INSTITUT DE STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION PAPER

0113

SIGNAL AND COVARIANCE ESTIMATION WITH MACROTILES

D. Donoho, S. Mallat, R. von Sachs and Y. Samuelides

http://www.stat.ucl.ac.be

Signal and Covariance Estimation with Macrotiles

David Donoho[1], Stéphane Mallat[2], Rainer von Sachs[3,4], Yann Samuelides[2]

March 25, 2001

Abstract

A macrotile estimation algorithm is introduced to remove additive noise from signals and to estimate the covariance of non-stationary processes. A macrotile algorithm uses a penalized method to optimize the partition of the space in orthogonal subspaces, and the estimation is computed with a projection operator. It is implemented by searching for a best basis among a dictionary of orthogonal bases, and by constructing an adaptive segmentation of this basis. Macrotile models are studied with local cosine bases to remove noise from sounds and to estimate the covariance matrices of locally stationary processes. The model selection and the estimation are implemented with a fast algorithm.

---

[1] Stanford University, Statistics Department, Stanford, CA 94305, U.S.A.

[2] Ecole Polytechnique, CMAP, 91128 Palaiseau Cedex, France

[3] Université catholique de Louvain, Institut de statistique, 1348 Louvain-la-Neuve, Belgium

[4] Financial support from the contract “Projet d’Actions de Recherche Concertes”no. 98/03-217 of the Belgian Government is gratefully acknowledged.

1 Introduction

In a non-parametric framework, to estimate a signal embedded in noise or to recover the covariance matrix of a process often requires to construct sparse representations. For noise removal, one can find an appropriate orthogonal basis where the signal is well approximated by few non-zero coefficients, and threshold these coefficients [9]. Best bases algorithms refine this approach by searching adaptively for an optimized basis, depending upon the signal [10].

The Karhunen-Loève basis which diagonalizes the covariance matrix provides a very sparse representation. For stationary processes, we know that the Karhunen-Loève basis is the Fourier basis, and we thus only need to estimate the diagonal coefficients which correspond to the power spectrum. For non-stationary processes, the Karhunen-Loève basis is not known in advance, but techniques have recently been studied in [13] and in [8] to estimate it among a dictionary of orthogonal bases. Analyzing the performance of this best basis search is more difficult than in the noise removal context, because the diagonalization of matrices has much more sophisticated properties than the decomposition of signals in orthogonal bases.

This paper introduces an adaptive projection subspace technique [17], which applies to noise removal and covariance estimation. Estimating an appropriate representation and the coefficients in this representation are two problems that are treated together, by searching for a macrotile model which divides the signal space in orthogonal subspaces. These orthogonal subspaces are constructed by grouping vectors in orthogonal bases that belong to a predefined dictionary of bases. The model is selected from the data by a penalization technique. Macrotile models constructed with local cosine vectors are used to remove noise from sounds and to estimate the covariance matrices of locally stationary processes. For signals or processes of size N, fast algorithms compute these estimators with $O(N(\log N)^2)$ operations.

Section 2 introduces the general principles of the macrotile algorithm to estimate signals embedded in noise and the covariance of processes. Macrotiles models constructed with local cosine bases provide a coarse tiling of the time-frequency plane. Applications to the removal of noises in sounds are studied in Section 3. The main result of the paper is the estimation of covariance matrices for locally stationary processes. Section 4 introduces an appropriate model for such processes and studies the performance of the macrotile estimation algorithm. The paper concentrates on algorithmic and numerical aspects of macrotile estimation as opposed to the mathematical proofs of consistency, which are long and technical, and can be found in [22] and in [21].

2 Penalized Macrotile Estimation

2.1 Projection Estimators

Let $x[n]$ be a signal of size N, contaminated by a zero-mean white Gaussian noise $\widehat{\theta}[n]$:

$$\widehat{x} = x + \widehat{\theta} \tag{1}$$

with $\mathsf{E}\{\widehat{\theta}[n]\widehat{\theta}[m]\} = \delta[n-m]\,\sigma^2$. The signal x is unknown but considered as deterministic so $x[n] = \mathsf{E}\{\widehat{x}[n]\}$. Given the empirical observations $\widehat{x}$, noise removal consists in finding an estimator $\tilde{x}$ of x which yields a small risk, that we shall compute with a Euclidean norm:

$$\mathsf{E}\{\|\tilde{x} - x\|^2\} = \sum_{n=1}^{N} \mathsf{E}\{|\tilde{x}[n] - x[n]|^2\} \ .$$

The estimation of the covariance operator Γ of a zero-mean Gaussian random vector $\widehat{y}[n]$ of size N can be casted as a similar estimation problem. This operator is characterized by a matrix of N^2 covariance coefficients

$$\gamma[n,m] = \mathsf{E}\{\widehat{y}[n]\,\widehat{y}^*[m]\} \quad \text{for } 1 \leq n, m \leq N \ . \tag{2}$$

Let $\widehat{\Gamma}$ be the operator whose matrix is composed of sample-mean estimators computed from R independent realizations $\{\widehat{y}_r\}_{1\leq r\leq R}$ of $\widehat{y}$:

$$\widehat{\gamma}[n,m] = \frac{1}{R}\sum_{r=1}^{R} \widehat{y}_r[n]\,\widehat{y}_r^*[m] \ . \tag{3}$$

It is clearly unbiased: $\mathsf{E}\{\widehat{\Gamma}\} = \Gamma$.

The risk is calculated with the Hilbert-Schmidt norm of $\Gamma - \widehat{\Gamma}$, which can be rewritten as a Euclidean norm in $\mathbb{C}^{N^2}$

$$\|\Gamma - \widehat{\Gamma}\|^2 = \sum_{n=1}^{N}\sum_{m=1}^{N} |\gamma[n,m] - \widehat{\gamma}[n,m]|^2 \ .$$

In signal processing, R is often much smaller than N, and we often have $R = 1$. The number of data points $N\,R$ is thus much smaller than the N^2 covariance coefficients that we intend to estimate, and as a result, $\|\Gamma - \widehat{\Gamma}\|^2$ is typically much larger than $\|\Gamma\|^2$. This means that estimating Γ as 0 would reduce the estimation risk. Improving this empirical estimator can be viewed as a denoising problem similar to (1)

$$\widehat{\Gamma} = \Gamma + \widehat{\theta} \ , \tag{4}$$

where the "noise" $\widehat{\theta} = \widehat{\Gamma} - \Gamma$ is the estimation error. A first difficulty is that $\widehat{\theta}$ is not Gaussian and has a complicated correlation structure. Moreover Γ is characterized by a matrix of N^2 coefficients that can be assimilated to an element of $\mathbb{C}^{N^2}$ as opposed to the much smaller signal space $\mathbb{C}^N$ in the denoising problem. Yet, we shall develop a similar strategy to perform the estimation in both cases.

The denoising and covariance estimation problems can be incorporated in the same framework, where we want to find an estimator $\tilde{z}$ of a vector $z \in \mathbb{C}^P$ given $\widehat{z} = z + \widehat{\theta}$. For denoising, $P = N$ and $z = x$ whereas for covariance estimation, $P = N^2$ and $z = \Gamma$. A simple approach consists in computing the projection of $\widehat{z}$ in a subspace $\mathcal{M}$ of dimension $\dim(\mathcal{M})$ much smaller than P:

$$\tilde{z} = P_{\mathcal{M}}(\widehat{z}) \ .$$

The risk can be decomposed according to the Pythagoras' theorem:

$$\|z - P_{\mathcal{M}}(\widehat{z})\|^2 = \|z - P_{\mathcal{M}}(z)\|^2 + \|P_{\mathcal{M}}(\widehat{\theta})\|^2 \ . \tag{5}$$

Since $\mathsf{E}\{\widehat{z}\} = z$, the first term $\|z - P_{\mathcal{M}}(z)\|$ is the estimation bias where as $\|P_{\mathcal{M}}(\widehat{\theta})\|^2$ carries the remaining noise. The space $\mathcal{M}$ can be considered as an estimation model.

