BAYESIAN APPROACH TO BEST BASIS SELECTION

J.-C. Pesquet, H. Krim‡, D. Leporini and E. Hamman

Laboratoire des Signaux et Systèmes,
CNRS/UPS and GDR ISIS, 91192 Gif sur Yvette Cédex, France,
Email: pesquet@lss.supelec.fr,
‡Stochastic Systems Group,
LIDS, MIT, Cambridge, MA 02139, USA

ABSTRACT

Wavelet packets and local trigonometric bases provide an efficient framework and fast algorithms to obtain a "best basis" or "best representation" of deterministic signals. Applying these deterministic techniques to stochastic processes may, however, lead to variable results. In this paper we revisit this problem and introduce a prior model on the underlying signal in noise and account for the contaminating noise model as well. We thus develop a Bayesian-based approach to the best basis problem, while preserving the classical tree search efficiency.

1. INTRODUCTION

Research interest in a "best representation" of a signal $x(t)$ when given a finite dictionary of possible bases, has greatly increased recently [1, 2, 3, 4]. In [1], it was established that wavelet packets and Malvar's wavelets (or local cosine bases) led to interesting choices of time-frequency dictionaries based on binary tree structures. These structures, in turn, allowed for a very efficient and fast dynamic programming methodology in searching for a "Best Basis" (BB). The relevance of these techniques in many real-world applications together with their highly variable performance in the presence of noise were the primary impetus in investigating their properties in a stochastic setting.

A few statistical approaches to the BB problem have recently appeared [2, 3, 4]. Most adopt the following model for the observed process,

$$y(t) = x(t) + w(t),$$

where $w(t)$ is Gaussian, with zero mean and finite variance, and for which estimation/recovery of the underlying *unknown* signal $x(t)$ is of interest. In [5, 6], thresholding denoising techniques have been introduced using a wavelet representation. More recently, other works [2, 3, 4] use the thresholding procedure as part of the BB search. It was shown in [7, 3] that the thresholding strategy was linked to the Minimum Description Length (MDL) criterion in selecting a model for a given set of observations. In [3], an analytical derivation of the statistical properties of the various existing criteria is also obtained and used to construct hypotheses tests to carry out the BB search. None of these BB search techniques, however, accounts for any potentially available prior information about the signal. Accounting for the prior information in the search algorithm may provide a solution to the asymptotic unboundedness of the optimal threshold [5]. The size of the threshold in large data sets may indeed be overwhelming for some features of the signal of interest.

In Section 2 of the paper, we give some relevant background. In Section 3, we address the BB search of a signal in noise taking into account any available statistical prior information. In Section 4, we propose a Bernoulli-Gaussian model for the components of a signal in its BB. In Section 5, we provide some simulation results to substantiate the proposed method. Finally, some concluding remarks are given in Section 6.

2. BB REPRESENTATIONS

The determination of the "best representation" of a signal in a wavelet packet or Malvar's wavelet basis generally relies on the minimization of an additive criterion. The unnormalized entropy is usually retained as a cost function but, as will be shown later, criteria which are often more meaningful from a statistical point of view can also be introduced. To obtain an efficient search of the BB, the dictionary $\mathcal{D}$ of possible bases is structured according to a binary tree. Each node (j, m) (with $j \in \{0, \ldots, J\}$ and $m \in \{0, \ldots, 2^j -$

The work of the second author was in part supported by grants from ARO(DAAL03-92-G-0115) (Center for Intellingent Control), AFOSR (F49620-95-1-0083).

$1\}$) of the tree then corresponds to a given orthonormal basis $\mathcal{B}_{j,m}$ of a vector subspace of $\ell^2(\{1,\ldots,K\})$.[1] An orthonormal basis of $\ell^2(\{1,\ldots,K\})$ is then $\mathcal{B}_{\mathcal{P}} = \cup_{(j,m)/I_{j,m}\in\mathcal{P}}\mathcal{B}_{j,m}$ where $\mathcal{P}$ is a partition of $[0,1[$ in intervals $I_{j,m} = [2^{-j}m, 2^{-j}(m+1)[$. By taking advantage of the property

$$\mathrm{Span}\{\mathcal{B}_{j,m}\} = \mathrm{Span}\{\mathcal{B}_{j+1,2m}\} \bar{\oplus} \mathrm{Span}\{\mathcal{B}_{j+1,2m+1}\},$$

a fast bottom-up tree search algorithm was developed in [1] to optimize the partition $\mathcal{P}$. For the sake of simplicity, we shall subsequently number each possible partition with an integer $n \in \{1,\ldots,N\}$.

