretical analysis.

The coefficient $\tilde{d}^j_{p,k}$ is an estimate of $d^j_{p,k} = E\{|<X, \phi^j_{p,k}>|^2\}$. It is averaged with other coefficients $\tilde{d}^{j'}_{p',k'}$ in the same local cosine basis, depending upon the distance in time and frequency of the two local cosine vectors $\phi^j_{p,k}$ and $\phi^{j'}_{p',k'}$

$$\hat{d}^j_{p,k} = \sum_{j',p',k'} w^j_{p,k}[j',p',k']\ \tilde{d}^{j'}_{p',k'}. \tag{58}$$

The weights $w^j_{p,k}[j',p',k']$ decrease when the distance between the time supports of of $\phi^j_{p,k}$ and $\phi^{j'}_{p',k'}$ increases, or when the distance between the support of their Fourier transform increases.

If M is the total number of samples of the signal, $\phi^j_{p,k}$ covers an interval of size $l_p = M2^{-j}$. The distance between the centers of the time support of $\phi^j_{p,k}$ and $\phi^{j'}_{p',k'}$ is thus

$$\Delta_t = M2^{-j}(p + \frac{1}{2}) - M2^{-j'}(p' + \frac{1}{2}).$$

The distance between the domains where their Fourier transform energy is mostly located is

$$\Delta_\omega = |\xi_{p,k} - \xi_{p',k'}| = \pi \left(\frac{k + \frac{1}{2}}{M2^{-j}} - \frac{k' + \frac{1}{2}}{M2^{-j'}} \right).$$

The averaging weights are computed with a Gaussian kernel $g(t)$ that is dilated in time and frequency proportionally to the time and frequency spread of $\phi^j_{p,k}$. The time and frequency scale factors are thus $M2^{-j}$ and $\frac{2\pi}{M2^{-j}}$

$$w^j_{p,k}[j', p', k'] = \lambda^j_{p,k}\ g(\frac{\Delta_t}{M2^{-j}})\ g(\frac{M2^{-j}}{2\pi}\Delta_\omega).$$

The factor $\lambda^j_{p,k}$ normalizes the sum of the weights

$$\sum_{j',p',k'} w^j_{p,k}[j', p', k'] = 1.$$

The variance of the Gaussian kernel $g(t)$ is a parameter that modifies the time-frequency spread of this averaging. The smaller the number of realizations N, the larger the variance of the estimators $\tilde{d}^j_{p,k}$ and the more averaging is needed. This also depends upon the expected time-frequency smoothness of the true coefficients $d^j_{p,k}$. The bottom of Figure 10 displays the amplitude of the smoothed coefficients $\hat{d}^j_{p,k}$ computed from the estimated coefficients $\tilde{d}^j_{p,k}$ for $N = 1$ realizations, shown at the top of Figure 10.

We denote by $\hat{D}^{\tilde{\alpha}}$ the diagonal operator in the basis $\mathcal{B}^{\tilde{\alpha}}$ whose diagonal coefficients are the smoothed estimates $\hat{d}^j_{p,k}$ defined by (58). The next to last column of Table 1 displays the expected error $E\{\|T - \hat{D}^{\tilde{\alpha}}\|^2_s\}$. It is much smaller than $E\{\|T - \tilde{D}^{\tilde{\alpha}}\|^2_s\}$, which shows that this smoothing decreases considerably the expected error when the number of realizations is small. Donoho and von Sachs [2] have proposed a different adaptive regularization of the best basis coefficients with a wavelet thresholding. They proved that for particular classes of locally stationary processes, the resulting algorithm yields a consistent estimator of the covariance.

A. Proof of Theorem 2.2. We estimate the norm of the error $U = T - B_K$ when approximating T with a band operator B_K. Let us denote $l_a = \max(l_p, l_q)$ and $l_b = \min(l_p, l_q)$. In a local cosine basis, the matrix coefficients of U are zero inside the band of B_K

$$(59)\; u_{p,k,q,j} = < U\phi_{p,k}, \phi_{q,j} > = \begin{cases} < T\phi_{p,k}, \phi_{q,j} > & \text{if } |p-q| > K \\ & \text{ or } |(k+\frac{1}{2})l_a l_p^{-1} - (j+\frac{1}{2})l_a l_q^{-1}| > K \\ 0 & \text{otherwise} \end{cases}$$

Since T is the covariance operator of a locally stationary process, the off-diagonal coefficients have a fast decay in a local cosine basis and for any $n \geq 2$ there exists Q_n such that

$$| < T\phi_{p,k}, \phi_{q,j} > | \leq \frac{Q_n}{(1 + |p-q|^n)(1 + |l_a(\xi_{p,k} - \xi_{q,j})|^n)}.$$

Replacing $\xi_{p,k}$ and $\xi_{q,j}$ by their expression (20) proves that for any $n \geq 2$ there exist constants D_n such that

$$(60) \qquad | < T\phi_{p,k}, \phi_{q,j} > | \leq \frac{D_n}{(1 + |p-q|^n)(1 + |(k+\frac{1}{2})l_a l_p^{-1} - (j+\frac{1}{2})l_a l_q^{-1}|^n)}$$

for all $(p, q, k, j) \in \mathbf{Z}^2 \times \mathbb{N}^2$. We use the following Schur lemma to derive an upper bound of $\|U\|_s$ from the amplitude of its coefficients.