Instead of choosing the subspace $\mathcal{M}$ a priori, one can try to optimize it depending upon $\widehat{z}$. We study a procedure where $\mathcal{M}$ is selected among a predefined family $\mathcal{F}$, which is optimized from prior information on z.

2.2 From Bases to Macrotile Models

We explain how to construct a family $\mathcal{F}$ of subspace models with a dictionary of bases and segmentations of these bases. Consider a dictionary $\mathcal{D} = \{\mathcal{B}_b\}_{b\in\mathcal{I}}$ where each $\mathcal{B}_b = \{g_m^b\}_{1\leq m\leq N}$ is an orthonormal basis of $\mathbb{C}^N$. Section 3.1 studies the examples of a local cosine dictionary. Given a particular basis $\mathcal{B}_b$, in order to define a subspace $\mathcal{M}$ of dimension $K = \dim(\mathcal{M})$, the N vectors $\{g_m^b\}_{1\leq m\leq N}$ are segmented into K groups, called macrotiles. A segmentation $S = \{I_k\}_{1\leq k\leq K}$ is a partition of the integers in $[1, N]$ into K disjoint integer sets I_k, each of which has a cardinal written $I_k^{\#}$. To the K macrotiles $\{g_m^b\}_{m\in I_k}$ we shall associate K orthogonal spaces $\mathbf{W}_k$ of dimension 1, in order to define a *macrotile model* space

$$\mathcal{M} = \oplus_{k=1}^{K} \mathbf{W}_k \ .$$

This model $\mathcal{M}$ is characterized by $\mathcal{B}_b$ and the segmentation S.

In the context of denoising, $z = x$ is a signal of size N and $\mathbf{W}_k$ is the one-dimensional space generated by the average vector $\sum_{m\in I_k} g_m^b$. The macrotile model $\mathcal{M} = \oplus_{k=1}^{K} \mathbf{W}_k$ contains all vectors of $\mathbb{C}^N$ whose coefficients in the basis $\mathcal{B}_b$ are constant over each macrotile $\{g_m^b\}_{m\in I_k}$. We easily verify that the orthogonal projection $P_{\mathcal{M}}(\widehat{x})$ has decomposition coefficients in $\mathcal{B}_b$ which are constant and equal to the average of the coefficients of $\widehat{x}$ on each macrotile:

$$P_{\mathcal{M}}(\widehat{x}) = \sum_{k=1}^{K} \frac{1}{I_k^{\#}} \left(\sum_{m\in I_k} \langle \widehat{x} \, , \, g_m^b \rangle \right) \sum_{m\in I_k} g_m^b \ . \tag{6}$$

For a given macrotile model $\mathcal{M}$, computing $P_{\mathcal{M}}(\widehat{x})$ is easy but one would like to use the ideal model $\mathcal{M} \in \mathcal{F}$, which minimizes $\|x - P_{\mathcal{M}}(\widehat{x})\|$ in $\mathcal{F}$. Since, x is unknown, this ideal model can not be exactly computed and must be estimated from the noisy data $\widehat{x}$. The next section describes a penalization approach which finds an empirical best model $\widehat{\mathcal{M}}$ which is nearly optimal, compared to the ideal model.

Computing a best empirical model amounts to finding a best basis in the dictionary $\mathcal{D}$ and a best segmentation of this basis, given the raw data. A macrotile algorithm can therefore be interpreted as a "best basis" denoising algorithm. However, as opposed to other "best basis" estimation algorithms [10], instead of estimating x by thresholding the coefficients of $\widehat{x}$ in a best basis, its coefficients are calculated with (6) using an adaptive averaging obtained by finding the best segmentation of the selected basis.

For covariance estimation, $z = \Gamma$ is an operator. Let us denote by $\mathbf{V}_k$ the space generated by the vectors $\{g_m^b\}_{m\in I_k}$ of a macrotile. In this case, we define $\mathbf{W}_k$ as the one-dimensional space of operators, which are diagonalized in $\mathcal{B}_b$ and whose eigenvalues are the same for all $\{g_m^b\}_{m\in I_k}$ and equal to zero for all other vectors of $\mathcal{B}_b$. The macrotile model $\mathcal{M} = \oplus_{k=1}^{K} \mathbf{W}_k$ is therefore the space of operators whose eigenspaces are the $\{\mathbf{V}_k\}_{1\leq k\leq K}$. The basis $\mathcal{B}_b$ diagonalizes the operators in $\mathcal{M}$ and their eigenvalues are constant over each macrotile $\{g_m^b\}_{m\in I_k}$. Let $P_{\mathbf{V}_k}$ be the orthogonal projection onto the vector space $\mathbf{V}_k$. Like in (6) we verify that the orthogonal projection $P_{\mathcal{M}}(\widehat{\Gamma})$ on the operator space $\mathcal{M}$ is given by

$$P_{\mathcal{M}}(\widehat{\Gamma}) = \sum_{k=1}^{K} \frac{1}{I_k^{\#}} \left(\sum_{m \in I_k} \langle \widehat{\Gamma} g_m^b, g_m^b \rangle \right) P_{\mathbf{V}_k} \quad , \tag{7}$$

and we derive from (3) that

$$\langle \widehat{\Gamma} g_m^b, g_m^b \rangle = \frac{1}{R} \sum_{r=1}^{R} |\langle \widehat{y}_r \, , \, g_m^b \rangle|^2 \ , \tag{8}$$

and that

$$\sum_{m \in I_k} \langle \widehat{\Gamma} g_m^b, g_m^b \rangle = \sum_{m \in I_k} \frac{1}{R} \sum_{r=1}^{R} |\langle \widehat{y}_r \, , \, g_m^b \rangle|^2 = \frac{1}{R} \sum_{r=1}^{R} \| P_{\mathbf{V}_k}(\widehat{y}_r) \|^2 \ . \tag{9}$$

If the vectors $\{g_m^b\}_{1 \leq m \leq N}$ of $\mathcal{B}_b$ are discrete Fourier vectors multiplied by square windows that localize them over subintervals of $[1, N]$ then (8) can be interpreted as a simple periodogram. Section 4.3 studies the application of local cosine bases computed with smooth windows, to estimate the spectrum of locally stationary processes.

In the general covariance context, computing a best empirical model means finding a best basis which nearly diagonalizes the covariance Γ. Within this basis, the segmentation is adapted to estimate the diagonal coefficients of Γ with an appropriate averaging of the empirical diagonal coefficients of $\widehat{\Gamma}$. This problem is clearly much more difficult than the denoising problem, and no statistical consistency result has been previously obtained on this problem. Yet, the next section shows that the same penalized approach enables us to estimate a best empirical model $\widehat{\mathcal{M}}$.

2.3 Penalization

A general framework for penalization estimation is presented in [2] and [3], for the estimation of probability densities. However, we cannot directly apply those results although the underlying concepts of our approach here are similar.

