Sparsity and persistence in time-frequency sound representations

Matthieu Kowalski, Bruno Torrésani

Sparsity and persistence in time-frequency sound representations

Matthieu Kowalski[♯†] and Bruno Torrésani[♯]

[*] LATP, Université de Provence, CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France;
[†]L2S, Supelec, Plateau du Moulon, 91192 Gif sur Yvette, France[*]

ABSTRACT

It is a well known fact that the time-frequency domain is very well adapted for representing audio signals. The main two features of time-frequency representations of many classes of audio signals are sparsity (signals are generally well approximated using a small number of coefficients) and persistence (significant coefficients are not isolated, and tend to form clusters). This contribution presents signal approximation algorithms that exploit these properties, in the framework of hierarchical probabilistic models.

Given a time-frequency frame (i.e. a Gabor frame, or a union of several Gabor frames or time-frequency bases), coefficients are first gathered into groups. A group of coefficients is then modeled as a random vector, whose distribution is governed by a hidden state associated with the group.

Algorithms for parameter inference and hidden state estimation from analysis coefficients are described. The role of the chosen dictionary, and more particularly its structure, is also investigated. The proposed approach bears some resemblance with variational approaches previously proposed by the authors (in particular the variational approach exploiting mixed norms based regularization terms).

In the framework of audio signal applications, the time-frequency frame under consideration is a union of two MDCT bases or two Gabor frames, in order to generate estimates for tonal and transient layers. Groups corresponding to tonal (resp. transient) coefficients are constant frequency (resp. constant time) time-frequency coefficients of a frequency-selective (resp. time-selective) MDCT basis or Gabor frame.

Keywords: Audio signals, time-frequency, sparse and structured approximation, hierarchical models

1. INTRODUCTION

Given a separable Hilbert space, and a dictionary (i.e. essentially a complete set of vectors) in this Hilbert space, a vector in this space is said to be sparsely represented in the dictionary when it may be expanded as a linear combination of the elements of the dictionary; in such a way that only a small percentage of the expansion coefficients is nonzero (or numerically significant). Sparsity has become a key concept in many domains of modern signal processing.

Audio signals, in particular musical signals, are known to possess such a sparsity property when an appropriate time-frequency dictionary is used (see for example [1] for a review). This is mainly a consequence of the way most audio signals are produced, i.e. involving resonating physical systems. This results in sounds that can naturally be decomposed as sums of (possibly delayed, damped, chirped...) sinusoids (see [2, 3] and references therein), together with additional components such as transients (sharply time-localized) and stochastic-like components. However, such elementary buiding blocks cannot be cast as *time-frequency atoms* in the general sense, as they are in some sense more macroscopic objects. In addition, there is already a vast variability within such a class building blocks, which makes it difficult to parameterize it (see however [4] for an example in particular situations).

[*] : starting September 2009
Further author information: (Send correspondence to B.T.)
M.K.: E-mail: kowalski@cmi.univ-mrs.com, Telephone: +33 (0)4 91 05 47 43
B.T.: E-mail: Bruno.Torresani@cmi.univ-mrs.fr, Telephone: +33 (0)4 91 05 46 78

In such situations, it makes sense to turn to simpler building blocks, (*sound atoms*), and gather them into more macroscopic sound objects, that could be called *sound molecules*. A sound molecule can roughly be defined as a linear combination of atoms, with variable coefficients. The underlying models are therefore hierarchical models, in which both groups and coefficients for a given group are to be modeled. Given such models, the problem is to find sparse signal expansions with respect to a given dictionary of atoms, that respect the molecule organizations.

Several approaches can be followed for such a sparse regression problem. Since they often lead to simple and efficient algorithms, variational approaches are currently very popular, and can be adapted to such a situation in various ways (see for example [5–9] and references therein). Pursuit methods also provide simple algorithmic approaches, that can also be adapted to the hierarchical situation (see [10, 11]). In this paper, we shall focus on explicit hierarchical modeling using probabilistic approaches, and classification of the so-called *analysis coefficients*, following the approach of [12]. Such approaches involve signal models of the form

$$x = \sum_{\lambda \in \Lambda} \alpha_\lambda \varphi_\lambda + r \ , \tag{1}$$

where the atoms $\{\varphi_\lambda, \lambda \in \Lambda\}$ constitute a dictionary in some reference Hilbert space, and the coefficients α_λ, called *synthesis coefficients* are random variables, whose distribution is controlled by some hidden state

$$X_\lambda \in \{s_0, \dots s_K\} \ .$$

r is some residual, which is not supposed to be sparse, and is generally modelled as Gaussian white noise.

The index set Λ can be equipped with a *structure*, which can take several forms, including

- A *hierarchy*, or *stratification*, defined in terms of groups and members. Namely, an index value is actually a pair $\lambda = (g, m)$, where g stands for "group" and m stands for "member". Coefficients which are members of the same group are assumed to be statistically dependent, while coefficients belonging to different groups are independent.
- A neighborhood system, assigning to each λ a neighborhood, neighboring coefficients being again assumed dependent. We shall not follow this way here, and we shall stick to the stratification approach.

The dependencies can be introduced in various ways. We shall describe here two approaches, that assume either dependent hidden states with uncorrelated synthesis coefficients (conditional to the hidden states, following [12–14]), or simplified hidden states (labelled only by groups) with "within group" correlations between synthesis coefficients. This approach somewhat extends the hierarchical Bernoulli model proposed in [12].

A main objective in such a context is to identify from a realization of the signal model the corresponding realization of hidden states. Assuming this goal has been reached, a multilayered decomposition of the signal can be obtained, by partial resynthesis from coefficients that are in the same state.

For sound signals, the dictionaries under consideration are generally redundant time-frequency dictionaries. Therefore, for a given signal x, the expansion (1) is not unique. In [14], a MCMC (Markov chain Monte Carlo) approach was proposed for calculating MMSE (Minimum Mean Squared Error) estimates; we shall rather focus here on estimates obtained from the *analysis coefficients*

$$a_\lambda = \langle x, \varphi_\lambda \rangle \ , \tag{2}$$

and study conditions under which the identification of hidden states (i.e. smoothing) is possible from these coefficients. After the identification of hidden states, a multilayered expansion of the signal is obtained by partial synthesis from coefficients associated with a fixed hidden state:

$$x_k = \sum_{\lambda : X_\lambda = s_k} \alpha_\lambda \varphi_\lambda \ . \tag{3}$$

In the context of sound signal decomposition, such multilayered expansions have been used to separate tonal, transient and "stochastic" layers. Several applications of such decompositions can be mentioned, including audio signal analysis (the physical characteristics of the systems that produced sound may be more easily readable from the layers after separation), coding (the different layers are most efficiently encoded in different waveform systems), transformation,...

