SPARSE REGRESSION WITH STRUCTURED PRIORS: APPLICATION TO AUDIO DENOISING

Cédric Févotte,[1] Laurent Daudet,[2] Simon J. Godsill [1] and Bruno Torrésani [3]

[1]Engineering Department,
University of Cambridge, UK
{cf269, sjg}@eng.cam.ac.uk

[2]LAM,
Univ. P&M Curie, France
daudet@lam.jussieu.fr

[3]LATP,
Université de Provence, France
Bruno.Torresani@cmi.univ-mrs.fr

ABSTRACT

We describe in this paper a fully Bayesian approach for sparse audio signal regression in an union of two bases, with application to audio denoising. One basis aims at modeling tonal parts and the other at modeling transients. The noisy signal is decomposed as a linear combination of atoms from the two basis, plus a residual part containing the noise. Conditionally upon an indicator variable which is either 0 or 1, one source coefficient is set to zero or given a hierarchical prior. Various priors can be considered for the indicator variables. In addition to non-structured Bernoulli priors we study the performance of structured priors which favor horizontal time-frequency structures for tonals and vertical structures for transients. A Gibbs sampler is used to sample from the parameters of the model. We present results over denoising of a piano sequence using a MDCT basis with long time resolution to model the tonals and a MDCT with short time resolution to model the transients.

1. INTRODUCTION

Sparse representations of signals are useful in a variety of current signal processing challenges, particularly in the audio domain : signal denoising, blind source separation, audio analysis and compression, for example. By sparse, we here mean that most of the energy of the signal $x = \sum s_i \Phi_i$ is concentrated in a few significant coefficients, allowing one to truncate the expansion to only these coefficients $x \approx \sum_{j=1}^{J} s_{i_j} \Phi_{i_j}$, while still maintaining to good approximation of the signal.

For music signals, lapped transforms such as the Modified Discrete Cosine Transform (MDCT) have proved their effectiveness and are nowadays used in the vast majority of high-quality audio coding algorithms (see *e.g* [1]). The MDCT has a number of desirable properties : it is orthogonal, is does not lead to blocking effects, and its implementation is based on the FFT. However, in general, a straightforward use of such orthonormal bases does not provide sufficiently sparse approximations, as for musical signals they fail to represent sparsely the combination of both tonal components (which require long windows) and fast transients (e.g. the attacks of the notes, which require small windows). To get around this problem, the simplest solution is to adapt the window size according to the local content of the signal. However, this usually requires *ad-hoc* methods for the adaptation decision; and more importantly such a binary decision does not account for the additive nature of audio signals: it is indeed quite common to have tones and transients simultaneously present at a given time.

To have a truly additive model, overcompleteness is required. We look for decompositions of the signal as a linear combination of elementary waveforms ("atoms") chosen from within a large dictionary Φ. However, given an overcomplete dictionary Φ, the uniqueness of the decomposition does not hold anymore. Consequently, finding the optimal or nearly-optimal J-term approximation of x requires a tradeoff between precision and computational complexity (indeed, finding *the* optimal J-term approximation is known to be an NP-complete problem). Many methods have been proposed, such as basis pursuit denoising (BPDN) [2] and matching pursuit [3]. In this context, a typical choice of dictionary is the union of two orthonormal bases, $\{\Phi_{1,k}\}_{k=1\ldots N}$ for the tonal content and $\{\Phi_{2,k}\}_{k=1\ldots N}$ for the transients:

$$x = \sum_{k=1}^{N} \tilde{s}_{1,k}\,\Phi_{1,k} + \sum_{k=1}^{N} \tilde{s}_{2,k}\,\Phi_{2,k} + e. \qquad (1)$$

As such, these techniques can give satisfactory results [4,5], but the denoised signals are often contaminated by a significant amount of so-called "musical noise". This is due to the fact that some components can be represented almost equally well (or badly !) in either basis Φ_1 or Φ_2. When the priors do not explicitly consider local dependencies between atom coefficients, isolated components can remain which may be perceived as artifacts. Indeed, one has to enforce some consistency using *structured* priors. For the tonal content, each partial appears on the spectrogram as a horizontal line. Transients, on the other hand, are characterized by vertical lines: most time-frequency bins are active for some short time duration. Therefore, enforcing structure should help reducing the number of isolated atoms, and thus the musical noise. In a Markovian framework, such structures have been proposed [6] for representing tones and transients, respectively in the MDCT and discrete wavelet domains. The approach is two-step: first identify tonal components from the analysis coefficients of the signal on the long window MDCT and then identify transients in the residual.