3. BAYESIAN APPROACH

3.1. PROBLEM STATEMENT

A natural way to incorporate available prior knowledge about an observed signal in noise is provided by the Bayesian statistical framework. We should note that this framework had previously been proposed in a plain wavelet estimation problem [8], and to the best of our knowledge, is novel and original for BB search techniques. The BB search proposed here provides, in a sense, an adaptive representation selected from the dictionary of bases $\mathcal{D} = \{\mathcal{B}_1,\ldots,\mathcal{B}_N\}$.
Let $\boldsymbol{Y}_n$, $\boldsymbol{X}_n$ and $\boldsymbol{W}_n$ respectively denote the K-dimensional vector of components of $y(t)$, $x(t)$ and $w(t)$ (*i.e.* $t = 1,2,\cdots,K$) in a basis $\mathcal{B}_n$. Using the linearity property of wavelet packets and Malvar's wavelets transforms, we have

$$\boldsymbol{Y}_n = \boldsymbol{X}_n + \boldsymbol{W}_n.$$

The signals $x(t)$ and $w(t)$ are assumed to be two mutually independent stochastic processes. We will further assume that there exists a $n_0 \in \{1,\ldots,N\}$ such that the Probability Density Function (PDF) of $\boldsymbol{X}_{n_0}$ is $f(\cdot)$ and that of $\boldsymbol{W}_{n_0}$ is $g(\cdot)$. The integer n_0 in this probabilistic model, which in fact indexes the BB $\mathcal{B}_{n_0}$, appears as a hyperparameter which must be estimated from the observed data. Toward this end, we propose two possible approaches.

3.2. MAXIMUM LIKELIHOOD METHOD

Using the independence property of the processes $\boldsymbol{X}_n$ and $\boldsymbol{W}_n$, we derive the law of the observations $\boldsymbol{Y}_n$. We thus obtain an estimate of n_0 as

$$\hat{n}_0 = \arg\max_n (f * g)(\boldsymbol{y}_n) \ , \qquad (1)$$

[1] Discrete decompositions on the interval are used in this paper.

where the symbol "*" stands for the convolution operation and lower case $\boldsymbol{y}_n$ denotes an observed realization of $\boldsymbol{Y}_n$. An estimate of $\boldsymbol{x}_n$ can subsequently be obtained by the Maximum *A Posteriori* (MAP) estimate

$$\hat{\boldsymbol{x}}_{\hat{n}_0} = \arg\max_{\boldsymbol{x}_{\hat{n}_0}} g(\boldsymbol{y}_{\hat{n}_0} - \boldsymbol{x}_{\hat{n}_0})\, f(\boldsymbol{x}_{\hat{n}_0}) \ . \qquad (2)$$

3.3. MAXIMUM GENERALIZED LIKELIHOOD METHOD

The principle of the Maximum Generalized Likelihood (MGL) method is to determine

$$(\hat{n}_0, \hat{\boldsymbol{x}}_{\hat{n}_0}) = \arg\max_{n,\boldsymbol{x}_n} p_{\boldsymbol{Y}_n,\boldsymbol{X}_n}(\boldsymbol{y}_n, \boldsymbol{x}_n).$$

If $\hat{\boldsymbol{x}}_n$ is the following MAP estimate,

$$\begin{aligned} \hat{\boldsymbol{x}}_n &= \arg\max_{\boldsymbol{x}_n} p_{\boldsymbol{Y}_n,\boldsymbol{X}_n}(\boldsymbol{y}_n, \boldsymbol{x}_n) \\ &= \arg\max_{\boldsymbol{x}_n} g(\boldsymbol{y}_n - \boldsymbol{x}_n)\, f(\boldsymbol{x}_n) \ , \end{aligned} \qquad (3)$$

the resulting estimate of n_0 is given by

$$\hat{n}_0 = \arg\max_n g(\boldsymbol{y}_n - \hat{\boldsymbol{x}}_n)\, f(\hat{\boldsymbol{x}}_n) \ . \qquad (4)$$

Unlike the Maximum Likelihood (ML) method, the MGL is not guaranteed to provide a convergent estimate of n_0 but its computational cost is generally lower, and it may lead to better estimates for data sets of relatively short size [9].