We describe in this paper a general framework for identifying hidden states from the analysis coefficients, and solve the corresponding sparse regression problem. We first describe the hierarchical models we are interested in, and study the behavior of corresponding analysis coefficients, in mainly two situations: unstructured dictionaries, and unions of orthonormal bases. We then describe corresponding estimation algorithms, and conclude with numerical results on sound signals.

2. HIDDEN STATES MODELS

We discuss in this section the general sparse regression problem, and its adaptation to the molecular case. Let $\mathcal{H}$ denote a (finite or infinite dimensional) separable real Hilbert space, and let $\mathcal{D} = \{\varphi_\lambda, \lambda \in \Lambda\}$ denote a complete dictionary in $\mathcal{H}$. Here, Λ denotes a generic index set. We shall assume that the dictionary is a normalized tight frame in $\mathcal{H}$, i.e. that $\|\varphi_\lambda\| = 1$ for all λ, and that for all $x \in \mathcal{H}$, one has the Parseval identity

$$\sum_{\lambda \in \Lambda} |\langle x, \varphi_\lambda \rangle|^2 = A\|x\|^2 \tag{4}$$

for some constant $A > 0$. If $\mathcal{D}$ is not a basis in $\mathcal{H}$, any $x \in \mathcal{H}$ admits infinitely many expansions in the form given in (1).

We are interested in signals that are *sparse* in the considered dictionary, i.e. signals $x \in \mathcal{H}$ which admit an expansion (1) involving only a small number of nonzero (or significant) synthesis coefficients. The corresponding index set is termed *significance map*.

Given such a sparse signal, the non-uniqueness of its expansion with respect to the dictionary makes it difficult to identify unambiguously the model (1). The approach we propose uses the *analysis coefficients* (2) and develops a strategy to estimate the relevant such coefficients, from which a sparse expansion may be identified. Namely, it may be proved that under suitable assumptions on the dictionary and the sparsity of the expansion, the analysis coefficients may be used to locate the significant synthesis coefficients, and therefore estimate a sparse signal expansion.

In the numerical applications to be described later, we shall often limit ourselves to unions of bases (the dictionary $\mathcal{D}$ is the union of two orthonormal bases $\mathcal{B}^1$ and $\mathcal{B}^2$), and to a specific pair of orthonormal bases: $\mathcal{B}^1$ is a local trigonometric (i.e. an MDCT –Modified Discrete Cosine Transform– basis, see for example [15]) basis (tuned in such a way to achieve good frequency resolution), and $\mathcal{B}^2$ is a local trigonometric basis with good time resolution. The index sets are then two-dimensional (a time index and a frequency index), and we write them as such when necessary. Other choices for the bases are possible (for example a combination of MDCT and wavelet bases, as in [13, 16]), as well as extensions to frames (that would however require significant modifications).

2.1 Hierarchical signal random models

Let us now introduce an explicit *model* for the sparse signal in (1). The ingredients of such models are essentially twofold: a model for the set of *hidden states* $\mathbf{X}$ and, conditional to $\mathbf{X}$, a model for the synthesis coefficients.

Definition 1. *Given a dictionary $\mathcal{D} = \{\varphi_\lambda, \lambda \in \Lambda\}$ of the Hilbert space $\mathcal{H}$ as above, a corresponding* hierarchical random model *is defined by*
i. *A discrete probability model for the significance map. The corresponding probability measures for the (random) map $\mathbf{X}$ will be denoted by $\mathbb{P}_{\mathbf{X}}$, and the expectations by $\mathbb{E}_{\mathbf{X}}$.*
ii. *A probability model for the synthesis coefficients $\{\alpha_\lambda, \lambda \in \Lambda\}$, conditional to the hidden states. The corresponding probability measure and expectation are denoted by $\mathbb{P}_0$ and $\mathbb{E}_0$. The global probability measure and expectation will be denoted by $\mathbb{P}$ and $\mathbb{E}$ respectively.*

The corresponding signal model takes the form (1).

The coefficients α_λ above are the *synthesis coefficients*. The *analysis coefficients* are given by

$$a_\lambda = \langle x, \varphi_\lambda \rangle = \sum_\mu \alpha_\mu \langle \varphi_\mu, \varphi_\lambda \rangle + \langle r, \varphi_\lambda \rangle \ . \tag{5}$$

For now on, we shall limit ourselves to the case of zero-mean, Gaussian synthesis coefficients. In such situations, the analysis coefficients are linear combinations of zero-mean, jointly Gaussian random variables: conditional to the hidden states, an analysis coefficient is a zero-mean Gaussian random variable, whose variance depends on the hidden states realization. More precisely, introducing the Gram matrix G of the dictionary (*i.e.* $G_{ij} = \langle \varphi_i, \varphi_j \rangle$), the analysis covariance matrix $\mathcal{C}^{(a)}_{\mathbf{X};\lambda\mu} = \mathbb{E}_0 \{a_\lambda \overline{a}_\mu\}$ is related to the synthesis covariance matrix $\mathcal{C}^{(s)}_{\mathbf{X};\lambda\mu} = \mathbb{E}_0 \{\alpha_\lambda \overline{\alpha}_\mu\}$ (remember that these matrices are random because they are defined conditional to the hidden states $\mathbf{X}$) by

$$\mathcal{C}^{(a)}_{\mathbf{X}} = G^T \mathcal{C}^{(s)}_{\mathbf{X}} G^T \ .$$

Clearly enough, to identify hidden states from analysis coefficients, $\mathcal{C}^{(a)}_{\mathbf{X}}$ should reflect the structure of $\mathcal{C}^{(s)}_{\mathbf{X}}$. This imposes constraints on the Gram matrix, and thus on the dictionary. We shall analyze below such requirements in more specific situations.