LEMMA A.1 (SCHUR). *Let O be an operator whose matrix elements in an orthonormal basis $\{\phi_n\}_{n\in\mathbb{N}}$ are $o_{n,m} = < O\phi_n, \phi_m >$. If there are two sequences of positive numbers $\{w_m\}$ and $\{\hat{w}_m\}$ and a constant B such that*

$$\sum_{m=0}^{+\infty} |o_{n,m} w_m| \leq B\hat{w}_n$$

and

$$\sum_{n=0}^{+\infty} |o_{n,m} \hat{w}_n| \leq B w_m$$

then

$$\|O\|_s \leq B.$$

To apply Schur lemma to $U = T - B_K$, for any $n \geq 2$ we define the two weight sequences

$$\hat{w}_{q,j} = w_{q,j} = \frac{1}{1 + \max(K, |q|)^n} \, \frac{1}{1 + \max(K, j+\frac{1}{2})^n}.$$

If we can prove that for any $n \geq 2$, there exists a constant C_n such that

$$(61) \qquad \sum_{p,k} |u_{p,k,q,j} w_{p,k}| \leq \frac{C_n}{1 + K^{n-1}} \hat{w}_{q,j}$$

and

$$\sum_{q,j} |u_{p,k,q,j} \hat{w}_{q,j}| \leq \frac{C_n}{1+K^{n-1}} w_{p,k}. \tag{62}$$

then Schur lemma proves that

$$\|U\|_s = \|T - B_K\|_s \leq \frac{C_n}{1+K^{n-1}}.$$

Since this is valid for all $n \geq 2$, we derive the theorem result (24). By setting $U = T$ we prove (23) with essentially the same derivations.

The proof of (61) and (62) is identical since U is a symmetric operator. We concentrate on the proof of (61) which uses upper bounds given by the following lemma.

LEMMA A.2. *For any $n \geq 2$, there exist constants H_n and G_n such that for any $K \geq 0$ and $q \in \mathbf{Z}$*

$$\sum_{p=-\infty}^{+\infty} \frac{1}{1+|p-q|^n} \frac{1}{1+\max(|p|,K)^n} \leq \frac{H_n}{1+\max(|q|,K)^n} \tag{63}$$

and

$$\sum_{\substack{p=-\infty \\ |p-q|>K}}^{+\infty} \frac{1}{1+|p-q|^n} \frac{1}{1+\max(|p|,K)^n} \leq \frac{G_n}{1+K^{n-1}} \frac{1}{1+\max(|q|,K)^n}. \tag{64}$$

The proof of this lemma is left to the reader. One must distinguish the case $K \leq |q|$ and $K > |q|$. The sums over p must also be divided in two pieces where $\frac{1}{1+|p-q|^n}$ and $\frac{1}{1+\max(|p|,K)^n}$ are respectively smaller.

To prove (61), we evaluate the sum $\sum_{p,k} |u_{p,k,q,j} w_{p,k}|$ by replacing the $u_{p,k,q,j}$ by its expression (59). The coefficients $u_{p,k,q,j}$ are non-zero if $|p-q| > K$ or $|(k+\frac{1}{2}) l_a l_p^{-1} - (j+\frac{1}{2}) l_a l_q^{-1}| > K$. The sum over p and k is divided in two sums I and II corresponding to $|p-q| \leq K$ and $|p-q| > K$

$$\sum_{p,k} |u_{p,k,q,j} w_{p,k}| = I + II. \tag{65}$$

For non-zero values $u_{p,k,q,j} = | < T\phi_{p,k}, \phi_{q,j} > |$, we use an upper bound that is slightly different from (60). For any $n \geq 2$, there exists $E_n > 0$ such that $\forall (p,q,k,j) \in \mathbf{Z}^2 \times \mathbb{N}^2$

$$| < T\phi_{p,k}, \phi_{q,j} > | \leq \frac{E_n}{(1+|p-q|^{n+2n\mu})(1+|(k+\frac{1}{2}) l_a l_p^{-1} - (j+\frac{1}{2}) l_a l_q^{-1}|^n)}, \tag{66}$$

where $\mu < 1$ is the constant that appears in definition 1. We thus derive that

$$I \leq \sum_{\substack{p=-\infty \\ |p-q|\leq K}}^{+\infty} \Bigg(\frac{E_n}{1+|p-q|^{n+2n\mu}} \frac{1}{1+\max(|p|,K)^n} \sum_{\substack{k=0 \\ |(k+\frac{1}{2}) l_a l_p^{-1} - (j+\frac{1}{2}) l_a l_q^{-1}|>K}}^{+\infty} \frac{1}{1+|(k+\frac{1}{2}) l_a l_p^{-1} - (j+\frac{1}{2}) l_a l_q^{-1}|^n} \frac{1}{1+\max(k+\frac{1}{2},K)^n} \Bigg)$$

and

$$II \leq \sum_{\substack{p=-\infty \\ |p-q|>K}}^{+\infty} \Bigg(\frac{E_n}{1+|p-q|^{n+2n\mu}} \frac{1}{1+\max(|p|,K)^n} \sum_{k=0}^{+\infty} \frac{1}{1+|(k+\frac{1}{2})l_a l_p^{-1} - (j+\frac{1}{2})l_a l_q^{-1}|^n} \frac{1}{1+\max(k+\frac{1}{2},K)^n} \Bigg).$$