Our approach is one-step: we aim at identifying the synthesis coefficients $\{\tilde{s}_{i,k}\}$ for all bases jointly, through application of a Bayesian hierarchical prior. An indicator variable is used to indicate whether a particular coefficient should be set to zero or not. Markov chain priors are used on the indicator variables to favor horizontal structures for tonals and vertical structures for the transients. A Gibbs sampler (a standard Markov chain Monte Carlo technique) is used to jointly estimate all of the parameters in the model. Our approach can be related to that of [7], in which horizontal and spatial Markov priors were adopted in a single-resolution overcomplete Gabor regression model. In our case the chosen overcomplete basis, in the framework of Eq. (1), is a dual-resolution union of two MDCT

bases, one with long windows $\{\Phi_{1,k}\}_{k=1...N}$ for the tonal content and one with short windows $\{\Phi_{2,k}\}_{k=1...N}$ for the transients. Although this technique can be applied to any of the above-mentioned applications, we focus here on denoising problems, which are more straightforward to evaluate (at least in terms of SNR).

The paper is organized as follows. Section 2 describes the signal model and priors used for all the parameters. Section 3 describes briefly the Gibbs sampler and gives update steps for the parameters. Section 4 gives results for denoising of a piano sequence, comparing different structured priors for the indicator variables. We show that the use of horizontal and vertical structures to model the tonals and the transients clearly reduces musical noise.

2. SIGNAL MODEL

2.1. Coefficients priors

The coefficients $\tilde{s}_{i,k}$, $i = 1, 2$, $k = 1, \ldots, N$ of Eq. (1) are given the following hierarchical prior:

$$
\begin{aligned}
p(\tilde{s}_{i,k}|\gamma_{i,k}, \alpha_i, \lambda_i) &= (1 - \gamma_{i,k})\,\delta_0(\tilde{s}_{i,k}) + \gamma_{i,k}\,\mathcal{N}(\tilde{s}_{i,k}|0, v_{i,k}) \\
p(v_{i,k}|\alpha_i, \lambda_i) &= \mathcal{IG}(v_{i,k}|\alpha_i, \lambda_i)
\end{aligned}
\tag{2}
$$

where $\mathcal{N}(u|\mu, v)$ and $\mathcal{IG}(u|\alpha, \beta)$ are the normal and inverted-Gamma distributions as defined in Appendix, $\delta_0(x)$ is the Dirac delta function and $\gamma_{i,k} \in \{0, 1\}$ is an indicator variable. When $\gamma_{i,k} = 0$, $\tilde{s}_{i,k}$ is set to zero; when $\gamma_{i,k} = 1$, $\tilde{s}_{i,k}$ has a normal distribution with zero mean and variance $v_{i,k}$, which is in turn assigned a conjugate inverted-Gamma prior.

2.2. Structured priors

In order to model persistencies of t-f coefficients of musical signal, structured priors are used for $\gamma_{i,k}$. We consider three scenarios:

1. *Bernoulli priors*: no structure is imposed on the indicator variables, which are assigned the following independent Bernoulli priors:

$$
P(\gamma_{i,k} = 1|P_i) = P_i \quad P(\gamma_{i,k} = 0|P_i) = 1 - P_i \tag{3}
$$

2. *"Horizontal" Markov model for tonals and Bernoulli prior on transients*: in order to model persistency in time of t-f coefficients corresponding to tonal parts, we give a prior horizontal structure to the indicator variable of the first basis (the one aimed at modeling tonals). More precisely, when a MDCT basis is used, the index $k = 1, \ldots, N$ is more conveniently replaced by (q, n) with $q = 1, \ldots, l_{\text{frame1}}$ being a frequency index and $n = 1, \ldots, n_{\text{frame1}}$ being a frame index, with $l_{\text{frame1}} \times n_{\text{frame1}} = N$. For a fixed frequency index q the sequence $\{\gamma_{1,q,n}\}_{n=1,\ldots,n_{\text{frame1}}}$ is modeled by a 2-state first order Markov chain with transition probabilities $P_{1,0\to0}$ and $P_{1,1\to1}$.

3. *"Horizontal" Markov model for tonals and "vertical" Markov model for transients*: in addition to horizontal structures we also favor vertical structures for the transients. For a fixed frame index n, the sequence $\{\gamma_{2,q,n}\}_{q=1,\ldots,l_{\text{frame2}}}$ is thus modeled by a 2-state first order Markov chain with transition probabilities $P_{2,0\to0}$ and $P_{2,1\to1}$.