The case of structured dictionaries. A particular case of interest is the case where the dictionary $\mathcal{D}$ is *structured* as the union of several orthonormal bases (for example, two local cosine bases with different time-frequency resolutions, as described in the introduction):

$$\mathcal{D} = \mathcal{B}^1 \cup \mathcal{B}^2 \cup \cdots \cup \mathcal{B}^K \ , \qquad \text{with} \quad \mathcal{B}^k = \{\psi^k_\ell, \ \ell = 1, 2, \dots\}$$

In such cases, the Gram matrix inherits a simpler block structure

$$G = \begin{pmatrix} I & \tilde{G}^{12} & \dots & \tilde{G}^{1K} \\ \tilde{G}^{21} & I & \dots & \tilde{G}^{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{G}^{K1} & \tilde{G}^{K2} & \dots & I \end{pmatrix} \ , \qquad \text{with} \quad I_{\lambda\mu} = \delta_{\lambda\mu} \ , \quad \tilde{G}^{kk'}_{\ell\ell'} = \langle \psi^k_\ell, \psi^{k'}_{\ell'} \rangle$$

Sparse significance map models. Sticking to the problematics of sparse signal expansion, we shall limit the present discussion to models in which coefficients are either significant or vanish. State s_0 corresponds to vanishing coefficients, while states $s_1, \dots s_K$ correspond to random synthesis coefficients with nonzero variances, denoted by $\sigma_1^2, \dots \sigma_K^2$. The significance map is the subset of the index state corresponding to states s_k, $k \neq 0$.

We shall mainly consider the two simpler models

M1. When the dictionary is not structured: we set

$$\mathbb{P}\{X_\lambda = s_k\} = p_k \ , \qquad \mathbb{P}\{X_\lambda = s_0\} = p_0 = 1 - \sum_{k=1}^K p_k \ . \tag{6}$$

For simplicity, we set $\sigma_0 = 0$. The simplest such model is a two-state model ($K = 1$).

M2. For structured dictionaries, for example unions of orthonormal bases, we introduce as many states as bases, and set

$$\mathbb{P}\big\{X^k_\ell = s_{k'}\big\} = p_k \delta_{kk'} + (1 - p_k)\delta_{k'0} \tag{7}$$

The model is thus as follows: a signal is expanded in the form

$$x = \sum_{k=1}^K \sum_\ell \alpha^k_\ell \psi^k_\ell + r \ , \tag{8}$$

and the distribution of α^k_ℓ is governed by X^k_ℓ.

In what follows, the residual signal (noise) is modelled as a Gaussian white noise with variance ϑ^2.

2.2 Behavior of analysis coefficients

As mentioned earlier, in such models, the analysis coefficients are zero mean, correlated Gaussian random variables. Their covariance matrix has a somewhat complicated expression, which we study now in a couple of specific situations. Let us recall that the goal is to estimate model parameters (variances σ_k^2 and probabilities p_k), and identify "active" atoms, i.e. atoms φ_λ such that $X_\lambda \neq s_0$. The generic algorithmic structure we shall work with is (Classification) Expectation Maximization –EM or CEM– type methods [17], in which parameter and hidden states estimates are recursively refined from results of previous iterations. We derive below conditional estimates that can be plugged in such algorithms.

2.2.1 Conditional independent synthesis coefficients

We follow the analysis of [12, 13] in which the synthesis coefficients were supposed independent conditional to the hidden states. In such a situation, it was shown that the analysis coefficients follow a Gaussian mixture distribution, from which the hidden states may sometimes be estimated. We now briefly describe this situation.

Let us then assume that conditional to the hidden states, the synthesis coefficients are either independent zero mean Gaussian variables, with variances $\sigma^2_{X_\lambda}$, or vanish. By convention, we denote $\sigma_{s_0} = 0$. Clearly,

$$\mathbb{E}_{\mathbf{X}}\{\sigma^2_{X_\lambda}\} = \sum_{k=1}^{K} p_k \sigma_k^2 \ .$$

It follows from the analysis above that the synthesis coefficients are sums of independent zero-mean Gaussian random variables, and are therefore (dependent) zero-mean Gaussian random variables, whose covariance depends upon the hidden states $\mathbf{X}$, and the redundancy of the dictionary

$$\mathbb{E}_0\left\{a_\lambda \overline{a}_{\lambda'}\right\} = \sum_{\mu} \sigma^2_{X_\mu} \langle \varphi_{\lambda'}, \varphi_\mu\rangle\langle \varphi_\mu, \varphi_\lambda\rangle + \vartheta^2 \ . \tag{9}$$

Disregarding the correlations between analysis coefficients, pick a fixed λ, and denote for the sake of simplicity by $\mathbf{X}_{-\lambda} = \{X_{\lambda'}, \lambda' \neq \lambda\}$ the set of all hidden states except X_λ. We have that

$$\mathbb{E}_0\left\{|a_\lambda|^2\right\} = \sigma^2_{X_\lambda} + \gamma_\lambda(\mathbf{X}_{-\lambda}) + \vartheta^2 = \left\{ \begin{array}{ll} \gamma_\lambda(\mathbf{X}_{-\lambda}) + \vartheta^2 & \text{if } X_\lambda = s_0 \\ \sigma^2_\lambda + \gamma_\lambda(\mathbf{X}_{-\lambda}) + \vartheta^2 & \text{otherwise} \end{array} \right. , \tag{10}$$

where we have introduced the random variables γ_λ (called *gamma weights* in [12]), defined by (recall that by convention, $\sigma_{s_0} = 0$)

$$\gamma_\lambda(\mathbf{X}_{-\lambda}) = \sum_{\mu \neq \lambda} \sigma^2_{X_\mu} |\langle \varphi_\lambda, \varphi_\mu\rangle|^2 \ . \tag{11}$$

Clearly, for all λ one has the bound

$$\gamma_\lambda(\mathbf{X}_{-\lambda}) \leq A \max_k \sigma_k^2 \ , \tag{12}$$

A being the frame constant, see (4).