The risk of a macrotile estimation from $\widehat{z} = z + \widehat{\theta}$ is

$$\|z - P_{\mathcal{M}}(\widehat{z})\|^2 = \|z - P_{\mathcal{M}}(z)\|^2 + \|P_{\mathcal{M}}(\widehat{\theta})\|^2 \ . \tag{10}$$

Given a family of macrotile models $\mathcal{F}$, we want to find a model $\widehat{\mathcal{M}} \in \mathcal{F}$ which reduces as much as possible this risk. For this purpose, we first estimate an upper bound of $\|P_{\mathcal{M}}(\widehat{\theta})\|^2$. This upper bound is then used by a penalization procedure that finds $\widehat{\mathcal{M}}$. We denote by $\mathcal{V}^{\#}$ the total number of different macrotile spaces $\mathbf{W}_k$ used to construct the models $\mathcal{M} = \oplus_{k=1}^{K} \mathbf{W}_k$ in $\mathcal{F}$.

In the denoising context, $\widehat{\theta}$ is a Gaussian white noise; $P_{\mathcal{M}}$ is an orthogonal projection onto a space of dimension $\dim(\mathcal{M})$, and therefore,

$$\mathsf{E}\{\|P_{\mathcal{M}}(\widehat{\theta})\|^2\} = \sigma^2 \dim(\mathcal{M}) \ .$$

Moreover, using a standard inequality on Gaussian distributions, since all models $\mathcal{M}$ are constructed from a family of $\mathcal{V}^{\#}$ macrotile spaces, one can prove [21] that the probability that

$$\|P_{\mathcal{M}}(\widehat{\theta})\| \leq \sigma \sqrt{3 \log \mathcal{V}^{\#} \dim(\mathcal{M})} \tag{11}$$

tends to 1 when $\mathcal{V}^{\#}$ tends to $+\infty$.

For covariance estimation, $\widehat{\theta} = \widehat{\Gamma} - \Gamma$ is a matrix of empirical estimation error, whose entries have χ^2 distributions. The variances of these errors are inversely proportional to the number R of realizations used to compute the sample means, and proportional to the covariance coefficients of Γ, which are bounded by the largest eigenvalue λ of Γ. In the Appendix, we derive that

$$\mathsf{E}\{\|P_{\mathcal{M}}(\widehat{\theta})\|^2\} \leq 2\,\lambda^2 \,\frac{\dim(\mathcal{M})}{R} \ . \tag{12}$$

Like in the denoising case, the variance of the projected noise is bounded by a constant proportional to the dimension of the model, and the largest eigenvalue λ of Γ plays the same role as the variance of the noise in the denoising problem. Moreover, in [22] it is shown that, as in (11), if Γ is the covariance of a Gaussian process, then $\widehat{\theta} = \Gamma - \widehat{\Gamma}$ satisfies

$$\|P_{\mathcal{M}}(\widehat{\theta})\| \leq \ 3\,\lambda \,\log \mathcal{V}^{\#} \sqrt{\frac{\dim(\mathcal{M})}{R}} \tag{13}$$

with a probability that tends to 1 when $\mathcal{V}^{\#}$ tends to $+\infty$.

Inserting (11) or (13) into (5) gives an upper bound for the estimation risk: With probability tending to 1 when $\mathcal{V}^{\#}$ tends to $+\infty$.

$$\|z - P_{\mathcal{M}}(\widehat{z})\|^2 \leq \|z - P_{\mathcal{M}}(z)\|^2 + C \dim(\mathcal{M}) \ , \tag{14}$$

with $C = 3\,\sigma^2 \,\log \mathcal{V}^{\#}$ in the denoising case and $C = 9\,\lambda^2 (\log \mathcal{V}^{\#})^2/R$ for covariance estimation. The penalization approach tries to find a model which nearly minimizes this upper bound, and which therefore produces a risk close to

$$\mathrm{Risk}_{\min} = \inf_{\mathcal{M}\in\mathcal{F}} \left\{ \|z - P_{\mathcal{M}}(z)\|^2 + C\,\dim(\mathcal{M}) \right\} \ . \tag{15}$$

Observe that the second term on the right side increases when $\dim(\mathcal{M})$ increases. Assume that the macrotile family $\mathcal{F}$ satisfies a refinement property such that for all models $\mathcal{M} \neq \mathbb{C}^P \in \mathcal{F}$, there exists a richer model $\mathcal{M}' \in \mathcal{F}$ with

$$\mathcal{M} \subset \mathcal{M}' \ \text{ and } \ \dim(\mathcal{M}) < \dim(\mathcal{M}') \leq 2\,\dim(\mathcal{M}) \tag{16}$$

It is then always possible to decrease the first term of (14) by increasing the dimension $\dim(\mathcal{M})$ of the model. The minimum is thus nearly reached when both terms are of the same order.

For estimating a model $\mathcal{M}$ that can nearly reach the minimum $\mathrm{Risk}_{\min}$, observe that

$$\|z - P_{\mathcal{M}}(z)\|^2 = \|z\|^2 - \|P_{\mathcal{M}}(z)\|^2 \ .$$

Therefore, $\|z - P_{\mathcal{M}}(z)\|^2$ can be replaced by $-\|P_{\mathcal{M}}(z)\|^2$ when minimizing the upper bound (14). Since $\|P_{\mathcal{M}}(z)\|^2$ is unknown, the penalization approach replaces it by $\|P_{\mathcal{M}}(\widehat{z})\|^2$ and we compute

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}\in\mathcal{F}} \left\{ -\|P_{\mathcal{M}}(\widehat{z})\|^2 + C'\,\dim(\mathcal{M}) \right\} \ . \tag{17}$$

Of course, the main difficult is to prove that the risk $\|z - P_{\widehat{\mathcal{M}}}(\widehat{z})\|^2$ obtained with the penalized model $\widehat{\mathcal{M}}$ is of the same order as $\mathrm{Risk}_{\min}$, with probability close to 1, despite the fact that we have approximated $\|P_{\mathcal{M}}(z)\|^2$ by $\|P_{\mathcal{M}}(\widehat{z})\|^2$. This requires to choose a penalization constant C' at least larger than C by a fixed factor.

In the context of noise removal where $z = x$ and $\widehat{z} = \widehat{x}$ are signals of size N, we derive from (6) that

$$\|P_{\mathcal{M}}(\widehat{z})\|^2 = \sum_{k=1}^{K} \frac{1}{I_k^{\#}} \Big| \sum_{m\in I_k} \langle \widehat{z}\,,\, g_m^b \rangle \Big|^2 \ . \tag{18}$$

Since $C = 3\,\sigma^2\,\log \mathcal{V}^{\#}$, we write

$$C' = C_p\,\sigma^2\,\log \mathcal{V}^{\#} \ .$$

Using techniques developed by Barron, Birge and Massart [2] based on Talagrand inequalities, if $C_p \geq 32\,\sigma^2$, one can prove [21] that the optimal penalized model

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}\in\mathcal{F}} \Big\{ -\|P_{\mathcal{M}}(\widehat{z})\|^2 + C_p\,\sigma^2\,\log \mathcal{V}^{\#}\,\dim(\mathcal{M}) \Big\} \tag{19}$$

has a nearly optimal risk

$$\|x - P_{\widehat{\mathcal{M}}}(\widehat{x})\|^2 \leq 4\, C_p\, \mathrm{Risk}_{\min} \ , \tag{20}$$

with a probability that tends to 1 when $\mathcal{V}^{\#}$ tends to infinity. This result is also similar to the one obtained for removing Gaussian white noise by thresholding coefficients in a best basis selected among a dictionary [10].