4. BERNOULLI-GAUSSIAN PRIORS

The PDF $f(\cdot)$ reflects our prior knowledge about the signal $x(t)$ represented in a basis $\mathcal{B}_{n_0}$. In general, such a prior is most simply expressed when $\mathcal{B}_{n_0}$ yields the most parsimonious representation (*i.e.* best matched basis) of the underlying signal $x(t)$. To account for the property of the expected energy concentration in the BB, we select a distribution which would reflect a certain amount of "spikyness" and thus be adapted to a basis representation, *e.g.* a Bernoulli-Gaussian distribution.
The model for $\boldsymbol{X}_n = (X_n^1,\ldots,X_n^K)^T$ is used in tandem with a hidden indicator vector $\boldsymbol{Q}_n = (Q_n^1,\ldots,Q_n^K)^T$ of independent binary random variables. More specifically, $(X_{n_0}^k)_{1\le k\le K}$ is an i.i.d. sequence whose conditional densities are

$$p_{X_{n_0}^k|Q_{n_0}^k=0}(x_{n_0}^k) = \delta(x_{n_0}^k) \ , \qquad (5)$$

$$p_{X_{n_0}^k|Q_{n_0}^k=1}(x_{n_0}^k) = \gamma(x_{n_0}^k \mid \sigma_x^2) \ , \qquad (6)$$

where $\delta(\cdot)$ is the Dirac distribution, and $\gamma(\cdot \mid s^2)$ defines throughout the paper a Gaussian PDF with zero-mean

and variance s^2. We further define a mixture parameter $\varepsilon = P(Q^k_{n_0} = 1) \in [0,1]$. The noise components are also assumed to be independent Gaussian random variables with zero-mean and variance σ^2. Under these hypotheses, we obtain by using the unitary transform property

$$\mathrm{E}[\boldsymbol{W}_n \boldsymbol{W}_n^T] = \sigma^2 \boldsymbol{I}_K \quad \text{for all } n,$$

where $\boldsymbol{I}_K$ denotes the $K \times K$ identity matrix.
With these assumptions, it is more convenient to use a maximum generalized marginal likelihood approach than a MGL one.
We proceed to determine the MAP estimate $\hat{\boldsymbol{q}}_n$ of $\boldsymbol{q}_n$ when n is guessed to be the true basis index n_0. It can be shown that this amounts to thresholding the components y^k_n of $\boldsymbol{y}_n$, by noting that

$$\begin{aligned} & p_{\boldsymbol{Y}_n, \boldsymbol{Q}_n}(\boldsymbol{y}_n, \hat{\boldsymbol{q}}_n) \\ &= \prod_{k=1}^{K} \int_{-\infty}^{\infty} \gamma(y^k_n - x^k_n \mid \sigma^2) p_{X^k_n \mid Q^k_n = \hat{q}^k_n}(x^k_n) \\ &\qquad\qquad P(Q^k_n = \hat{q}^k_n) dx^k_n \\ &= \prod_{k=1}^{K} \gamma(y^k_n \mid \sigma^2_{\hat{q}^k_n}) P(Q^k_n = \hat{q}^k_n) \ , \end{aligned}$$

where

$$\sigma^2_{\hat{q}^k_n} = \begin{cases} \sigma^2 + \sigma^2_x & \text{if } \hat{q}^k_n = 1, \\ \sigma^2 & \text{if } \hat{q}^k_n = 0. \end{cases}$$

The corresponding threshold value $\chi \geq 0$ is given by

$$\chi^2 = \max\left\{ \frac{2\sigma^2(\sigma^2 + \sigma^2_x)}{\sigma^2_x} \ln\left(\frac{\sqrt{\sigma^2 + \sigma^2_x}(1-\varepsilon)}{\sigma\varepsilon} \right), 0 \right\}. \tag{7}$$

Note that in contrast with the result in [5], this value is independent of the data length K.
The BB index $\hat{n}_0$ is then estimated as the integer n which maximizes the PDF $p_{\boldsymbol{Y}_n, \boldsymbol{Q}_n}(\boldsymbol{y}_n, \hat{\boldsymbol{q}}_n)$. Equivalently, $\hat{n}_0$ also results by minimizing the following criterion which is additive with respect to the components y^k_n of $\boldsymbol{y}_n$,

$$\varepsilon(n) = -\sum_{k=1}^{K} \ln[\gamma(y^k_n \mid \sigma^2_{\hat{q}^k_n}) P(Q^k_n = \hat{q}^k_n)]. \tag{8}$$