To compute an upper bound for I, observe that

$$\sum_{\substack{k=0 \\ |(k+\frac{1}{2})l_a l_p^{-1} - (j+\frac{1}{2})l_a l_q^{-1}|>K}}^{+\infty} \frac{1}{1+|(k+\frac{1}{2})l_a l_p^{-1} - (j+\frac{1}{2})l_a l_q^{-1}|^n} \frac{1}{1+\max(k+\frac{1}{2},K)^n} \leq$$

$$\sum_{\substack{k=0 \\ |(k+\frac{1}{2}) - (j+\frac{1}{2})l_p l_q^{-1}|>K l_p {l_a}^{-1}}}^{+\infty} \frac{1}{1+|(k+\frac{1}{2}) - (j+\frac{1}{2})l_p l_q^{-1}|^n} \frac{1}{1+\max(k+\frac{1}{2},K l_p {l_a}^{-1})^n}$$

Applying (64) gives

$$\sum_{\substack{k=0 \\ |(k+\frac{1}{2})l_a l_p^{-1} - (j+\frac{1}{2})l_a l_q^{-1}|>K}}^{+\infty} \frac{1}{1+|(k+\frac{1}{2})l_a l_p^{-1} - (j+\frac{1}{2})l_a l_q^{-1}|^n} \frac{1}{1+\max(k+\frac{1}{2},K)^n} \leq$$

$$\frac{G_n}{1+(K l_p {l_a}^{-1})^{n-1}} \frac{1}{1+\max((j+\frac{1}{2})l_p l_q^{-1}, K l_p {l_a}^{-1})^n}.$$

We thus derive that

$$I \leq \sum_{\substack{p=-\infty \\ |p-q|\leq K}}^{+\infty} \frac{E_n}{1+|p-q|^{n+2n\mu}} \frac{1}{1+\max(|p|,K)^n} \frac{G_n}{1+(K l_p {l_a}^{-1})^{n-1}} \frac{1}{1+\max((j+\frac{1}{2})l_p l_q^{-1}, K l_p {l_a}^{-1})^n}.$$

In definition 1, condition (21) guarantees the existence of $A > 0$ such that $|p-q|^{2n\mu} \geq {l_a}^{2n} {l_b}^{-2n} A^{-2n}$. Since $l_p {l_a}^{-1} \leq 1$ and $l_b {l_a}^{-1} l_p l_q^{-1} \leq 1$, we derive the existence of R_n such that

$$I \leq \sum_{\substack{p=-\infty \\ |p-q|\leq K}}^{+\infty} \frac{R_n}{1+|p-q|^n} \frac{1}{1+\max(|p|,K)^n} \frac{1}{1+K^{n-1}} \frac{1}{1+\max(j+\frac{1}{2},K)^n}.$$

We now use (63) to evaluate the sum over p and prove that there exists D_n^1 such that

$$I \leq \frac{D_n^1}{1+\max(K,|q|)^n} \frac{1}{1+K^{n-1}} \frac{1}{1+\max(K,j+\frac{1}{2})^n} = \frac{D_n^1}{1+K^{n-1}} \hat{w}_{q,j}.$$

With a similar approach, the reader can also verify that there exists D_n^2 such that

$$II \leq \frac{D_n^2}{1+K^{n-1}} \frac{1}{1+\max(K,|q|)^n} \frac{1}{1+\max(K,j+\frac{1}{2})^n} = \frac{D_n^2}{1+K^{n-1}} \hat{w}_{q,j}.$$

Inserting these two upper bounds in (65) completes the proof of (61).

B. Proof of Theorem 2.3. Theorem 2.3 is proved by constructing a local cosine basis in which the covariance operator T has matrix coefficients that satisfy the off-diagonal decay condition (22) of definition 1. The first part of the proof specifies this local cosine basis and proves that the windows lengths satisfy the slow variation condition (21) of definition 1. The second part proves (22).

Each window of a local cosine basis covers an interval $[a_p, a_{p+1}]$. The size l_p of any such interval is set to $l(a_p)$ or $l(a_{p+1})$, which is the scale of variation of the symbol $\Lambda_1(t, \omega)$ of T in this interval. We choose $a_0 = 0$ and if $p > 0$

$$a_{p+1} = a_p + l(a_p),$$

whereas if $p < 0$

$$a_p = a_{p+1} - l(a_{p+1}).$$

The rising and decaying intervals are stretched to their maximum

$$\eta_p = \frac{\min(l_p, l_{p-1})}{2}. \tag{67}$$

The rising and decaying profiles are specified by dilating a C^∞ function $\beta(t)$ such that

$$\beta(t) = \begin{cases} 0 & \text{if } t < -1 \\ 1 & \text{if } t > 1, \end{cases}$$

with

$$\beta^2(t) + \beta^2(-t) = 1.$$

The window $g_p(t)$ is defined by

$$g_p(t) = \begin{cases} \beta(\frac{t-a_p}{\eta_p}) & \text{if } t < a_p + \eta_p \\ 1 & \text{if } a_p + \eta_p \le t < a_{p+1} - \eta_{p+1} \\ \beta(\frac{a_{p+1}-t}{\eta_{p+1}}) & \text{if } t > a_{p+1} - \eta_{p+1} \end{cases} \tag{68}$$

The following lemma proves that the length l_p satisfies the slow variation condition (21) in definition 1.