For covariance estimation, we already mentioned that the problem is much more difficult and no mathematical result guarantees that in general, the penalized model yields a risk that is close to the minimum risk. However, Section 4.3 explains that when estimating locally stationary processes in a dictionary of local cosine bases, this near-optimality result is valid.

3 Denoising with Local Cosine Macrotiles

3.1 Local Cosine Macrotile Models

For noise removal, we know that a penalized macrotile search yields a risk that is of the same order as the minimum risk

$$\mathrm{Risk}_{\min} = \inf_{\mathcal{M} \in \mathcal{F}} \left\{ \|x - P_{\mathcal{M}}(x)\|^2 + C_p\, \sigma^2 \, \log_2 \mathcal{V}^{\#} \, \dim(\mathcal{M}) \right\} \ .$$

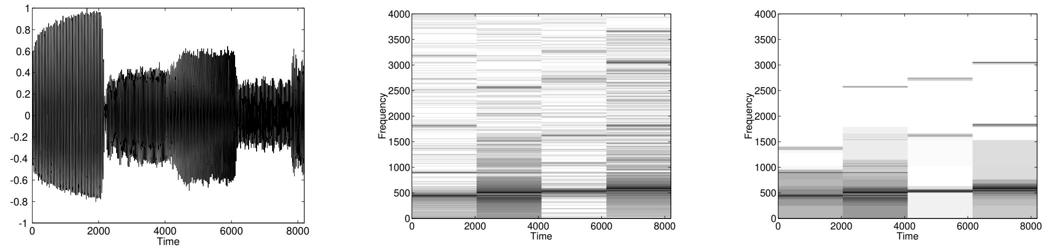
The main issue is now to construct families of macrotiles models which can guarantee that $\mathrm{Risk}_{\min}$ is small.

The time-frequency properties of many classes of signals, including sounds, motivates the use of local cosine macrotiles. If $\mathcal{F}$ satisfies the refinement property (16), then $\mathrm{Risk}_{\min}$ minimizes the sum of two terms which respectively decrease and increase when $\dim(\mathcal{M})$ increases. Its minimum is thus achieved, up to a factor 2, when

$$\|x - P_{\mathcal{M}}(x)\|^2 = C_p \sigma^2 \, \log \mathcal{V}^{\#} \, \dim(\mathcal{M}) \ .$$

The resulting risk is small if one can reconstruct a precise approximation of x from coefficients in a basis which take a very limited number of different values. For sound signals, this is indeed the case in an appropriate local cosine basis. The performance of audio compression standards such as MPEG-2-AAC, which is based on an adapted local cosine basis, shows that audio sounds can often be well approximated by few non-zero coefficients. This shows that one can construct local cosine macrotile models, of small dimension, that efficiently approximate such signals.

A local cosine basis divides an interval $[0, N]$ with windows of varying sizes, which are multiplied by cosine vectors of varying frequencies. Let



(a) (b) (c)

Figure 1: a) Recording f of four notes played on a clarinet. b) Tiling corresponding to the local cosine coefficients $|\langle f\,,\,g_{p,q}\rangle|$. The darker the grey level of the tile the larger the amplitude. c) Macrotiles obtained by averaging the tiles over groups of varying sizes for each interval.

$1/2 = a_1 < a_2 < \ldots < a_{L+1} = N + 1/2$ be half-integers that define a partition of $[0, N]$ in L intervals. Let $l_p = a_{p+1} - a_p$, and $\eta \leq \min_p l_p/2$. One can construct for each $p \in [1, L]$ a regular window $g_p[n]$ with support $[a_p - \eta, a_{p+1} + \eta]$ such that the local cosine family

$$\left\{ g_{p,q}[n] = g_p[n]\,\sqrt{\frac{2}{l_p}}\,\cos\left[\pi\,(q+\frac{1}{2})\,\frac{n-a_p}{l_p}\right]\right\}_{1\leq p\leq L\,,\,0\leq q\leq l_p} \qquad (21)$$

is an orthonormal basis of $\mathbb{C}^N$ [5, 15, 14]. Each local cosine vector $g_{p,q}$ has a support nearly localized in the interval $[a_p\,,\,a_{p+1}]$, and its Fourier transform has its energy mostly localized in the interval $[q\pi/l_p\,,\,(q+1)\pi/l_p]$. It can thus be represented by the Heisenberg rectangle $[a_p, a_{p+1}] \times [q\pi/l_p\,,\,(q+1)\pi/l_p]$ in a time-frequency plane. This Heisenberg box is called a *tile* because the union of the Heisenberg rectangles associated to a local cosine basis defines an exact tiling of the time-frequency plane. This is illustrated by Figure 1 which displays the local cosine coefficients of an audio recording $x[n]$. Darker tiles correspond to high amplitude coefficients $|\langle x\,,\,g_{p,q}\rangle|$ whereas white tiles coefficients $|\langle x\,,\,g_{p,q}\rangle|$ that are nearly zero.

A difficult issue is to find a time partition which defines a local cosine basis which is well adapted to an estimation problem. Coifman and Meyer [5, 14] have thus constructed a dyadic dictionary of local cosine bases, which includes orthonormal bases defined over partitions whose intervals have sizes which are powers of 2. The dictionary is organized as a tree. For $a_p = pN2^{-j} + 1/2$ and $a_{p+1} = (p+1)N2^{-j} + 1/2$, we denote by g_p^j the window whose support is $[a_p - \eta\,,\,a_{p+1} + \eta]$ and by $\mathcal{B}_p^j$ the corresponding family of

$l_p = N\,2^{-j}$ local cosine vectors

$$\mathcal{B}_p^j = \left\{ g_{p,q,j}[n] = g_p^j[n]\,\sqrt{\frac{2}{2^{-j}N}}\,\cos\left[\pi\Big(q+\frac{1}{2}\Big)\frac{n-p2^{-j}N-1/2}{2^{-j}N}\right]\right\}_{0\le q<N2^{-j}} .$$

The family of cosine vectors $\mathcal{B}_p^j$ is stored at the depth j and position p (from left to right) of a binary tree. The 2^j local cosine families $\mathcal{B}_p^j$ at the depth j divide the interval $[0, N]$ in 2^j overlapping intervals. The local cosine family $\mathcal{B}_0^0$ at the root of the local cosine tree has N local cosine vectors that cover the whole signal support. The maximum depth of this tree is $J = \log_2(N/(2\eta))$ because $\eta = \min_p l_p/2 = 2^{-J-1}N$. There are thus $J\,N \le N\,\log_2 N$ local cosine vectors stored in the tree.

A local cosine basis in this dictionary is constructed with an admissible subtree, whose nodes have either 0 or 2 children. Let $b = \{p_l, j_l\}_{1\le l\le L}$ be the position and depth of the L leaves of an admissible subtree. One can verify that the family

$$\mathcal{B}_b = \{\mathcal{B}_{p_l}^{j_l}\}_{1\le l\le L} \tag{22}$$

is a local cosine basis of $\mathbb{C}^N$ [14]. It defines a particular segmentation of the interval $[0, N]$ into windows of various sizes.

Remember that a family of macrotile models $\mathcal{F}$ is defined with a dictionary of bases $\mathcal{D}$ and a set of admissible segmentations. For a local cosine basis $\mathcal{B}_b = \{\mathcal{B}_{p_l}^{j_l}\}_{1\le l\le L}$, we shall consider segmentations that are performed independently within each family $\mathcal{B}_p^j$, and which satisfy the refinement property (16).