This clearly preserves a fast tree search structure for both wavelet packets and Malvar's wavelets.
We subsequently obtain an estimate $\hat{\boldsymbol{x}}_{\hat{n}_0}$ of the coefficients in the BB as the vector $\boldsymbol{x}_{\hat{n}_0}$ maximizing the joint PDF

$$\begin{aligned} & p_{\boldsymbol{Y}_{\hat{n}_0}, \boldsymbol{Q}_{\hat{n}_0}, \boldsymbol{X}_{\hat{n}_0}}(\boldsymbol{y}_{\hat{n}_0}, \hat{\boldsymbol{q}}_{\hat{n}_0}, \boldsymbol{x}_{\hat{n}_0}) \\ &= \prod_{k=1}^{K} \gamma(y^k_{\hat{n}_0} - x^k_{\hat{n}_0} \mid \sigma^2) p_{X^k_{\hat{n}_0} \mid Q^k_{\hat{n}_0} = \hat{q}^k_{\hat{n}_0}}(x^k_{\hat{n}_0}) P(Q^k_{\hat{n}_0} = \hat{q}^k_{\hat{n}_0}). \end{aligned}$$

The solution is given by

$$\hat{x}^k_{\hat{n}_0} = \begin{cases} \dfrac{\sigma^2_x}{\sigma^2 + \sigma^2_x} y^k_{\hat{n}_0} & \text{if } \hat{q}^k_{\hat{n}_0} = 1, \\ 0 & \text{if } \hat{q}^k_{\hat{n}_0} = 0. \end{cases}$$

Note that this estimate is very reminiscent in form of a Wiener estimate, albeit nonlinear.
As previously noted, the ML approach can also be applied. The selection of n_0 is also carried out by the minimization of an additive criterion.

5. SIMULATION RESULTS

5.1. BERNOULLI-GAUSSIAN SIGNALS

The above procedure has been implemented for an observed noisy process of length 131072 samples and with a Signal to Noise ratio and mixture parameter ε as indicated in Fig. 1. The original signal is Bernoulli-Gaussian in the wavelet basis. The resulting reconstruction error using the Bayesian BB approach is compared to the MDL-based BB search (see [6, 3]), and the significant gain in performance, nicely corroborates our original conjecture.

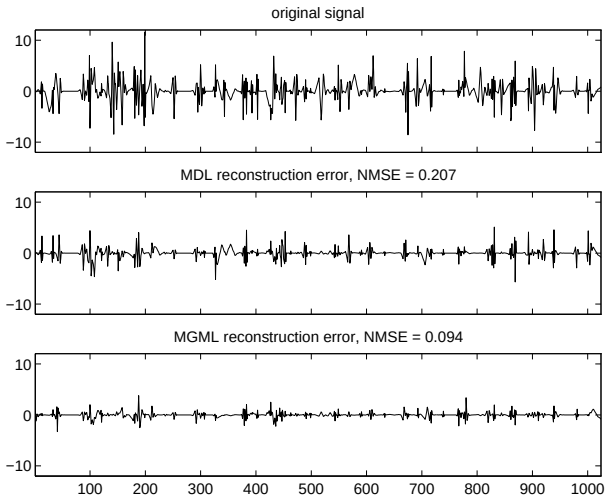

Figure 1: Wavelet packet denoising with a Bernoulli-Gaussian model ($\sigma = 1$, $\sigma_x = 5$ and $\varepsilon = 0.1$).

5.2. ARBITRARY SIGNALS

In our second example, we do not impose the statistical structure on the signal and choose an arbitrary noisy process (*e.g.* we chose signals from Donoho and Johnstone's database of sample signals), for which we estimate the parameters.

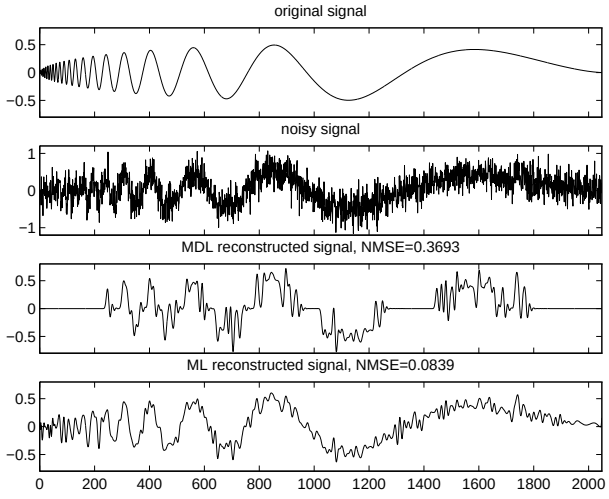

Figure 2: Wavelet packet denoising of a doppler signal with a 3-level tree search.