LEMMA B.1. *There exists $A > 0$ such that for any $p \neq q$*

$$\frac{\max(l_p, l_q)}{\min(l_p, l_q)} \le A|p - q|^\mu, \tag{69}$$

where μ is related to the constant $\alpha < \frac{1}{2}$ in hypothesis (29) of the theorem by

$$\mu = \frac{\alpha}{1 - \alpha} < 1. \tag{70}$$

Proof of lemma B.1 To prove (69), we verify that there exists $C > 0$ such that for any $k \in \mathbb{N}$

$$\frac{\max(l_p, l_{p+k})}{\min(l_p, l_{p+k})} \leq C(k+1)^{\mu} \tag{71}$$

which implies (69) for $A = C2^{\mu}$. We suppose without loss of generality that $l_{p+k} \geq l_p$. Property (71) is proved by induction on k.

For $k = 0$ (71) is clearly valid for $C \geq 1$. Suppose that (71) is true for all $n < k$, with $k > 0$. We want to prove that

$$l_{k+p} \leq C l_p (k+1)^{\mu}. \tag{72}$$

We only consider the case where $p \geq 0$, the other one being identical. The window length is then specified by $l_{k+p} = l(a_{k+p})$ and hence

$$l_{k+p} = l\left(a_p + \sum_{j=0}^{k-1} l_{p+j}\right).$$

Hypothesis (29) of the theorem implies that

$$l_{k+p} \leq l(a_p) + \left(\sum_{j=0}^{k-1} l_{p+j}\right)^{\alpha}$$

Applying the induction hypothesis for $j < k$ gives

$$\begin{aligned} l_{k+p} &\leq l_p + \left(C l_p \sum_{j=0}^{k-1} (j+1)^{\mu}\right)^{\alpha} \\ &\leq l_p + C^{\alpha} l_p^{\alpha} \frac{(k+1)^{(\mu+1)\alpha}}{(\mu+1)\alpha}. \end{aligned}$$

The hypothesis (30) also supposes that

$$\inf_{t \in \mathbb{R}} l(t) = \eta > 0,$$

so $l_p = l(a_p) \geq \eta$. We thus obtain

$$l_{k+p} \leq l_p \left(1 + \frac{\eta^{\alpha-1} C^{\alpha}}{(\mu+1)\alpha} (k+1)^{(\mu+1)^{\alpha}}\right).$$

The constant μ in (70) satisfies $(\mu+1)\alpha = \mu$. We choose the constant C big enough so that

$$1 + \frac{\eta^{\alpha-1} C^{\alpha}}{(\mu+1)^{\alpha}} \leq C,$$

which verifies the induction hypothesis (72). This finishes the proof of (71). $\square$

In this second part of the proof of theorem 2.3, we verify that the matrix coefficients of the operator T satisfy the off-diagonal decay imposed by definition 1 for locally stationary processes

$$|< T\phi_{p,k}, \phi_{q,j} >| \leq \frac{Q_n}{(1+|p-q|^n)(1+|\max(l_p,l_q)(\xi_{p,k}-\xi_{q,j})|^n)}. \tag{73}$$

Instead of working with cosine modulated windows, we introduce

$$\psi_{p,k}(t) = \frac{1}{\sqrt{l_p}} g_p(t) e^{-i\xi_{p,k}t}. \tag{74}$$

The local cosine basis vectors can be written

$$\phi_{p,k}(t) = \frac{e^{i\theta_{p,k}}}{\sqrt{2}}\psi_{p,k}(t) + \frac{e^{-i\theta_{p,k}}}{\sqrt{2}}\psi_{p,-k}(t), \tag{75}$$

with $\theta_{p,k} = \xi_{p,k}a_p$. If we can prove that for any $n \geq 2$ there exists Q_n^1 such that for all $(p,k,q,j) \in \mathbf{Z}^4$

$$|< T\psi_{p,k}, \psi_{q,j} >| \leq \frac{Q_n^1}{(1+|p-q|^n)(1+|\max(l_p,l_q)(\xi_{p,k}-\xi_{q,j})|^n)}, \tag{76}$$

we then easily derive (73) by inserting (75) in (76). We now concentrate on proving (76).

Let us recall that

$$Tf(t) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \Lambda_1(t,\omega)\hat{f}(\omega)e^{i\omega t}d\omega.$$

Hence,

$$|< T\psi_{p,k}, \psi_{q,j} >| = \frac{1}{2\pi}\left|\int\int_{-\infty}^{+\infty} \hat{\psi}_{p,k}(\omega)\Lambda_1(t,\omega)e^{i\omega t}\psi_{q,j}^*(t)dtd\omega\right|. \tag{77}$$

Let $h_p(t) = g_p(t+a_p)$ be the window whose support is translated back in the neighborhood of $t=0$. Inserting (74) in (77) gives

$$|< T\psi_{p,k}, \psi_{q,j} >| = \frac{1}{2\pi\sqrt{l_pl_q}}\left|\int\int_{-\infty}^{+\infty} \hat{h}_p(\omega+\xi_{p,k})e^{-i\omega a_p}\Lambda_1(t,\omega)e^{i\omega t}e^{i\xi_{q,j}t}h_q(t-a_q)dtd\omega\right|. \tag{78}$$