The distribution of analysis coefficients, and thus the possibility of discriminating between $X_\lambda = s_0$ and $X_\lambda \neq s_0$ depends clearly on the distribution of the γ variables. An explicit calculation yields

LEMMA 1. *The first moments of the γ weights read*

$$\mathbb{E}_{\mathbf{X}}\left\{\gamma_\lambda(\mathbf{X}_{-\lambda})\right\} = (A-1)\sum_{k=1}^{K} p_k \sigma_k^2$$

and their variance

$$\mathrm{var}_{\mathbf{X}}\{\gamma_\lambda(\mathbf{X}_{-\lambda})\} = \mathrm{var}_{\mathbf{X}}\{\sigma^2_{X_\lambda}\} \sum_{\mu \neq \lambda} |\langle \varphi_\lambda, \varphi_\mu\rangle|^4$$

As we shall see, the smaller the latter quantity compared with $\sigma^2_{X_\lambda}$, the easier the estimation of X_λ. It is interesting to look at the quantities that control the sizes of these averages. Low values of $\mathbb{E}_{\mathbf{X}}\{\gamma_\lambda(\mathbf{X}_{-\lambda})\}$ can be obtained with small values of the variances σ_k^2 and small values for the probabilities p_k, $k \neq 0$ (which is a sparsity requirement) and a small value for the frame constant A (which is a low coherence assumption). Small values for the variance of γ_λ are ensured by small values of the 4-Babel function

$$B_4 = \max_\lambda \sum_{\mu \neq \lambda} |\langle \varphi_\lambda, \varphi_\mu \rangle|^4$$

(which is also a low coherence assumption).

We now specialize to the case $K = 1$ for the sake of simplicity. Noticing that conditional to $\mathbf{X}_{-\lambda}$, the distribution of the analysis coefficient is a Gaussian mixture

$$a_\lambda | \mathbf{X}_{-\lambda} \sim p\mathcal{N}(0, w^2_{\lambda,1}) + (1-p)\mathcal{N}(0, w^2_{\lambda,0}) \ , \tag{13}$$

where we have introduced the new (random) variances

$$w^2_{\lambda;0} = \gamma_\lambda(\mathbf{X}) + \vartheta^2 \ , \qquad w^2_{\lambda;1} = w^2_{\lambda;0} + \sigma_1^2 \ . \tag{14}$$

For simplicity, let us introduce the following threshold function

$$\tau_0 : (w_0, w_1, p) \longmapsto \sqrt{\ln\left[\frac{1-p}{p}\frac{w_1}{w_0}\right]} \ , \tag{15}$$

which is well-defined as soon as $p \in (0,1)$ and

$$w_1 > \frac{p}{1-p}\, w_0 \ . \tag{16}$$

Let us also denote by $\delta(w_0, w_1)$ the harmonic difference of the squared numbers w_0 and w_1 (such that $w_1 > w_0$)

$$\delta(w_0, w_1) = \sqrt{\frac{2}{w_0^{-2} - w_1^{-2}}} \ . \tag{17}$$

The conditional MAP estimate for X_λ can be obtained as follows

PROPOSITION 1. *Consider model M1, with $K = 1$, and assume i.i.d. hidden states X_λ. Assume that for all λ, condition* (16) *is satisfied. With the notations above, set*

$$\tau(\lambda) = \delta(w_{\lambda;0}, w_{\lambda;1})\tau_0(w_{\lambda;0}, w_{\lambda;1}, p) \ .$$

Then the maximum a posteriori estimator for X_λ conditional to $\mathbf{X}_{-\lambda}$ is given by

$$\hat{X}_\lambda = \begin{cases} s_1 & \text{if } |a_\lambda| \geq \tau(\lambda) \\ s_0 & \text{otherwise} \end{cases}$$

The proof can be adapted from Proposition 3 of [12]. Type I and II error rates can also be derived from this result.

REMARK 1. Estimates for the variance of the γ weights are important for the following reason. If $\mathrm{var}_{\mathbf{X}}\{\gamma_\lambda(\mathbf{X}_{-\lambda})\}$ is large, then the variance of the thresholds τ_λ is large too, which results in high error rates.

Given estimates $\hat{\mathbf{X}}$ for the hidden states, estimates for the other parameters σ_1 and p are readily obtained:

$$\hat{p} = \frac{\#\{\lambda : \hat{X}_\lambda = s_1\}}{\dim(\mathcal{H})} \ , \qquad \hat{\sigma}_1^2 = \frac{1}{\hat{p}\dim(\mathcal{H})} \sum_{\lambda : \hat{X}_\lambda = s_1} \left(|a_\lambda|^2 \ - \gamma_\lambda(\hat{\mathbf{X}}_{-\lambda}) - \vartheta^2 \right) \ . \tag{18}$$

It is worth noticing that in such a scheme, the noise variance ϑ^2 has to be known in advance. It must then be estimated separately, or used as a tuning parameter that controls the sparsity of the expansion.

REMARK 2. Replacing $\gamma_\lambda(\mathbf{X}_{-\lambda})$ by its expectation $\mathbb{E}_{\mathbf{X}}\{\gamma_\lambda(\mathbf{X}_{-\lambda})\}$ given in Lemma 1 yields simpler estimates, which we call *mean field estimates*.

This naturally leads to a simple CEM algorithm.

ALGORITHM 1.

- *Initialize the parameters σ_1 and p and the hidden states $\mathbf{X}$, using the mean field estimates.*
- *Iterate the following steps*
  1. *Compute the $\gamma_\lambda(\mathbf{X}_{-\lambda})$*
  2. *Re-estimate hidden states* (Maximization Step. Can be done by classification)
  3. *Re-estimate parameters σ_1 and p* (Expectation Step).
- *Estimate the significant coefficients α_λ with $X_\lambda = s_1$ by regression.*