Sound signals can include frequency tones that are localized over a time intervals and have narrow frequency bandwidth. Let $\mathcal{B}_j^p = \{g_{p,q,j}\}_{0\le q<N\,2^{-j}}$ be a local cosine family whose window covers a time interval $[p\,2^{-j}\,N,\ (p+1)\,2^{-j}\,N]$ where the signal has several narrow frequency tones. To efficiently approximate such a signal with a macrotile model of small dimension, it is necessary to define a frequency segmentation that uses narrow macrotiles in the neighborhood of the frequency tones, to approximate them efficiently. Constructing a local cosine macrotile model is similar to finding a piecewise constant approximation of the local signal coefficients $\langle f\,,\,g_{p,q,j}\rangle$ when the frequency index q varies for p and j fixed. For each p, j, we thus consider any possible grouping, which corresponds to a piecewise constant sequence of the $2^{-j}N$ local cosine coefficients. We shall however impose that each macrotile has a size which is a power of two. For a given p, j, a segmentation in $K_{j,p}$ macrotile frequency intervals, is thus defined by a sequence of length $S = \{2^{s_k}\}_{1\le k\le K_{j,p}}$. Imposing that each group has a number of elements

which is a power of 2 has no asymptotic effect on the performance of the macrotile averaging, and allows one to optimize this segmentation with a fast CART algorithm [4]. A macrotile model $\mathcal{M}$ is specified by the indices of the local cosine basis $\mathcal{B}_b = \{\mathcal{B}^{j_l}_{p_l}\}_{1 \leq l \leq L}$ and the adaptive segmentation S_l for each $\mathcal{B}^{j_l}_{p_l}$

$$\mathcal{M} = \{j_l\,,\, p_l\,,\, S_l\}_{1 \leq l \leq L}\ . \tag{23}$$

Figure 1 displays the representation of $P_{\mathcal{M}}(x)$ for one such macrotile model. This macrotile model is just an example and is not optimal in any sense with respect to this signal. The gray level of each macrotile indicates the amplitude of the average inner product corresponding to a particular grouping of local cosine tiles. For a local cosine tree of depth $J \leq \log_2 N$, one can verify that the total number of different macrotile models $\mathcal{M}$ is larger than $2^{N/2}$, but all these models are constructed from only a number $\mathcal{V}^{\#}$ of macrotiles which satisfies

$$(2J-1)\,N \leq \mathcal{V}^{\#} \leq 2J\,N\ .$$

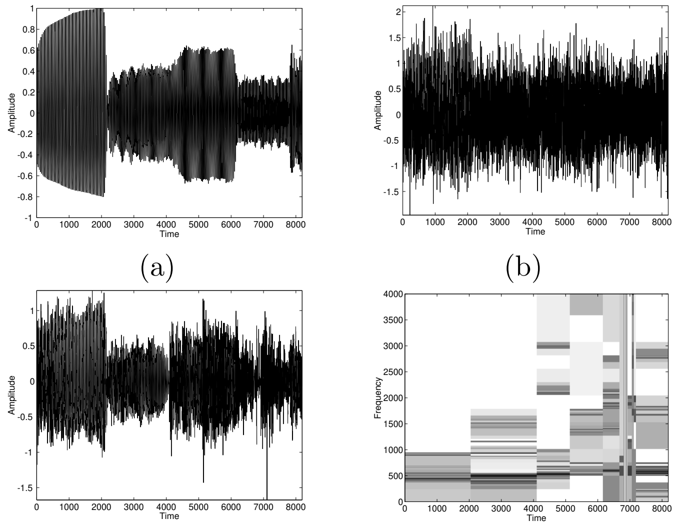
3.2 Denoising with Fast Model Search

We now study the application of local cosine macrotiles to noise removal with a fast implementation using the tree structure of local cosine dictionaries. Given a noisy signal $\widehat{x} = x + \widehat{\theta}$, a best local cosine model is selected according to (19). Since $\log_2 \mathcal{V}^{\#} \approx \log_2 N$, this amounts to computing

$$\widehat{\mathcal{M}} = \arg \min_{\mathcal{M} \in \mathcal{F}} \left\{ -\|P_{\mathcal{M}}(\widehat{x})\|^2 + C_p\, \sigma^2\, \log N\, \mathrm{dim}(\mathcal{M}) \right\}\ . \tag{24}$$

Figures 2 illustrates this macrotile denoising for the music signal with four notes, shown in Figure 1a. Figure 2b is the signal contaminated by a white noise for an SNR of 0 db. Figure 2c gives the macrotile estimation $P_{\widehat{\mathcal{M}}}(\widehat{x})$ calculated with the macrotile denoising algorithm. The resulting SNR is 2 db, which corresponds to a gain of 7 db. Figure 2d displays the macrotiles in the time-frequency plane. The harmonics appear as darker and more narrow macrotiles; the time-frequency structures of the signal are clearly preserved. Two other experiments were performed with the same original signal and other white noises corresponding to SNR of 6.5 db and -6 db; the gains were respectively of 6 db and 8 db. The audio quality of the signal is considerably improved by this macrotile denoising, yet, this example is only an illustration, and further refinements in the models are needed to optimize the noise removal for music signals.

When the family of models is constructed from dictionaries of orthonormal bases having a tree structure, like a local cosine dictionary, the best



(a) (b)

(c) (d)

Figure 2: a) Music signal f. b) Noisy signal $\widehat{x} = x + \widehat{\theta}$. c) Penalized macrotile estimation $P_{\widehat{\mathcal{M}}}(\widehat{x})$ calculated with (24). d) Representation of the time-frequency macrotiles of $P_{\widehat{\mathcal{M}}}(\widehat{x})$. The darkest macrotiles correspond to high energy coefficients.

penalized model is computed with a fast dynamic programming algorithm, which takes advantage of the additivity property of the penalization cost for partial models. A partial model $\mathcal{M}$ is characterized by an orthonormal basis $\mathcal{B}$ of a subspace $\mathbf{V}$ of $\mathbb{C}^N$, as opposed to a basis of the whole space, and a segmentation S of $\mathcal{B}$. The cost of this model is defined by

$$\mathrm{Cost}(\mathcal{M}) = -\|P_{\mathcal{M}}(\widehat{x})\|^2 + C_p\,\sigma^2\,\log N\,\dim(\mathcal{M})\ ,$$

where $\|P_{\mathcal{M}}(\widehat{x})\|^2$ is calculated with (18) by summing only the coefficients restricted to the basis $\mathcal{B}$ of $\mathbf{V}$. Let $\mathcal{M}_1 = (\mathcal{B}_1\,,\,S_1)$ and $\mathcal{M}_2 = (\mathcal{B}_2\,,\,S_2)$ be two partial models where $\mathcal{B}_1$ and $\mathcal{B}_2$ are two orthogonal families of vectors. One can define a new model $\mathcal{M} = (\mathcal{B}_1 \cup \mathcal{B}_2\,,\,S_1 \cup S_2)$ where $S_1 \cup S_2$ is the segmentation of $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ obtained by segmenting $\mathcal{B}_1$ with S_1 and $\mathcal{B}_2$ with S_2. Clearly

$$\mathrm{Cost}(\mathcal{M}) = \mathrm{Cost}(\mathcal{M}_1) + \mathrm{Cost}(\mathcal{M}_2)\ . \tag{25}$$

Using this additive property in a tree dictionary, one can now identify the model that minimizes the cost with the fast bottom up strategy of Coifman and Wickerhauser [6].