2.3. Residual model

The residual signal e is assumed i.i.d zero-mean Gaussian with variance σ^2 given an inverted-Gamma (conjugate) prior $p(\sigma^2|\alpha_\sigma, \beta_\sigma) = \mathcal{IG}(\sigma^2|\alpha_\sigma, \beta_\sigma)$.

2.4. Hyperparameters priors

The scale parameters λ_i in each basis are given independent Gamma (conjugate) priors $p(\lambda_i|\alpha_{\lambda_i}, \beta_{\lambda_i}) = \mathcal{G}(\lambda_i|\alpha_{\lambda_i}, \beta_{\lambda_i})$, allowing an automatic adaptation to the scaling of the coefficients in each basis. The degrees of freedom α_i can be fixed to a certain value or estimated like in [8]. However in our simulations the value of α_i happened to have little influence on the results and in practice we fixed it to 1. The probabilities P_i in the Bernoulli models and $P_{i,0\to0}$, $P_{i,1\to1}$ in the Markov models are given uniform priors on $[0\ 1]$, which may be routinely extended to Beta priors if required to favor certain values over others.

3. MCMC INFERENCE

We propose to sample from the posterior distribution of the parameters $\boldsymbol{\theta} = \{\tilde{s}_i, v_i, \alpha_i, \lambda_i\}_{i=1,2} \cup \sigma^2$, using a Gibbs sampler. The Gibbs sampler is a standard Markov Chain Monte Carlo technique which simply requires to sample from the conditional distributions of each parameter upon the others [9]. Point estimates can then be computed from the obtained samples of the posterior distribution $p(\boldsymbol{\theta}|x)$. In contrast with EM-like methods which aim directly at point estimates (ML or MAP), MCMC approaches are very robust because they scan the full posterior distribution and are thus unlikely to fall into local minima. This is however at the cost of higher computational burden.

3.1. Alternate sampling of $(\gamma_1, \tilde{s}_1)$ and $(\gamma_2, \tilde{s}_2)$

One approach is to sample $\gamma = [\gamma_1\ \gamma_2]$ and $\tilde{s} = [\tilde{s}_1\ \tilde{s}_2]$ successively. This strategy requires the storage and inversion of the $2N \times 2N$ matrix $(\Phi^T\,\Phi + \sigma^2 \operatorname{diag}([v_1\ v_2])^{-1})$ at each iteration of the sampler (where $\Phi = [\Phi_1\ \Phi_2]$), which might not be feasible for large N. The structure of our dictionary Φ, however, facilitates alternative block sampling of $(\gamma_1, \tilde{s}_1)$ and $(\gamma_2, \tilde{s}_2)$, in the fashion of [10]. Indeed, because the Euclidean norm is invariant under rotation, the likelihood of the observation x can be written as

$$
\begin{aligned}
p(x|\boldsymbol{\theta}) &= (2\,\pi\,\sigma^2)^{-N/2} \exp -\frac{1}{2\,\sigma^2} \|x - \Phi_1\,\tilde{s}_1 - \Phi_2\,\tilde{s}_2\|_2^2 \\
&= (2\,\pi\,\sigma^2)^{-N/2} \exp -\frac{1}{2\,\sigma^2} \|\underbrace{\Phi_2^T\,(x - \Phi_1\,\tilde{s}_1)}_{\tilde{x}_{2|1}} - \tilde{s}_2\|_2^2 \\
&= (2\,\pi\,\sigma^2)^{-N/2} \exp -\frac{1}{2\,\sigma^2} \|\underbrace{\Phi_1^T\,(x - \Phi_2\,\tilde{s}_2)}_{\tilde{x}_{1|2}} - \tilde{s}_1\|_2^2
\end{aligned}
$$

This means that conditionally upon $(\gamma_2, \tilde{s}_2)$ (resp. $(\gamma_1, \tilde{s}_1)$) and the other parameters, inferring $(\gamma_1, \tilde{s}_1)$ (resp. $(\gamma_2, \tilde{s}_2)$) is a simple regression problem with i.i.d data $\tilde{x}_{1|2}$ (resp. $\tilde{x}_{2|1}$), and does not require any matrix inversion.

3.2. Gibbs steps

We now give the expression for the update steps of the parameters. In the following most of the derivations have been skipped, further details can be found in [5]. Note that all the conditional posterior distributions of all the parameters can be easily sampled from.