The situation becomes simpler in the case of a structured dictionary, as in model M2. Assume $\mathcal{D}$ is the union of K orthonormal bases. Then we have

$$\left\{ \begin{array}{ll} \alpha_k^\ell \sim \mathcal{N}(0, \sigma_k^2) & \text{if } X_\ell^k = s_k \\ \alpha_k^\ell = 0 & \text{otherwise} , \end{array} \right. \tag{19}$$

and as before, conditional to the hidden states $\mathbf{X}_{-k} = \{X_{\ell'}^{k'},\ k' \neq k\}$, the analysis coefficients $a_\ell^k = \langle x, \psi_\ell^k \rangle$ are distributed according to a Gaussian mixture, with zero mean and variances

$$\mathbb{E}_0\left\{|a_\ell^k|^2\right\} = \sigma^2_{X_\ell^k} + \gamma_\ell^k(\mathbf{X}_{-k}) + \vartheta^2 = \left\{ \begin{array}{ll} \gamma_\ell^k(\mathbf{X}_{-k}) + \vartheta^2 & \text{if } X_\ell^k = s_0 \\ \sigma_k^2 + \gamma_\ell^k(\mathbf{X}_{-k}) + \vartheta^2 & \text{if } X_\ell^k = s_k \end{array} \right. , \tag{20}$$

where as before we have introduced random variables $\gamma_\ell^k(\mathbf{X}_{-k})$, defined by

$$\gamma_\ell^k(\mathbf{X}_{-k}) = \sum_{k' \neq k} \sum_{\ell'} \sigma^2_{X_{\ell'}^{k'}} |\langle \psi_\ell^k, \psi_{\ell'}^{k'} \rangle|^2 . \tag{21}$$

LEMMA 2. *Assume that for each k, the X_ℓ^k are i.i.d. Then the first moments of the $\gamma_\ell^k(\mathbf{X}_{-k})$ weights read*

$$\mathbb{E}_{\mathbf{X}}\{\gamma_\ell^k(\mathbf{X}_{-k})\} = \sum_{k' \neq k} p_k \sigma_k^2 \sum_{\ell'} |\langle \psi_\ell^k, \psi_{\ell'}^{k'} \rangle|^2 = \sum_{k' \neq k} p_{k'} \sigma_{k'}^2 ,$$

and

$$\mathrm{var}_{\mathbf{X}}\{\gamma_\ell^k(\mathbf{X}_{-k})\} = \sum_{k'} \sigma_{k'}^4 p_{k'}(1 - p_{k'}) \sum_{\ell'} |\langle \psi_{\ell'}^{k'}, \psi_\ell^k \rangle|^4 .$$

As before, conditional MAP estimates for the hidden states are obtained by adaptive thresholding. Set

$$w_{0;\ell}^k = \sqrt{\gamma_\ell^k(\mathbf{X}_{-k}) + \vartheta^2} , \qquad w_{1;\ell}^k = \sqrt{\sigma_k^2 + \gamma_\ell^k(\mathbf{X}_{-k}) + \vartheta^2} . \tag{22}$$

PROPOSITION 2. *Consider model M2, and assume i.i.d. hidden states X_ℓ^k. Assume that for all k, ℓ, parameters $w_{0;\ell}^k, w_{1;\ell}^k, p_k$ are such that condition* (16) *is fulfilled. With the notations above, set*

$$\tau^k(\ell) = \delta\left(w_{0;\ell}^k, w_{1;\ell}^k\right) \tau_0\left(w_{0;\ell}^k, w_{1;\ell}^k, p_k\right) .$$

Then the maximum a posteriori estimator for X_ℓ^k conditional to $\mathbf{X}_{-k}$ is given by

$$\hat{X}_\ell^k = \begin{cases} s_k & \text{if } |a_\ell^k| \geq \tau^k(\ell) \\ s_0 & \text{otherwise} \end{cases}$$

As before, estimates for the parameters (except the noise variance ϑ^2) can be obtained, as in (18). Also, mean-field estimates are obtained replacing the γ weights by their expected values. All this results in an algorithm that parallels ALGORITHM 1:

ALGORITHM 2.

- *Initialize the parameters σ_k and p_k and the hidden states $\mathbf{X}$, using the mean field estimates.*
- *Iterate the following steps*
  1. *Compute the $\gamma_\ell^k(\mathbf{X}_{-k})$*
  2. *Re-estimate hidden states* (Maximization Step. Can be done by classification)
  3. *Re-estimate parameters σ_k and p_k* (Expectation Step).
- *Estimate the significant coefficients α_ℓ^k such that $X_\ell^k = s_k$ by regression.*

Before turning to a more complex situation, it is worth analyzing the results obtained so far and the corresponding algorithms. Propositions 1 and 2 actually provide expressions for adaptive thresholds, which are used inside iterative algorithms for hidden states estimations. Such an approach is to be compared with LASSO type variational approaches, in which regression is performed through iterative thresholding algorithms. The main difference with the current approach lies in the fact that here, the thresholds are coefficient dependent, and thresholding is performed on the *analysis* coefficients at each iteration: the goal of this approach is to obtain an estimate of the significance map while the LASSO gives an estimates of the *synthesis* coefficients.

2.2.2 Dependent coefficients

It turns out that a similar analysis can be carried out in situations where correlations between synthesis coefficients are introduced, provided they are introduced in a stratified way. We now develop a new scheme in which dependencies between coefficients are taken into account. We will see that in such a situation, generalized adaptive thresholding rules are obtained, together with an iterative thresholding algorithm, to be compared with the corresponding algorithm that was derived in a variational framework [8] (the so-called group-LASSO regression problem [18]).

For the sake of simplicity, let us first discuss the case of dictionaries structured as union of orthonormal bases $\mathcal{B}^k = \{\psi_\lambda^k, \lambda \in \Lambda_k\}$. The stratification is introduced by assuming that the index set of the basis is actually a double index: $\lambda = (g, m)$, g being a group index, and m being a membership index. We assume that the hidden states are structured in the sense that they are constant within a given group:

$$X_{gm}^k = X_g^k \ , \quad \forall m \ . \tag{23}$$

Furthermore, we assume that the synthesis coefficients belonging to the same group are (correlated) zero-mean Gaussian vectors. Assuming again a two-states model, we write

$$x = \sum_{k=1}^{K} \sum_g \sum_m \alpha_{g,m}^k \psi_{g,m}^k + r \ , \tag{24}$$

where for each k, the synthesis coefficients $\alpha_{g,m}^k$ are, conditional to the hidden state X_g, multivariate Gaussian random variables, with covariance matrix Σ^k.

$$\underline{\alpha}_g^k = \{\alpha_{g,m}^k, m = 1, 2, \dots\} \sim \begin{cases} \mathcal{N}(0, \Sigma^k) & \text{if } X_g^k = s_k \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

Again, the residual layer r is modelled as a Gaussian white noise, with zero mean and variance ϑ^2.