The parameter vector $\boldsymbol{\theta}_n = (\sigma_x^2, \varepsilon)_n$ and the hidden process $\boldsymbol{q}_n$ have been estimated by a MGL method:

$$(\hat{\boldsymbol{\theta}}_n, \hat{\boldsymbol{q}}_n) = \arg \max_{\boldsymbol{\theta}_n, \boldsymbol{q}_n} p_{\boldsymbol{Y}_n, \boldsymbol{Q}_n}(\boldsymbol{y}_n, \boldsymbol{q}_n \mid \boldsymbol{\theta}_n). \quad (9)$$

Starting from an initial guess $\hat{\boldsymbol{\theta}}_n^{(0)}$, a cross-optimization technique provides the MGL estimate

$$\begin{cases} \hat{\boldsymbol{q}}_n^{(\ell)} = \arg \max_{\boldsymbol{q}_n} p_{\boldsymbol{Y}_n, \boldsymbol{Q}_n}(\boldsymbol{y}_n, \boldsymbol{q}_n \mid \hat{\boldsymbol{\theta}}_n^{(\ell-1)}), \\ \hat{\boldsymbol{\theta}}_n^{(\ell)} = \arg \max_{\boldsymbol{\theta}_n} p_{\boldsymbol{Y}_n, \boldsymbol{Q}_n}(\boldsymbol{y}_n, \hat{\boldsymbol{q}}_n^{(\ell)} \mid \boldsymbol{\theta}_n), \end{cases}$$

which gives

$$\begin{aligned} (\hat{\varepsilon})_n^{(\ell)} &= \frac{\mathbf{1}^T \hat{\boldsymbol{q}}_n^{(\ell)}}{K}, \\ (\hat{\sigma}_x^2)_n^{(\ell)} &= \frac{\boldsymbol{y}_n^T \hat{\boldsymbol{q}}_n^{(\ell)}}{\mathbf{1}^T \hat{\boldsymbol{q}}_n^{(\ell)}} - \sigma^2. \end{aligned}$$

with

$$\mathbf{1}^T = [1 1 \cdots 1].$$

It should be noted that this method only guarantees the convergence to a local optimum. This nevertheless results in estimates which appear to be more robust than those based on an MDL criterion, particularly when a low Signal to Noise ratio prevails or when the number of levels in the decomposition tree is limited (Fig. 2). This also may be of interest when dealing with arbitrarily long signals.

6. CONCLUSION

The issue of finding a "best representation" for stochastic signals has been considered in a Bayesian framework using "spikyness" priors via Bernoulli-Gaussian mixtures. Both MGL and ML methods lead to the classical tree search algorithm developed for wavelet packets and Malvar's wavelets, allowing a fast computation of the BB. Several extensions addressing alternative approaches for estimating the model parameters can be envisaged and addressed in [10].

7. REFERENCES

[1] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Informat. Theory*, vol. IT-38, pp. 713–718, Mar. 1992.

[2] D. L. Donoho and I. M. Johnstone, "Ideal denoising in an orthogonal basis chosen from a library of bases." to appear in C. R. Acad. Sci. Paris, 1994.

[3] H. Krim and J.-C. Pesquet, "On the statistics of best bases criteria," in *Wavelets and statistics* (A. Antoniadis, ed.), Lecture Notes in Statistics, Springer Verlag, 1995.

[4] H. Krim, S. Mallat, D. Donoho, and A. Willsky, "Best basis algorithm for signal enhancement," in *ICASSP*, (Detroit, MI), IEEE, May 1995.

[5] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, Sept. 1994.

[6] D. L. Donoho, "Denoising by Soft-Thresholding," *IEEE Trans. Informat. Theory*, vol. 41, pp. 613–627, May 1995.

[7] P. Moulin, "A wavelet regularization method for diffuse radar target imaging and speckle noise reduction," *Journ. Math. Imaging and Vision*, vol. 3, pp. 123–134, 1993.

[8] B. Vidakovic, "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors." Internal Report, Duke University, 1995.

[9] F. Champagnat and J. Idier, "An alternative to standard maximum likelihood for Gaussian mixtures," in *Proc. IEEE Conf. Acoust., Speech, Signal Processing*, (Detroit, USA), pp. 2020–2023, May 9-12 1995.

[10] D. Leporini, J.-C. Pesquet, and H. Krim, "Best basis representations based on prior statistical models." Tech. Rep., Laboratoire des Signaux et Systèmes, France, in preparation.