3.2.1. Update of $(\gamma_i, \tilde{s}_i)$

The coefficients $\tilde{s}_{i,k}$ can be integrated out from the posterior distribution of $\gamma_{i,k}$, leading to

$$p(\gamma_{i,k} = 0 | \sigma^2, v_{i,k}, \tilde{x}_{i|-i,k}) \quad = \quad 1/(1 + \tau_{i,k}) \qquad (4)$$

$$p(\gamma_{i,k} = 1 | \sigma^2, v_{i,k}, \tilde{x}_{i|-i,k}) \quad = \quad \tau_{i,k}/(1 + \tau_{i,k}) \qquad (5)$$

with

$$\tau_{i,k} = \sqrt{\frac{\sigma^2}{\sigma^2 + v_{i,k}}} \, \exp\left(\frac{\tilde{x}_{i|-i,k}^2 \, v_{i,k}}{2 \, \sigma^2 (\sigma^2 + v_{i,k})} \right) \frac{p(\gamma_{i,k} = 1 | \gamma_{i,-k})}{p(\gamma_{i,k} = 0 | \gamma_{i,-k})}$$
$$(6)$$

where $\gamma_{i,-k}$ denotes the set of all indicator variables $\{\gamma_{i,l}\}_{l=1,\dots,N}$ except $\gamma_{i,k}$ and where $\tilde{x}_{i|-i}$ similarly denotes either $\tilde{x}_{1|2}$ or $\tilde{x}_{2|1}$. The expression of the ratio $p(\gamma_{i,k} = 1 | \gamma_{i,-k})/p(\gamma_{i,k} = 0 | \gamma_{i,-k})$ changes according to the chosen prior for the indicator variables. When $\gamma_{i,k}$ has a Bernoulli prior, this ratio is simply $P_i/(1 - P_i)$. When $\gamma_{1,k}$ has a Markov horizontal structure and $k = (q, n)$, this ratio depends on the values of $\gamma_{1,q,n-1}$ and $\gamma_{1,q,n+1}$. The exact expressions are standard results from the Markov chain literature (see *e.g*, [11]). The posterior distribution of $\tilde{s}_{i,k}$ is written as

$$p(\tilde{s}_{i,k} | \gamma_{i,k}, v_{i,k}, \sigma^2, \tilde{x}_{i|-i,k}) = \begin{cases} \delta_0(\tilde{s}_{i,k}) & \text{if } \gamma_{i,k} = 0 \\ \mathcal{N}(\tilde{s}_{i,k} | \mu_{\tilde{s}_{i,k}}, \sigma^2_{\tilde{s}_{i,k}}) & \text{if } \gamma_{i,k} = 1 \end{cases}$$
$$(7)$$

with $\sigma^2_{\tilde{s}_{i,k}} = \left(1/\sigma^2 + 1/v_{i,k}\right)^{-1}$ and $\mu_{\tilde{s}_{i,k}} = (\sigma^2_{\tilde{s}_{i,k}}/\sigma^2) \, \tilde{x}_{i|-i,k}$.

3.2.2. Update of v_i

The conditional posterior distribution of $v_{i,k}$ is

$$p(v_{i,k} | \gamma_{i,k}, \tilde{s}_{i,k}, \alpha_i, \lambda_i) =$$
$$\begin{cases} \mathcal{IG}\left(v_{i,k} | \alpha_i, \lambda_i\right) & \text{if } \gamma_{i,k} = 0 \\ \mathcal{IG}\left(v_{i,k} | \frac{1}{2} + \alpha_i, \frac{\tilde{s}_{i,k}^2}{2} + \lambda_i\right) & \text{if } \gamma_{i,k} = 1 \end{cases} \qquad (8)$$

3.2.3. Update of σ^2

The conditional posterior distribution of σ^2 is given by

$$p(\sigma^2 | \tilde{s}_1, \tilde{s}_2, x) =$$
$$\mathcal{IG}\left(\sigma^2 | \frac{N}{2} + \alpha_\sigma, \frac{\|x - \Phi_1 \tilde{s}_1 - \Phi_2 \tilde{s}_2\|_2^2}{2} + \beta_\sigma \right) \qquad (9)$$

3.2.4. Update of the hyperparameters

- The posterior distribution of the scale parameters is $p(\lambda_i | v_i) = \mathcal{G}\left(\lambda_i | N \alpha_i + \alpha_{\lambda_i}, \sum_k 1/v_{i,k} + \beta_{\lambda_i}\right)$,