Suppose $\mathbf{X}_{-k} = \{x_{gm}^{k'}, k' \neq k\}$ is fixed. Conditional to $\mathbf{X}_{-k}$, the analysis coefficients $a_{g,m}^k = \langle x, \psi_{g,m}^k \rangle$ are distributed according to a multivariate zero mean normal distribution. Limiting our investigations to a fixed group g, we readily obtain

$$\mathbb{E}_0 \left\{ a_{gm}^k \overline{a}_{gm'}^k \right\} = I_g^k \Sigma_{mm'}^k + \left[\Gamma_g^k(\mathbf{X}_{-k}) \right]_{mm'} + \vartheta^2 \delta_{mm'} \ , \tag{26}$$

where we have introduced the indicator variable

$$I_g^k = \begin{cases} 1 & \text{if } X_g^k = s_k \\ 0 & \text{otherwise} \end{cases}$$

and the random covariance matrices $\Gamma_g^k(\mathbf{X}_{-k})$ defined by

$$\left[\Gamma_g^k(\mathbf{X}_{-k}) \right]_{mm'} = \sum_{k' \neq k} \sum_{g'} I_{g'}^{k'} \sum_{n,n'} \Sigma_{nn'}^{k'} \langle \psi_{g'n}^{k'}, \psi_{gm}^k \rangle \langle \psi_{gm'}^k, \psi_{g'n'}^{k'} \rangle \ . \tag{27}$$

The matrix $\Gamma_g^k(\mathbf{X}_{-k})$ represents the contribution of the other layers $k' \neq k$ to the covariance of the considered layer k. $\Gamma_g^k(\mathbf{X}_{-k})$ is positive semi-definite by construction. In analogy with the previous situation, set

$$W_{0;g}^k = \Gamma_g^k(\mathbf{X}_{-k}) + \vartheta^2 I \ , \qquad W_{1;g}^k = W_{0;g}^k + \Sigma_g^k \ . \tag{28}$$

In what follows, we shall need to use the inverses of these matrices.

LEMMA 3. *Assume that for all k, the covariance matrix Σ^k is (strictly) positive-definite. If $\vartheta \neq 0$, or if there exists $k' \neq k$ and g such that $X_g^{k'} = s_{k'}$, then $W_{0;g}^k$ and thus $W_{1;g}^k$ are positive-definite, and thus invertible.*

In this new situation, estimates for the hidden states cannot be obtained by simple coefficient thresholding as before. However, under suitable assumptions, such estimates can be obtained using generalized adaptive thresholding, very much in the spirit of the group-LASSO regression methods [18].

PROPOSITION 3. *Let k be a given layer, and assume $\mathbf{X}_{-k}$ is fixed. Assume that the matrices $W_{0;g}^k$ and $W_{1;g}^k$ are invertible, and set*

$$C_g^k = \left(W_{0;g}^k \right)^{-1} - \left(W_{1;g}^k \right)^{-1} \ . \tag{29}$$

Assume further that $W_{0;g}^k$ and $W_{1;g}^k$ are such that $w_0 = \det(W_{0;g}^k)$, $w_1 = \det(W_{1;g}^k)$ and p_k fulfill condition (16), and set

$$\tilde{\tau}^k(g) = \tau_0(\det(W_{0;g}^k), \det(W_{1;g}^k), p_k) \ .$$

1. *Conditional to $\mathbf{X}_{-k}$, the MAP estimate for the hidden states X_g^k is given as follows*

$$\hat{X}_g^k = \begin{cases} s_k & \text{if } \left| \langle \underline{a}_g^k, C_g^k \underline{a}_g^k \rangle \right| \geq \tilde{\tau}^k(g) \\ s_0 & \text{otherwise} \ . \end{cases}$$

2. *Assume C_g^k is positive definite. Then the estimate above becomes*

$$\hat{X}_g^k = \begin{cases} s_k & \text{if } \left\| \left(C_g^k \right)^{-1/2} \underline{a}_g^k \right\|^2 \geq \tilde{\tau}^k(g) \\ s_0 & \text{otherwise} \ . \end{cases}$$

REMARK 3. Notice the similarity of the so-obtained thresholding rule with the thresholding rule obtained in [8]: for a given layer k, a group g is selected (or not) based on the value of the L^2 norm of the corresponding vector of analysis coefficients: here this L^2 norm is replaced with a more general quadratic form, which reduces to a squared norm in the positive definite case.

In very much the same spirit as before, parameters (except again the noise variance) can be estimated conditional to the hidden states. This results in the following general scheme

ALGORITHM 3.

- *Initialize the parameters* Σ^k *and* p_k *and the hidden states* $\mathbf{X}$*, using the mean field estimates.*
- *Iterate the following steps*
  1. *Compute the matrices* $\Gamma_g^k(\mathbf{X}_{-k})$
  2. *Re-estimate the hidden states* $X_{g;m}^k$ (Maximization Step. Can be done by classification)
  3. *Re-estimate parameters* Σ^k *and* p_k (Expectation Step).
- *Estimate the significant coefficients* $\underline{\alpha}_g^k$ *such that* $X_g^k = s_k$ *by regression.*

Notice that in this algorithm (as in the previous one) we do not specify the regression method used to estimate the significant coefficients. Standard L^2 regression can be used, as well as sparse regression.

3. NUMERICAL RESULTS

We now illustrate the approaches we described above in sound signal applications. The goal is to obtain multilayered audio signal expansions that can synthesize separately transient and tonal layers. Following earlier works, we consider a dictionary constructed as the union of two time-frequency bases (i.e. MDCT bases, see [15]). Hence, we follow here the model M2 with $K = 2$.

We recall that an MDCT basis is an orthonormal basis of functions of the form

$$\psi_{\tau\nu}(t) = \sqrt{\frac{2}{L}}\, \omega(t - k\ell) \cos\left(\pi \frac{\nu + 1/2}{L}(t - \tau L)\right) , \tag{30}$$

for an appropriate choice of the window function ω and the parameter L (see [15] for details). The choice of the window determines the time-frequency resolution of the window. We thus choose two windows ω^1 and ω^2 of different sizes, for describing the tonal layer (wide window, narrow band) and the transient layer (wide window, narrow band) respectively.

For the correlated case, we need to construct a stratification of the index set.