To any node of the local cosine tree, is associated a family of local cosine vectors $\mathcal{B}_p^j$ which generates a vector space $\mathbf{V}_p^j$. Let $\mathcal{M}_p^j$ be the optimal model of $\mathbf{V}_p^j$ which minimizes $\mathrm{Cost}(\mathcal{M})$ among all models $\mathcal{M}$ defined by $(\mathcal{B}, S)$ where $\mathcal{B}$ is a local cosine basis of $\mathbf{V}_p^j$. One possible candidate is the model $\mathcal{M}_j^{p*} = (\mathcal{B}_p^j, S_p^j)$ where the segmentation S_p^j is adjusted to minimize the cost associated with the basis $\mathcal{B}_p^j$. However other models can also be considered by building other bases of $\mathbf{V}_p^j$ from the local cosine vectors located in the descendant nodes. The space $\mathbf{V}_p^j$ can be decomposed into two orthogonal subspaces $\mathbf{V}_{2p}^{j+1}$ and $\mathbf{V}_{2p+1}^{j+1}$ generated by the two local cosine families $\mathcal{B}_{2p}^{j+1}$ and $\mathcal{B}_{2p+1}^{j+1}$ located at the children nodes in the binary tree. Using the additivity property (25) we easily verify that

$$\mathcal{M}_p^j = \begin{cases} \mathcal{M}_j^{p*} & \text{if } \mathrm{Cost}(\mathcal{M}_p^{j*}) \leq \mathrm{Cost}(\mathcal{M}_{2p}^{j+1}) + \mathrm{Cost}(\mathcal{M}_{2p+1}^{j+1}) \\ \mathcal{M}_{2p}^{j+1} \cup \mathcal{M}_{2p+1}^{j+1} & \text{if } \mathrm{Cost}(\mathcal{M}_p^{j*}) > \mathrm{Cost}(\mathcal{M}_{2p}^{j+1}) + \mathrm{Cost}(\mathcal{M}_{2p+1}^{j+1}) \end{cases} . \tag{26}$$

At the leaves of a local cosine tree of depth J, the only possible choice is $\mathcal{M}_p^J = \mathcal{M}_p^{J*}$ since $\mathcal{B}_p^J$ is the only basis available for $\mathbf{V}_p^J$. Using the aggregation relation (26), the optimal model is computed by going up the tree along each branch, until the root where $\mathbf{V}_0^0 = \mathbb{C}^N$. The resulting optimal model is $\mathcal{M}_0^0 = \widehat{\mathcal{M}}$ and is the model that minimizes the cost over $\mathbb{C}^N$. The overall computational complexity of the algorithm is driven by the number of operations to compute the models $\mathcal{M}_p^{j*} = (\mathcal{B}_p^j, S_p^j)$ at all nodes of the tree.

For this purpose, we first compute the $J\,N$ local cosine coefficients of the noisy signal $\widehat{\alpha}_p^j[q] = \langle \widehat{x}, g_{p,q,j} \rangle$, for all local cosine vectors $g_{j,q,p}$ in the tree. With a fast lapped cosine transform [15], it requires $O(J\,N \log_2 N)$ operations. Then for each j and p, we compute the segmentation S_p^j which yields a model $\mathcal{M}_p^{j*}$ of minimum cost for local cosine family $\mathcal{B}_p^j$. The cost is calculated from the $N2^{-j}$ coefficients $\{\widehat{\alpha}_p^j[q]\}_{1 \le q \le N2^{-j}}$ associated to $\mathcal{B}_p^j$.

The optimal segmentation S_p^j can be computed with $O(2^{-j}N)$ operations with a CART algorithm [4]. In the following we write $\mathrm{Cost}(\mathcal{S})$ the cost of a partial macrotile model corresponding to a segmentation $\mathcal{S}$ of $\mathcal{B}_p^j$. By definition S_p^j is the segmentation which minimizes this cost. To compute it, we use a segmentation tree, which sub-decompose any segmentation in partial segmentations whose cost are summed. At the level l of this segmentation tree, each node of index $0 \le k < 2^l$ corresponds to segmentations of the group of $N2^{-l-j}$ coefficients $\widehat{\alpha}_p^j[q]$ for $kN2^{-l-j} < q \le (k+1)N2^{-l-j}$. Let $\mathcal{S}_l^k$ be the optimal segmentation of this group, which yields a partial model of minimum cost, that we write $\mathrm{Cost}(\mathcal{S}_l^k)$. One possible candidate is the segmentation $\mathcal{S}_l^{k*}$ where all coefficients of the group are averaged into one macrotile whose average value is

$$\bar{\alpha}_l[k] = \frac{1}{N2^{-l-j}} \sum_{q=kN2^{-l-j}+1}^{(k+1)N2^{-l-j}} \widehat{\alpha}_p^j[q] \ .$$

The cost of the partial model associated with this segmentation is

$$\mathrm{Cost}(\mathcal{S}_l^{k*}) = -N\,2^{-l-j}\,|\bar{\alpha}_l[k]^2| + C_p \log N \ .$$

As in (26), using the additivity property of the cost function, one can verify that

$$\mathcal{S}_k^l = \left\{ \begin{array}{ll} \mathcal{S}_k^{l*} & \text{if } \mathrm{Cost}(\mathcal{S}_k^{l*}) \le \mathrm{Cost}(\mathcal{S}_{2k}^{l+1}) + \mathrm{Cost}(\mathcal{S}_{2k+1}^{l+1}) \\ \mathcal{S}_{2k}^{l+1} \cup \mathcal{S}_{2k+1}^{l+1} & \text{if } \mathrm{Cost}(\mathcal{S}_k^{l*}) > \mathrm{Cost}(\mathcal{S}_{2k}^{l+1}) + \mathrm{Cost}(\mathcal{S}_{2k+1}^{l+1}) \end{array} \right. . \tag{27}$$

At the maximum depth $L = \log_2 N - j$ of this tree, the only possible segmentation is $\mathcal{S}_k^L = \mathcal{S}_k^{L*}$ because there is only one coefficient in each group. Using (27), a bottom-up strategy requires $O(2^{-j}N)$ operations to compute $\mathcal{S}_0^0$, which is the optimal segmentation associated to all coefficients $\{\widehat{\alpha}_p^j[q]\}_{1 \le q \le N2^{-j}}$, and hence $S_p^j = \mathcal{S}_0^0$. This segmentation defines $\mathcal{M}_p^{j*}$, and by definition

$$\mathrm{Cost}(\mathcal{M}_p^{j*}) = \mathrm{Cost}(\mathcal{S}_0^0) \ .$$

Since $\mathcal{M}_p^{j*}$ is computed with $O(2^{-j}N)$ operations, the optimal model $\widehat{\mathcal{M}}$ model is obtained with $O(J\,N)$ operations using (26). Since $J \leq \log_2 N$, the penalization algorithm which optimizes the model over adaptive segmentations thus requires $O(N\,(\log_2 N)^2)$ operations.

4 Estimation of Covariance of Locally Stationary Processes

The central result of this paper is the application of the penalized macrotile procedure to covariance estimation. Next section summarizes the algorithm and shows that it can be interpreted as a simultaneous estimation of an approximate Karhunen-Loève basis as well as the diagonal coefficients in this basis. The consistency and numerical performance of this estimator is studied for locally stationary processes. Section 4.2 reviews the property of these processes and defines a precise model.