The change of variables $\omega' = \omega + \xi_{p,k}$ and $t' = t - a_q$ yields

$$|< T\psi_{p,k}, \psi_{q,j} >| = \frac{1}{2\pi\sqrt{l_pl_q}}\left|\int\int_{-\infty}^{+\infty} \hat{h}_p(\omega')h_q(t')\Lambda_1(t'+a_q,\omega'-\xi_{p,k})e^{i\omega't'}e^{i\omega'(a_q-a_p)}e^{-it'(\xi_{p,k}-\xi_{q,j})}dtd\omega\right|. \tag{79}$$

Let us define

$$\Gamma_{p,q,k,j}(t,\omega) = \hat{h}_p(\omega)h_q(t)e^{it\omega}\Lambda_1(t+a_q,\omega-\xi_{p,k}).$$

The upper-bound (76) is obtained with an integration by parts in (79) which separates $\Gamma_{p,q,k,j}(t,\omega)$ and the remaining complex exponentials.

$$\begin{aligned} |< T\psi_{p,k}, \psi_{q,j} >| &= \frac{1}{2\pi\sqrt{l_p l_q}} \left| \int\int_{-\infty}^{+\infty} \partial_t^n \partial_\omega^m \Gamma_{p,q,k,j}(t,\omega) \frac{e^{i\omega(a_q - a_p)}}{|a_p - a_q|^m} \frac{e^{it'(\xi_{p,k} - \xi_{q,j})}}{|\xi_{p,k} - \xi_{q,j}|^n} dt d\omega \right| \\ &\leq \frac{1}{2\pi\sqrt{l_p l_q}} \frac{1}{|a_p - a_q|^m |\xi_{p,k} - \xi_{q,j}|^n} \int\int_{-\infty}^{+\infty} |\partial_t^n \partial_\omega^m \Gamma_{p,q,k,j}(t,\omega)| dt d\omega. \qquad (80) \end{aligned}$$

Let us denote

$$l_a = \max(l_p, l_q) \quad \text{and} \quad l_b = \min(l_p, l_q).$$

We prove later in lemma B.2 that there exists $A_{n,m}$ such that for all p, q, k, j

$$\int\int_{-\infty}^{+\infty} |\partial_t^n \partial_\omega^m \Gamma_{p,q,k,j}(t,\omega)|\, dt d\omega \leq A_{n,m} {l_a}^m {l_b}^{-(n-1)}. \tag{81}$$

Since $l_b \leq l_p, l_q \leq l_a$, inserting (81) in (80) shows that

$$|< T\psi_{p,k}, \psi_{q,j} >| \leq \frac{A_{n,m}}{2\pi} \sqrt{\frac{l_a}{l_b}} \frac{1}{\left|\frac{a_p - a_q}{l_a}\right|^m} \frac{1}{|l_b(\xi_{p,k} - \xi_{q,j})|^n} \tag{82}$$

and hence

$$|< T\psi_{p,k}, \psi_{q,j} >| \leq \frac{A_{n,m}}{2\pi} \frac{{l_a}^{n+\frac{1}{2}}}{{l_b}^{n+\frac{1}{2}}} \frac{1}{\left|\frac{a_p - a_q}{l_a}\right|^m} \frac{1}{|l_a(\xi_{p,k} - \xi_{q,j})|^n}. \tag{83}$$

To finish the upper bound computation, we show that there exists $C > 0$ such that

$$\frac{|a_p - a_q|}{l_a} \geq C|q - p|^{1-\mu}. \tag{84}$$

If $p < q$ then

$$\frac{|a_q - a_p|}{l_a} = \sum_{k=0}^{q-p-1} \frac{l_{p+k}}{l_a}.$$

Whether $l_a = l_p$ or $l_a = l_q$, we derive from (69) that

$$\frac{|a_q - a_p|}{l_a} \geq 1 + \sum_{k=1}^{q-p-1} A^{-1} k^{-\mu} \geq 1 + \frac{|q-p|^{1-\mu}}{A(1-\mu)},$$

which proves (84).

Inserting (84) in (83) gives

$$|< T\psi_{p,k}, \psi_{q,j} >| \leq \frac{A_{n,m}}{2\pi C^m} \frac{{l_a}^{n+\frac{1}{2}}}{{l_b}^{n+\frac{1}{2}}} \frac{1}{|p-q|^{m(1-\mu)}} \frac{1}{|l_a(\xi_{p,k} - \xi_{q,j})|^n}.$$

Property (69) shows that

$$\frac{l_a}{l_b} \leq A|p-q|^{\mu},$$

and hence

$$| < T\psi_{p,k}, \psi_{q,j} > | \leq \frac{A_{n,m} A^{n+\frac{1}{2}}}{2\pi C^m} \frac{1}{|p-q|^{m(1-\mu)-(n+\frac{1}{2})\mu}} \frac{1}{|l_a(\xi_{p,k}-\xi_{q,j})|^n}.$$

If m is large enough so that

$$m(1-\mu) - (n+\frac{1}{2})\mu \geq n$$

then

$$| < T\psi_{p,k}, \psi_{q,j} > | \leq \frac{A_{n,m} A^{n+\frac{1}{2}}}{2\pi C^m} \frac{1}{|p-q|^n} \frac{1}{|l_a(\xi_{p,k}-\xi_{q,j})|^n}.$$

By integrating directly (79), one can also prove that there exists $B > 0$ such that

$$| < T\psi_{p,k}, \psi_{p,k} > | \leq B.$$

We thus derive that for any $n \geq 0$ there exists Q_n^1 such that

$$| < T\psi_{p,k}, \psi_{q,j} > | \leq \frac{Q_n^1}{1+|p-q|^n} \frac{1}{1+|l_a(\xi_{p,k}-\xi_{q,j})|^n}.$$

Next lemma finishes the theorem proof by verifying (81).