- For the tonal layer, synthesis coefficients are expected to be correlated in time. The group index is then the frequency index ν, and the membership index is the time index k.
- For the transient layer, the situation is opposite (correlations across frequencies for fixed time). The group index is then the time index ν, and the membership index is the frequency index k.

The musical signal used for the experiments is an excerpt of *mamavatu* from Shusheela Raman. The signal is sampled at 44.1 KHz and is about 12 s long (2^{19} samples). The two window functions we choose are respectively 256 samples (6 ms) and 4096 samples (93 ms) long for the transient and the tonal layers.

The results shown here are essentially obtained using the mean-field version of the analysis, which has the advantage of being numerically efficient. The CEM algorithms proposed above are actually very costly in terms of computations, because the γ weights have to be updated at each iteration. Efficient update for the γ weights require the storage of the dictionary's Gram matrix, which is very demanding in terms of storage capacities for long signals. However, the structure of the MDCT basis makes it possible to simplify it, as discussed in [12].

A mean-field version can be also derived within a CEM algorithm. We provide the algorithm for the correlated case, assuming the groups are distributed according to a Bernoulli law. Notice that in this simple version, both layers are estimated independently of each other, which is far from optimal. The algorithm is described in the case of the transient layer.

ALGORITHM 4.

1. *Fix a basis* $\{\psi_{\tau\nu}\}$ *and compute the analysis coefficients* $a_{\tau\nu} = \langle x, \psi_{\tau\nu} \rangle$.
2. *Initialize the significance map (for example, with a thresholding on the* ℓ^2 *norm of the vectors* $\underline{a}_\tau = \{a_{\tau\nu}, \nu = 1, 2, \dots\}$*).*
3. *Classification of the analysis coefficients with CEM*
   - *Estimation of the memberships probabilities (denoted by* p_1 *and* p_2*) by computing respectively the proportion of* $\hat{X}_\tau = 1$ *and* $\hat{X}_\tau = 0$.
   - *Estimation of the covariance matrices of* $\{a_{\tau\nu}, X_\tau = 1\}$ *and* $\{a_{\tau\nu}, X_\tau = 0\}$ *(denoted by* Σ_1 *and* Σ_2*)*

   (a) E step: *compute the membership probabilities of* $a_{\tau\nu}$ *for each weighted normal law.*

   (b) M step by classification: $\hat{X}_k = 1$ *if* $\mathbb{P}\{a_{\tau\nu} \sim p_1 \mathcal{N}(0, \Sigma_1)\} > \mathbb{P}\{a_{\tau\nu} \sim p_2 \mathcal{N}(0, \Sigma_2)\}$*, and then estimates the membership probabilities and the covariance matrices.*

We first use the Bernoulli model described in [12] for both the tonal and the transient layers. In this model, each time-frequency atom of each basis $k = 1, 2$ can be used with probability p_k. The mean field estimates of the significance maps of each layer are provided in figure 1.

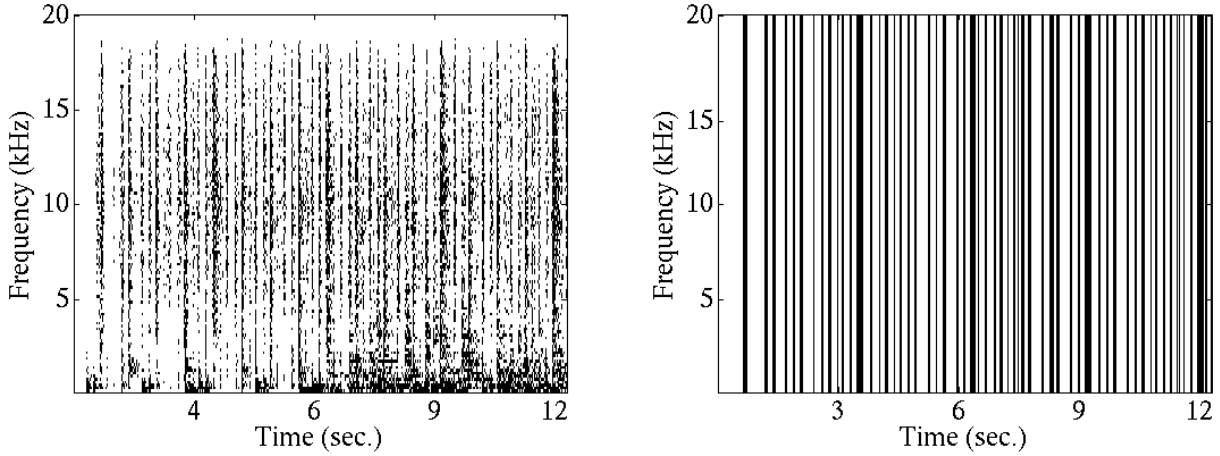

Figure 1. Decomposition of the mamavatu signal into two layers. Left: significance map of the transient layer using the Bernoulli M2 model. Right: significance map of the transient layer using the structured model.

As can be seen in Figure 1, the mean field Bernoulli approach has difficulties separating the two layers correctly. The estimated tonal layer (not shown here) keeps a lot of transient information, and the estimated transient layer retains too many low frequency atoms that "belong" to the tonal layer. The two layers being estimated independently of each other, they share too much information: the tonal layer "sounds" too "transient", and the transient layer "sounds" too "tonal". The right hand side image represents the estimated significance map obtained using the "vector" model. As expected, the significance map is less sparse (entire vectors are retained), but sharper. In addition, the low frequency part is not present any more.

As a result, the vector algorithm also provides estimates for the correlation (i.e. covariance divided out by standard deviations) matrices for the two layers. Focusing again on the transient layer estimate, we display in Figure 2 the correlation matrix for the transient layer and the non transient layer (using a narrow analysis window). As can be seen, the dominance of the diagonal is much stronger for the non transient part, which reveals the frequency correlations that have been captured by the algorithm. One may also notice that the rows of the transient correlation matrix are much sparwer than the rows of the other one. This remark may be used for initializing the classification algorithms: mark as "transient" the fixed time coefficient vectors $\{a_{k\nu}, \nu = 1, 2, \dots\}$ whose ℓ^1 norm (or any other diversity measure) is below a certain threshold.