4.1 Macrotile Estimation of Covariance

Let $\mathcal{F}$ be a family of models constructed from a dictionary of orthonormal bases $\mathcal{D} = \{\mathcal{B}_b\}_b$, which are segmented in macrotiles. To estimate the covariance Γ of a process $\widehat{y}[n]$ of size N, according to (17), we compute the best penalized model

$$\widehat{\mathcal{M}} = \arg\min_{\mathcal{M}\in\mathcal{F}} \left\{ -\|P_{\mathcal{M}}(\widehat{\Gamma})\|^2 + C_p \frac{\lambda^2}{R} \left(\log \mathcal{V}^{\#}\right)^2 \dim(\mathcal{M}) \right\} , \qquad (28)$$

where λ is an upper bound of the largest eigenvalue of Γ that we suppose known a priori. We derive from (7) that

$$\|P_{\mathcal{M}}(\widehat{\Gamma})\|^2 = \sum_{k=1}^{K} \frac{1}{I_k^{\#}} \Big| \sum_{m\in I_k} \langle \widehat{\Gamma} g_m^b, g_m^b \rangle \Big|^2 , \qquad (29)$$

where we recall (8), i.e.

$$\langle \widehat{\Gamma} g_m^b, g_m^b \rangle = \frac{1}{R} \sum_{r=1}^{R} |\langle \widehat{y}_r\,, g_m^b \rangle|^2 . \qquad (30)$$

The main difficulty is now to evaluate the resulting risk $\|\Gamma - P_{\mathcal{M}}(\widehat{\Gamma})\|$ depending upon the properties of the process $\widehat{y}[n]$.

Let $\mathcal{B}_{\widehat{b}}$ be the orthonormal basis corresponding to the selected penalized model $\widehat{\mathcal{M}}$. The estimator $P_{\mathcal{M}}(\widehat{\Gamma})$ of Γ is an operator which is diagonal in $\mathcal{B}_{\widehat{b}}$ and whose eigenvalues have been averaged in $\dim(\widehat{\mathcal{M}})$ groups. The best basis $\mathcal{B}_{\widehat{b}}$ can thus be interpreted as an approximate Karhunen-Loève basis of Γ which is optimized to reduce the risk $\|\Gamma - P_{\mathcal{M}}(\widehat{\Gamma})\|$.

The maximization of a Hilbert-Schmidt norm has already been studied for finding an approximate Karhunen-Loève basis in a dictionary of orthonormal bases [13]. This approach does not incorporate simultaneously an appropriate estimation of the diagonal coefficients in the selected basis, and thus can not lead to a consistent estimator. This algorithm has been improved by also estimating the diagonal coefficients with a soft thresholding of the coefficients (8) in a Haar wavelet basis [8]. The performance of this estimator has been studied in the context of locally stationary processes, defined by a model similar to our set-up of the following section, with dictionaries of local cosine bases [8], and some result on the ideal risk of this estimation has been derived. Unfortunately, the soft thresholding introduces too large a bias in the covariance estimation to come close to this ideal risk. As a result, the numerical and mathematical consistency results are not significantly better than the ones obtained with simpler linear estimator, and hence the proposed non-linear method of [8], could not be shown to reach the near-optimal risk. In the following, we evaluate both numerically and mathematically the efficiency of the macrotile penalized estimator, in the context of locally stationary processes.

4.2 Locally Stationary Processes

Locally stationary processes appear in many physical systems where random fluctuations are produced by a mechanism which varies slowly in time or which has few brutal transitions. Such processes can be locally approximated by stationary processes. A locally stationary process can thus be qualitatively defined as a process which is approximately stationary over sufficiently short time intervals.

A process $\widehat{y}[n]$ for $n \in \mathbb{Z}$ is wide-sense stationary if its covariance $\gamma[n,m] = \mathsf{E}\{\widehat{y}[n]\,\widehat{y}[m]\}$ satisfies $\gamma[n,m] = c[n-m]$. The corresponding covariance operator Γ is a convolution operator, diagonalized in the Fourier basis, and whose eigenvalues (power spectrum) are given by

$$P(\omega) = \sum_{\tau} c[\tau]\, e^{-i\,\omega\,\tau} \ .$$

If the process is not strictly stationary, we can still write the covariance

$$c[n,\tau] = \gamma[n, n-\tau] \ .$$

Priestley [19], Martin and Flandrin [16] have proposed several definitions of locally stationary processes by introducing a time-varying power spectrum defined by

$$P[n,\omega] = \sum_{\tau\in\mathbb{Z}} c[n,\tau]\, e^{-i\omega\,\tau} \ .$$

It can also be rewritten as the expected value of the Wigner-Ville distribution of the process $\widehat{y}$:

$$\widetilde{P}[n,\omega] = \mathsf{E}\left\{ \sum_{\tau\in\mathbb{Z}} \widehat{y}[n]\,\widehat{y}[n+\tau]\, e^{-i\,\omega\,\tau} \right\} \ .$$

If the process is locally stationary then one can expect that $P[n,\omega]$ has slow variations as a function of the position parameter n. If $\widehat{y}$ also has a fast decorrelation, which means that $c[n,\tau]$ decays quickly when τ increases, then $P[n,\omega]$ is a smooth function along the frequency variable ω. This time-varying spectrum has been further studied in [7, 18], and in [13], where it is observed that if $P[n,\omega]$ is nearly constant over the time-frequency Heisenberg rectangle of a local cosine vector, then this vector is nearly an eigenvector of the covariance operator Γ. More precisely, if $P[n,\omega]$ is approximatively equal to a constant $P_{p,q}$ over the Heisenberg rectangle $[a_p, a_{p+1}] \times [q\pi/l_p\, ,\, (q+1)\pi/l_p]$ of a local cosine vector $g_{p,q}[n]$ defined in (21) then

$$\Gamma g_{p,q}[n] \approx P_{p,q}\, g_{p,q}[n] \ .$$

In this case, the eigenvalues are given by the time-varying power spectrum.

Despite the interesting properties of the time-varying spectrum, we formulate the conditions to develop consistency of the statistical estimation procedure not in the spectral domain (as was done in [7] and [18]). We rather characterize the local stationarity by imposing directly a condition on the covariance $c[n,\tau]$ as in [8] - a condition which has the advantage that it remains non-asymptotic in its nature.

Being locally stationary means that $c[n,\tau]$ can be approximated by the covariance of a stationary processes $c_m[\tau]$ for n in a neighborhood of a point m. We thus need a condition that expresses the fact that $c[n,\tau]$ can be approximated by a $\tilde{c}[n,\tau]$ which is piecewise-constant when n varies, over intervals that are the same for all τ. The choice of the stationarity intervals depends upon each particular locally stationary process. A sufficient condition imposes that $c[n,\tau]$ has a uniformly bounded variation when n varies [8]. This means that there exists a constant Q_1 such that

$$\sum_{n=1}^{N-1} \|c[n+1,.]-c[n,.]\| \leq Q_1 \ , \tag{31}$$

with

$$\|c[n+1,.]-c[n,.]\|^2 = \sum_{\tau\in\mathbb{Z}} |\, c[n+1,\tau]-c[n,\tau]\,|^2 \ .$$

This condition allows some brutal variations of $c[n,\tau]$ along n, but few of them so that the total variation remains bounded. A standard approximation result [8] shows that this condition is sufficient to impose a good approximation of $c[n,\tau]$ with a signal $\tilde{c}[n,\tau]$ which is piecewise-constant as a function of n.