Lemma B.2. *There exists $A_{n,m}$ such that for all p, q, k, j*

$$\int\int_{-\infty}^{+\infty} |\partial_t^n \partial_\omega^m \Gamma_{p,q,k,j}(t,\omega)| \, dt d\omega \leq A_{n,m} {l_a}^m {l_b}^{-(n-1)}. \tag{85}$$

Proof of lemma B.2 By definition

$$\Gamma_{p,q,k,j}(t,\omega) = \hat{h}_p(\omega) h_q(t) e^{it\omega} \Lambda_1(t+a_q, \omega - \xi_{p,k}).$$

We expand $\partial_t^n \partial_\omega^m \Gamma_{p,q,k,j}(t,\omega)$ into a sum of partial derivatives of $\hat{h}_p(\omega) h_q(t) e^{it\omega}$ and of $\Lambda_1(t+a_q, \omega+\xi_{p,k})$, and we prove that for any integers $c \geq 0$ and $d \geq 0$ there exists $D_{c,d}$ such that

$$\int\int_{-\infty}^{+\infty} \left|\partial_t^c \partial_\omega^d \Lambda_1(t+a_q, \omega+\xi_{p,k})\right| \left|\partial_t^{m-c} \partial_\omega^{n-d} [\hat{h}_p(\omega) h_q(t) e^{it\omega}]\right| dt d\omega \leq D_{c,d} {l_a}^m {l_b}^{-(n-1)}. \tag{86}$$

Property (28) guarantees that for any integers $c \geq 0$ and $d \geq 0$

$$|\partial_t^c \partial_\omega^d \Lambda_1(t+a_q, \omega+\xi_{p,k})| \leq B_{c,d} \; l(t+a_q)^{d-c}.$$

Since $\eta_q = \frac{\min(l_q, l_{q-1})}{2}$, the support of $h_q(t)$ is included in $[-\frac{l_q}{2}, \frac{3l_q}{2}]$. Hypothesis (29) of the theorem proves that over this support

$$|l(t+a_q) - l(a_q)| \leq A|t|^\alpha \leq A\frac{3^\alpha}{2^\alpha} l_q^\alpha,$$

with $\alpha < \frac{1}{2}$. Since $l_q = l(a_q) \geq \inf_{t \in \mathbb{R}} l(t) = \eta$,

$$l(t+a_q) \leq l_q(1 + A l_q^{\alpha-1}\frac{3^\alpha}{2^\alpha}) \leq l_q(1 + A\eta^{\alpha-1}\frac{3^\alpha}{2^\alpha})$$

so there exists $C_{c,d}$ such that

$$|\partial_t^c \partial_\omega^d \Lambda_1(t + a_q, \omega + \xi_{p,k})| \leq C_{c,d}\ l_q^{d-c}. \tag{87}$$

This proves that

$$\int\int_{-\infty}^{+\infty} \left|\partial_t^c \partial_\omega^d \Lambda_1(t + a_q, \omega + \xi_{p,k})\right| \left|\partial_t^{n-c}\partial_\omega^{m-d}[\hat{h}_p(\omega) h_q(t) e^{it\omega}]\right| dt d\omega \quad \leq$$
$$C_{c,d}\ l_q^{d-c} \int\int_{-\infty}^{+\infty} \left|\partial_t^{n-c}\partial_\omega^{m-d}[\hat{h}_p(\omega) h_q(t) e^{it\omega}]\right| dt d\omega.$$

To derive (86), we verify that for any j and l there exists $D_{j,l}$ such that

$$\int\int_{-\infty}^{+\infty} \left|\partial_t^j \partial_\omega^l [\hat{h}_p(\omega) h_q(t) e^{it\omega}]\right| dt d\omega \leq D_{j,l}\ {l_a}^l {l_b}^{-(j-1)}.$$

By expanding the partial derivatives $\partial_t^j \partial_\omega^l [\hat{h}_p(\omega) h_q(t) e^{it\omega}]$, we derive this last property from the next lemma. The details of this verification are left to the reader.

Lemma B.3. *For all $k \geq 0$ and $m \geq 0$, there exist a constant $E_{m,k}$ such that*

$$\int_{-\infty}^{+\infty} |t|^k |\partial_t^m h_p(t)| dt \leq E_{m,k}\ l_p^{k-m+1} \tag{88}$$

and a constant $F_{m,k}$ such that

$$\int_{-\infty}^{+\infty} |\omega|^k |\partial_\omega^m \hat{h}_p(\omega)| d\omega \leq F_{m,k}\ l_p^{m-k}. \tag{89}$$

Proof of lemma B.3: Let us denote $h_p^s(t) = h_p(l_p t)$. Since the support of $h_p(t)$ is included in $[-\frac{l_p}{2}, \frac{3l_p}{2}]$, the support of $h_p^s(t)$ is included in $[-\frac{1}{2}, \frac{3}{2}]$. With the change of variable $t' = \frac{t}{l_p}$ we obtain

$$\int_{-\infty}^{+\infty} \left|\frac{t}{l_p}\right|^k |\partial_t^m h_p(t)| \frac{dt}{l_p} = \int_{-\frac{1}{2}}^{3/2} |t'|^k |\partial_t^m h_p(l_p t')| dt' = l_p^{-m} \int_{-\frac{1}{2}}^{3/2} |t|^k |\partial_t^m h_p^s(t)| dt. \tag{90}$$