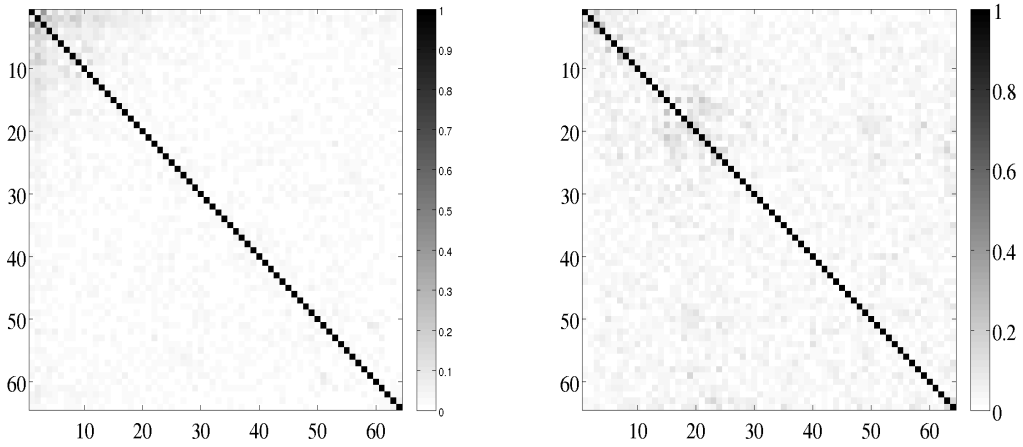

Figure 2. Correlation matrices of the fixed-time analysis coefficient vectors, estimated using the CEM algorithm: non-transient (left) and transient layers (right).

4. CONCLUSIONS

We have described in this paper some generalizations of results previously obtained in [12], motivated by the need of modelling more closely sound signals. The proposed approach fits nicely into EM or CEM estimation algorithms. The theoretical results are fairly similar to the ones obtained in [12], and clearly show the role of the dictionary and sparsity assumptions on the possibility of recovering sparse signal expansions from analysis coefficients. In the Bernoulli case, we have for simplicity limited the discussion to the case where all coefficients belonging to the same state have the same variance, but this assumption can be relaxed as was done in [12].

Besides the generalization to arbitrary dictionaries, a main aspect was to extend the previous strategy, which was essentially based upon synthesis coefficient decorrelation assumptions, to the correlated case. Interestingly enough, the obtained algorithms bear some resemblance with iterative thresholding approaches developed to solve the LASSO regression problem (decorrelated case), and the generalized version that was developed in the context of group-LASSO regression (correlated case): a group of coefficients is considered significant and thus selected if its image by a given quadratic form (the squared L^2 norm for group-LASSO) exceeds a given threshold.

While the proposed procedure is particularly relevant for transient modeling, it should be improved further for modelling tonals. Indeed, when it comes to long signals, it doesn't make sense time-independent tonal significance maps, and some time-variations have to be included. This can be made by introducing an extra time scale in the algorithm, as in [12]. The latter may also be used to update the parameters of the model (including the parameters of the transient layer) as a function of time.

REFERENCES

[1] Daudet, L. and Torrésani, B., "Sparse adaptive representations for musical signals," in [*Signal Processing Methods for Music Transcription*], Klapuri, A. and Davy, M., eds., 65–98, Springer, New York (2006).

[2] Goodwin, M. M., [*Adaptive signal models : theory, algorithms and audio applications*], vol. 467 of *International Series on Engineering and Computer Sciences*, Kluwer (1998).

[3] Boyer, R. and Abed-Meraim, K., "Audio modeling based on delayed sinusoids," *IEEE Transactions on Speech and Audio Processing* **12**(2), 110–120 (2004).

[4] Vincent, E., *Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux*, PhD thesis, Université de Paris VI, France (2004).

[5] Fornasier, M. and Rauhut, H., "Recovery algorithm for vector-valued data with joint sparsity constraints," *SIAM Journal on Numerical Analysis* **46**(2), 577–613 (2007).

[6] Tropp, J., Gilbert, A., and Strauss, M., "Algorithms for simultaneous sparse approximation. part II: Convex relaxation," *Signal Processing* **86**, 589–602 (April 2006). special issue "Sparse approximations in signal and image processing".

[7] Teschke, G. and Ramlau, R., "An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector valued regimes and an application to color image inpainting," *Inverse Problems* **23**, 1851–1870 (October 2007).

[8] Kowalski, M., "Sparse regression using mixed norms," *Applied and Computational Harmonic Analysis* (2009). In press.

[9] Cotter, S., Rao, B., Engan, K., and Kreutz-Delgado, K., "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing* **53**(7), 2477–2488 (2005).

[10] Tropp, J., Gilbert, A., and Strauss, M., "Algorithms for simultaneous sparse approximation. part I: Greedy pursuit," *Signal Processing* **86**, 572–588 (April 2006). special issue "Sparse approximations in signal and image processing".

[11] Gribonval, R., Rauhut, H., Schnass, K., and Vandergheynst, P., "Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms," *Journal of Fourier Analysis and Applications* **14**(5-6), 655–687 (2008).

[12] Kowalski, M. and Torrésani, B., "Random models for sparse signals expansion on unions of bases with application to audio signals," *IEEE Transactions on Signal Processing* **58**(8), 3468–3481 (2008).

[13] Molla, S. and Torrésani, B., "An hybrid audio scheme using hidden Markov models of waveforms," *Applied and Computational Harmonic Analysis* **18**(2), 137–166 (2005).

[14] Févotte, C., Torrésani, B., Daudet, L., and Godsill, S. J., "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Transactions on Audio Speech and Language Processing* **16**(1), 174–185 (2008).

[15] Vetterli, M. and Kovacević, J., [*Wavelets and Subband Coding*], Signal Processing Series, Prentice Hall, Englewood Cliffs, NJ (1995).

[16] Daudet, L. and Torrésani, B., "Hybrid representations for audiophonic signal encoding," *Signal Processing* **82**(11), 1595–1617 (2002). Special issue on Image and Video Coding Beyond Standards.

[17] Govaert, G. and Celeux, G., "A classification em algorithm for clustering and two stochastic versions," *Computational Statistics and Data Analysis* **14**(3), 315–332 (1992).

[18] Yuan, M. and Lin, Y., "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Serie B* **68**(1), 49–67 (2006).