- When the indicator variables are given a Bernoulli prior, the posterior distribution of P_i is simply $p(P_i) = \mathcal{B}(P_i | \#\gamma_i + 1, N - \#\gamma_i + 1)$, where $\mathcal{B}(x | \alpha, \beta)$ is the Beta distribution defined in the Appendix and $\#\gamma_i$ is the number of values of $\gamma_{i,k}$ equal to 1. Similarly, when the indicator variables are given Markov priors, the posterior distributions of the transition probabilities can be sampled using a Metropolis-Hasting step as in [7]. In this work we simply update them as the number of transitions from 0 to 0 and 1 to 1 divided by N.

1. Bernoulli tonals + Bernoulli transients	20.7 dB
2. MC tonals + Bernoulli transients	21.6 dB
3. MC tonals + MC transients	21.6 dB
Jeffrey's + EM	15.3 dB

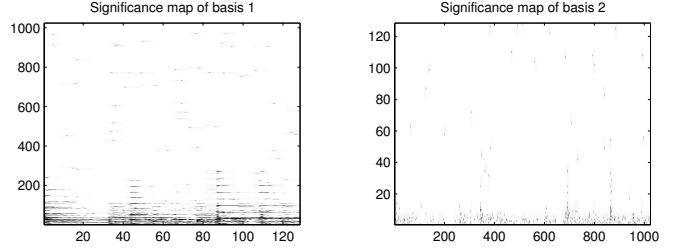**Table 1**. Output SNR of the signal estimate for various input SNR.

Fig. 1. Significance maps of the selected atoms in each basis using Bernoulli priors on γ_1 and γ_2.

4. RESULTS

We present results for denoising of a melodic piano sequence of length $N = 131072$, sampled at 44.1kHz ($\approx 3s$). White Gaussian noise was added to the clean source with 10dB input SNR. We used MDCT bases with sine bell analysis window and with 50% overlap. The time resolutions (half the window length) are $l_{frame1} = 1024$ samples ($\approx 23.2ms$) and $l_{frame2} = 128$ samples ($\approx 2.9ms$).

We ran 500 iterations of our Gibbs sampler, and Minimum Mean Square Error estimates of the parameters where computed by averaging the last 100 samples. Using a MATLAB implementation running on a 1.25 GHz Powerbook G4 with 1Go RAM this requires approximatively 30min when using Bernoulli priors only for γ_i, and a few minutes more when using Markov chains priors on each basis. The MDCT operation $\tilde{y} = \Phi_i^T y$ and inverse MDCT operation $y = \Phi_i \tilde{y}$ were performed using an implementation based on the Wavelab MATLAB toolbox [12]. The horizontal Markov chains were initialized with arbitrary probabilities $P(\gamma_{1,q,1} = 1) = 0.1$ and the vertical Markov chains with $P(\gamma_{2,1,n} = 1) = 0.05$.

Our approaches are compared with the one in [10], in which a Jeffrey's inverse prior $p(\tilde{s}_{i,k}) \propto 1/|\tilde{s}_{i,k}|$ is employed. This heavy-tailed (improper) prior leads to very sparse signal representations. It admits a hierarchical Gaussian formulation with Jeffrey's inverse prior on the variances. An Expectation Conditional Maximization algorithm is used to find a Maximum A Posteriori estimate of $\{\tilde{s}_1, \tilde{s}_2, \sigma^2\}$, treating $\{v_1 \, v_2\}$ as the missing data and using alternate update of $\tilde{s}_1$ and $\tilde{s}_2$ like in Section 3.1.

Table 1 shows the output SNR $20 \log_{10} \|\hat{s} - s\| / \|s\|$ of the signal estimate $\hat{s}$ obtained with each method. Audio files can also be found at `http://www-sigproc.eng.cam.ac.uk/~cf269/icassp06/sound_files.html`.

Fig. 1, 2 and 3 show the significance maps of each basis, *i.e* the Maximum A Posteriori estimates of γ_1 and γ_2 for the three scenarios described in Section 2.2.