Since $h_p^s(t) = g(l_p(t + a_p))$, we derive from the expression (68) that for any $m \geq 0$

$$|\partial_t^m h_p^s(t)| \leq l_p^m \min(\eta_p, \eta_{p+1})^{-m} \sup_{t \in [-1,1]} |\partial_t^m \beta(t)|. \tag{91}$$

We proved in (69) that

$$\frac{l_p}{\min(l_p, l_{p-1})} \leq A \quad \text{and} \quad \frac{l_p}{\min(l_p, l_{p+1})} \leq A \quad ,$$

so

$$\min(\eta_p, \eta_{p+1}) = \frac{\min(l_{p-1}, l_p, l_{p+1})}{2} \geq \frac{l_p}{2A}.$$

We thus derive from (91) that there exists a constant B_m independent from l_p such that

$$|\partial_t^m h_p^s(t)| \leq B_m. \tag{92}$$

Coming back to (90), we obtain

$$\int_{-\infty}^{+\infty} \left|\frac{t}{l_p}\right|^k |\partial_t^m h_p(t)| \frac{dt}{l_p} \leq l_p^{-m} B_m \int_{-\frac{1}{2}}^{3/2} |t|^k dt = l_p^{-m} E_{m,k}, \tag{93}$$

which proves (88).

To prove the second equation (89), observe that the Fourier transform of $(it)^k \partial_t^m h_p^s(t)$ is $(-i\omega)^m \partial_\omega^k \hat{h}_p^s(\omega)$. Since the modulus of the Fourier transform is bounded by the $\mathbf{L^1}(\mathbb{R})$ norm of the function, we see from (92) that

$$|\omega|^m |\partial_\omega^k \hat{h}_p^s(\omega)| \leq \int_{-\infty}^{+\infty} |t|^k |\partial_t^m h_p^s(t)| dt = \int_{-\frac{1}{2}}^{\frac{3}{2}} |t|^k |\partial_t^m h_p^s(t)| dt \leq E_{m,k}.$$

The same property applied to $m' = m + 2$ proves that

$$|\omega|^m |\partial_\omega^k \hat{h}_p^s(\omega)| \leq \min(\frac{E_{m+2,k}}{\omega^2}, E_{m,k}).$$

We derive the existence of $F_{m,k}$ such that

$$\int_{-\infty}^{+\infty} |\omega|^m |\partial_\omega^k \hat{h}_p^s(\omega)| d\omega \leq F_{m,k}. \tag{94}$$

Since $h_p^s(t) = h_p(l_p t)$

$$|\hat{h}_p^s(\omega)| = \frac{1}{l_p} |\hat{h}_p(\frac{\omega}{l_p})|.$$

We finally prove (89) with the change of variable $\omega' = \frac{\omega}{l_p}$ in (94). □

C. Proof of Theorem 2.4. To prove that the process $X(t)$ is locally stationary, we must construct a local cosine basis in which the decomposition coefficients of $T = LL^t$ satisfy the off-diagonal decay condition (22) of definition 1.

The proof of theorem 2.3 does not use explicitly the fact that the covariance operator is symmetric. Since the symbol $A(t, \omega)$ of L satisfies the same hypothesis as the symbol $\Lambda_1(t, \omega)$ of T, appendix B gives a procedure to construct a local cosine basis $\{\phi_{p,k}\}$ whose window length satisfy condition (21) of definition 1, and such that for any $n \geq 2$ there exist Q_n with

$$(95) \qquad |b_{p,k,q,j}| = |< L\phi_{p,k}, \phi_{q,j} >| \leq \frac{Q_n}{1+|p-q|^n} \frac{1}{1+|\max(l_p,l_q)(\xi_{p,k}-\xi_{q,j})|^n}.$$

The matrix coefficients of $T = LL^T$ are

$$(96) \qquad < T\phi_{p,k}, \phi_{q,j} >= a_{p,k,q,j} = \sum_{r=-\infty}^{+\infty} \sum_{v=0}^{+\infty} b_{p,k,r,v} b_{q,j,r,v}.$$

Let us prove that these coefficients satisfy a decay property similar to (95).

Since $\xi_{p,k} = \pi(k+\frac{1}{2})l_p^{-1}$, inserting (95) in (96) gives

$$(97) \qquad |a_{p,k,q,j}| \leq |Q_n|^2 \sum_{r=-\infty}^{+\infty} \frac{1}{1+|p-r|^n} \frac{1}{1+|q-r|^n} \times I$$

with

$$\begin{aligned} I &= \sum_{v=0}^{+\infty} \frac{1}{1+\left|\max(l_p,l_r)((k+\frac{1}{2})l_p^{-1}-(v+\frac{1}{2})l_r^{-1})\right|^n} \frac{1}{1+\left|\max(l_q,l_r)((j+\frac{1}{2})l_q^{-1}-(v+\frac{1}{2})l_r^{-1})\right|^n} \\ &\leq \sum_{v=0}^{+\infty} \frac{1}{1+|(k+\frac{1}{2})l_r l_p^{-1}-(v+\frac{1}{2})|^n} \frac{1}{1+|(j+\frac{1}{2})l_r l_q^{-1}-(v+\frac{1}{2})|^n} \end{aligned}$$