To further characterize the class of process, it is necessary to impose a condition on the decorrelation of $\widehat{y}[n]$ and hence on the decay of $c[n,\tau]$ as a function of τ. We first consider processes which have a fast decorrelation, such that there exists a constant Q_2 and $1/2<\beta<1$ with

$$\forall n \ \ , \ \ \sum_{\tau\in\mathbb{Z}} (1+2|\tau|^{\beta})^2\, |c[n,\tau]|^2 \leq Q_2 \ . \tag{32}$$

One can prove that if (31) and (32) are satisfied then the largest eigenvalue λ of Γ is bounded by a constant that depends only upon Q_1, Q_2 and β. The fast decorrelation locally stationary model corresponding to the conditions (31) and (32) is appropriate for a wide range of signals, including unvoiced speech recordings or seismic signals [1]. This model is used to prove the consistency of macrotile estimation of covariance operators, with local cosine bases [22].

Not all locally stationary processes satisfy the fast decorrelation condition (32). For examples, the sounds produced by wind instruments can be modeled as realizations of locally stationary processes, yet the existence of narrow frequency peaks indicates the existence of long range correlations within the signal. In this case, the fast decorrelation property (32) of $c[n,\tau]$, which implies that the time-varying spectrum $P[n,\omega]$ has uniformly smooth variation along ω, has to be replaced by a less stringent assumption which admits the existence of local peaks. One approach is to impose a weaker regular property, namely that $P[n,\omega]$ has bounded variations along ω

$$\forall n \in [1,N], \ \ \int_{-\pi}^{\pi} \left| \frac{dP[n,\omega]}{d\omega} \right| d\omega \leq Q_2 \ . \tag{33}$$

Weak regularity properties similar to this one have already been used to construct consistent estimators of the power spectrum of stationary processes [11].

4.3 Estimation of Locally Stationary Covariances

Let us now study the covariance estimation of locally stationary processes with a penalized macrotile algorithm. This section focuses on locally processes that have a fast decorrelation and whose covariance thus satisfy (31) and (32). An appropriate family of local cosine macrotile models is described together with a fast algorithm to compute the best penalized model.

For locally stationary processes, we mentioned that a local cosine vector is nearly the eigenvector of locally stationary covariance if over its time-frequency Heisenberg rectangle the time-varying power spectrum $P[n,\omega]$ remains nearly constant. This constant is the eigenvalue associated to this eigenvector. It has thus been proposed to search for an approximate Karhunen-Loève basis among a dictionary of local cosine bases [13]. To specify a family of macrotiles $\mathcal{F}$ we still need to choose the set of all possible segmentations for each basis. For processes with a fast decorrelation, the time-varying power spectrum $P[n,\omega]$ is uniformly smooth as a function of ω. This suggests to average uniformly the sample-mean estimator to estimate the diagonal covariance coefficients.

We saw in (22) that a local cosine basis can be written $\mathcal{B}_b = \{\mathcal{B}^{j_l}_{p_l}\}_{1\leq l\leq L}$ where each $\mathcal{B}^j_p$ regroups $2^{-j}N$ cosine vectors whose support are $[p\,2^{-j}\,N-\eta\,,\ (p+1)\,2^{-j}\,N+\eta]$. Among these vectors, we choose segmentations that perform a uniform averaging. This means that local cosine coefficients in this interval are averaged along frequencies in consecutive groups of 2^s coefficients, where 2^s is a fixed length with $1 \leq 2^s \leq 2^{-j}N$. Observe that this is more restrictive than in the local cosine macrotile family defined in Section 3.1, where the segmentation was not necessarily uniform. The number of macrotile models is therefore smaller, but the total number $\mathcal{V}^{\#}$ of macrotiles combined to construct these models is the same in both cases and $\log_2 \mathcal{V}^{\#}$ is nearly equal to $\log_2 N$.

Following (28) the best local cosine macrotile $\widehat{\mathcal{M}}$ is calculated by minimizing

$$\mathrm{Cost}(\mathcal{M}) = -\|P_{\mathcal{M}}(\widehat{\Gamma})\|^2 + C_p\,\frac{\lambda^2}{R}\,\Big(\log N\Big)^2\,\dim(\mathcal{M})\ . \qquad (34)$$

Each model $\mathcal{M}$ is specified by a local cosine basis $\mathcal{B}_b$ and its segmentation. The norm $\|P_{\mathcal{M}}(\widehat{\Gamma})\|^2$ is computed from the decomposition coefficients of each realization $\widehat{y}_r$ in the basis $\mathcal{B}_b$ with (29) and (30). The minimization of this cost in a local cosine dictionary is implemented with minor modifications of the fast algorithm described in Section 3.2, because it has the same additivity properties.

The best model $\widehat{\mathcal{M}}$ is constructed from partial optimal models $\mathcal{M}_p^j$ associated to the local cosine spaces $\mathbf{V}_p^j$ along the tree nodes, using the aggregation property (26). The only difficulty is to construct the model $\mathcal{M}_p^{j*}$ where the segmentation S_p^j is adjusted to minimize the cost associated with the basis $\mathcal{B}_p^j$.

For each realization $\widehat{y}_r$ of the process, we first compute the $J\,N$ coefficients $\langle \widehat{y}_r, g_{p,q,j}\rangle$ corresponding to all local cosine vectors $g_{j,q,p}$ in the tree depth J. With a fast lapped cosine transform [15], for R realizations, it requires $O(R\,J\,N\,\log_2 N)$ operations. Following (30) we then compute

$$\widehat{\alpha}_p^j[q] = \langle \widehat{\Gamma} g_{p,q,j}\,,\, g_{p,q,j}\rangle = \frac{1}{R}\sum_{r=1}^{R} |\langle \widehat{y}_r, g_{p,q,j}\rangle|^2$$

with $O(R\,J\,N)$ operations.

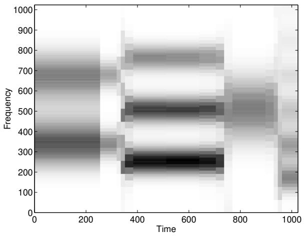
To compute the segmentation S_p^j which minimizes the cost associated with the basis $\mathcal{B}_p^j$ is a much simpler operation than in Section 3.2. Indeed, we are restricted to uniform segmentations which average the $2^{-j}N$ coefficients $\widehat{\alpha}_p^j[q]$ by intervals of same size 2^s, for any $1 \leq 2^s \leq 2^{-j}N$. The resulting averaged coefficients are

$$\bar{\alpha}_s[k] = \frac{1}{2^s}\sum_{q=k2^s+1}^{(k+1)2^s} \alpha_p^j[q] \quad \text{for } 0 \leq k \leq (2^{-s-j}N - 1)\ .$$

With a standard dyadic averaging cascade by groups of 2, this is calculated for all $0 \leq s \leq \log_2 N - j$ with $O(2^{-j}N)$ operations. The cost associated with the only model $\mathcal{M}_s$ of dimension $K = 2^{-s-j}N$ is

$$\mathrm{Cost}(\mathcal{M}_s) = -2^s\sum_{k=1}^{K} |\bar{\alpha}_s[k]|^2 + C_p\,(\log N)^2\frac{K}{R}\ .$$

Finding the model $\mathcal{M}_p^{j*}$ that minimizes this cost over s thus requires $O(2^{-j}N)$ operations. The total number of operations to compute all models $\mathcal{M}_p^{j*}$ in the tree is thus $O(J\,N)$. With the aggregation property (26) the optimal model $\widehat{\mathcal{M}}$ is then also computed with $O(J\,N)$ operations. The overall complexity of the penalization algorithm is therefore dominated by the number


Frequency
1000
900
800
700
600
500
400
300
200
100
0
0
200
400
600
800
1000
Time