The audio examples and Table 1 show that our approach, consisting of using hierarchical model for the source coefficients and a MCMC approach, outperforms the Jeffrey's + EM approach, even when using Bernoulli priors only. The main reason for this is that
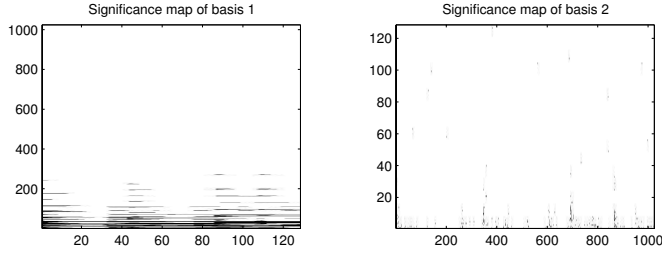
Fig. 2. Significance maps of the selected atoms in each basis using horizontal Markov prior on γ_1 and Bernoulli prior on γ_2.
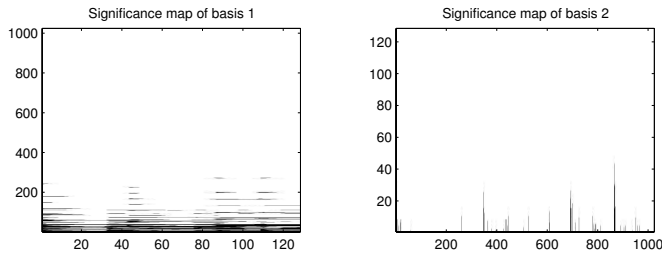

Fig. 3. Significance maps of the selected atoms in each basis using horizontal Markov prior on γ_1 and vertical prior on γ_2.

our prior is more flexible than the Jeffrey's prior. However, the Jeffrey's + EM approach has the advantage of lower computational cost (the results were obtained in 2 min).

5. CONCLUSIONS

We have proposed a fully Bayesian approach for sparse signal regression in union of basis, with application to audio denoising. The use of a mixture prior of a Dirac centered at 0 and a hierarchical prior for the source coefficients involves an indicator variable which can be given desired structures to model tonals and transients of audio signals. Table 1 does not reveal large differences of output SNR between scenarios 1, 2 and 3. However the audio files show that musical noise is considerably reduced when structures are used. In particular, though the SNR values obtained with scenario 2 and 3 are equal, the audio files show that the transient part obtained with scenario 3 contains much less musical noise than with scenario 2. The transients having a rather small energy, their contribution to the output SNR is low, hence the importance of audio qualitative assessment.

Other experiments carried out on polyphonic recordings showed that the transients estimate obtained from scenario 3 reliably follows the rhythm of the musical extracts and could be used for other applications such as tempo extraction or audio segmentation. Perspective of this work will involve studying the influence of the initialization of the Markov chains as well as exploring other transients models, such as hidden Markov trees priors used with wavelet bases, in a similar fashion to [6].

A. STANDARD DISTRIBUTIONS

Normal $\qquad \mathcal{N}(x|u, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp{-\frac{(x-u)^2}{2\sigma^2}}$

Beta $\qquad \mathcal{B}(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad x \in [0, 1]$

Gamma $\qquad \mathcal{G}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x), \ x \in [0, +\infty)$

inv-Gamma $\ \mathcal{IG}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp(-\frac{\beta}{x}), \ x \in [0, +\infty)$

The inverted-Gamma distribution is the distribution of $1/X$ when X is Gamma distributed.

B. REFERENCES

[1] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1998.

[2] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[3] S. Mallat and S. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, Dec 1993.

[4] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, Nov. 2002.

[5] C. Févotte and S. J. Godsill, "Sparse linear regression in unions of bases via Bayesian variable selection," *IEEE Signal Processing Letters*, 2005, Accepted for publication - Preprint available at http://www-sigproc.eng.cam.ac.uk/~cf269/.

[6] S. Molla and B. Torrésani, "An hybrid audio scheme using hidden Markov models of waveforms," *Applied and Computational Harmonic Analysis*, vol. 18, no. 2, pp. 137–166, Mar 2005.

[7] P. J. Wolfe, S. J. Godsill, and W.-J. Ng, "Bayesian variable selection and regularisation for time-frequency surface estimation," *J. R. Statist. Soc. Series B*, 2004.

[8] C. Févotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Trans. Speech and Audio Processing*, In press - Preprint available at http://www-sigproc.eng.cam.ac.uk/~cf269/.

[9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PAMI-6, no. 6, pp. 721–741, Nov 1984.

[10] M. E. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Processing*, In press.

[11] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, , no. 77, pp. 257–286, 1989.

[12] D. Donoho, M. R. Duncan, X. Huo, and O. Levi, "Wavelab 802," Toolbox for Matlab, http://www-stat.stanford.edu/~wavelab/.