With the change of variable $v' = v + \frac{1}{2} - (j+\frac{1}{2})l_r l_q^{-1}$ by setting $K = 0$ in (63) we derive that

$$(98) \qquad I \leq \frac{H_n}{1+|(k+\frac{1}{2})l_r l_p^{-1}-(j+\frac{1}{2})l_r l_q^{-1}|^n}.$$

We also proved in lemma B.1 that the properties of $l(t)$ imply the existence of $0 < \mu < 1$ such that

$$\frac{\max(l_p,l_q)}{\min(l_p,l_q)} \leq A|p-q|^\mu.$$

Hence

$$\frac{1}{|p-r|^{\mu n}} \frac{1}{|q-r|^{\mu n}} \leq \frac{A^{2n}}{\left(\frac{\max(l_p,l_r)}{\min(l_p,l_r)} \frac{\max(l_q,l_r)}{\min(l_q,l_r)}\right)^n},$$

We can thus derive the existence of D_n such that

$$|a_{p,k,q,j}| \leq D_n \sum_{r=-\infty}^{+\infty} \frac{1}{1+|p-r|^{(1-\mu)n}} \frac{1}{1+|q-r|^{(1-\mu)n}} \frac{1}{1+|l_{q,p,r}((k+\frac{1}{2})l_p^{-1}-(j+\frac{1}{2})l_q^{-1})|^n}$$

with

$$l_{q,p,r} = \frac{\max(l_p,l_r)}{\min(l_p,l_r)} \frac{\max(l_q,l_r)}{\min(l_q,l_r)} l_r \geq \max(l_p,l_q).$$

Since

$$\sum_{r=-\infty}^{+\infty} \frac{1}{1+|p-r|^m} \frac{1}{1+|q-r|^m} \leq \frac{H_m}{1+|q-p|^m},$$

for $m = (1-\mu)n$ we derive the existence of C_n such that

$$|a_{p,k,q,j}| \leq C_n \frac{1}{1+|q-p|^{(1-\mu)n}} \frac{1}{1+|\max(l_p, l_q)((k+\frac{1}{2})l_p^{-1} - (j+\frac{1}{2})l_q^{-1})|^n}.$$

Since this is valid for any $n \geq 2$, it implies that for any $n \geq 2$ there exists B_n such that

$$|a_{p,k,q,j}| \leq \frac{B_n}{1+|p-q|^n} \frac{1}{1+|\max(l_p, l_q)(\xi_{p,k} - \xi_{q,j})|^n}.$$

This proves that the operator T satisfies all the conditions of the local stationarity definition 1.

REFERENCES

[1] A. Benassi, S. Jaffard, D. Roux, "Elliptic Gaussian random processes," Preprint.

[2] D. Donoho, S. Mallat and R. von Sachs, "Estimating covariances of locally stationary processes: consistency of best basis methods," Proc. of Time Frequency and Time-Scale Symposium, Paris, July 1996.

[3] M. B. Priestley, "Evolutionary spectra and non-stationary processes," J. Roy. Stat. Soc. Ser. B, vol. 27, pp.204-237, 1965.

[4] M. B. Priestley, "Design relations for non-stationary processes," J. Roy. Stat. Soc. Ser. B, vol. 28, pp.228-240, 1965.

[5] K. S. Riedel, "Optimal data-based kernel estimation of evolutionary spectra," IEEE Trans. on Signal Proc., vol. 41, no. 7, pp. 2439-2447, July 1993.

[6] R.Coifman and V. Wickerhauser, "Entropy-based algorithms for best basis selection", IEEE Trans. on Information Theory, vol. 38, no. 2, March 1992.

[7] H. S. Malvar, "The LOT: transform coding without block effects", IEEE Trans. Acoust., Speech, Signal Process., vol 37, pp. 553-559, 1989.

[8] Y. Meyer, "Wavelets and Operators", Proceedings of Symposia in Applied Mathematics, vol 47, pp. 35-57, 1993.

[9] R. Coifman and Y. Meyer, "Remarques sur l'analyse de Fourier à fenêtre", C. R. Acad. Sci. Paris Sér. I, pp. 259-261, 1991.

[10] W. Martin and P. Flandrin, "Wigner-Ville spectral analysis of non-stationary processes", IEEE Trans. on Acoust. Speech and Sig. Proc., v. 33, no. 3, pp. 1461-1470, December 1985.

[11] Y. Meyer, "Wavelets — Algorithms and Applications", SIAM, 1993.

[12] M. Asch, W. Kohler, G. Papanicolaou, M. Postel and B. White, "Frequency content of randomly scattered signals", SIAM Review, vol. 33, No. 4, pp 519-625, December 1991.

[13] A.M. Sayeed and D. L. Jones, "Optimal kernels for nonstationary spectral estimation", IEEE Transactions on Signal Processing, v. 43, no. 2, pp. 478-491, February 1995.

[14] R. von Sachs and K. Schneider, "Wavelet smoothing of evolutionary spectra by nonlinear thresholding," Journal of Appl. and Comput. Harmonic Analysis, vol. 3, no. 3, pp. 268-282, July 1993.