A NEW BAYESIAN MODEL AVERAGING FRAMEWORK FOR WAVELET-BASED SIGNAL PROCESSING

Yi Wan and Robert D. Nowak

Department of Electrical and Computer Engineering
Rice University, USA

ABSTRACT

This paper develops a new signal modeling framework using Bayesian model averaging formulation and the redundant or translation-invariant wavelet transform. The aim of this framework is to provide a paradigm general enough to effectively treat fundamental problems arising in wavelet-based signal processing, segmentation, and modeling. Unlike many other attempts to mitigate the translation-dependent nature of wavelet analysis and processing, this framework is based on a well-defined statistical model averaging paradigm and improves over standard translation-invariant schemes for wavelet denoising. In addition to deriving new and more powerful signal modeling and denoising schemes, we demonstrate that certain existing methods are special suboptimal solutions of our proposed model averaging criterion. Experimental results demonstrate the promise of this framework.

1. INTRODUCTION

Wavelet-based signal processing has achieved tremendous success in theory and application in areas such as denoising, compression, and pattern recognition. However, the undesirable feature of translation-dependence, meaning that different shifts of the same input signal will have different representations in the wavelet transform, can lead to problems such as blocky artifacts in denoising. In [1, 4] a simple averaging method is proposed to partially solve this problem. That is, apply the wavelet transform to each translated version of the input signal, carry out a wavelet domain signal processing procedure, apply the inverse wavelet transform followed by inverse translation, then in the end take the average of such processed signals. It's described mathematically as

$$\widehat{\mathbf{x}} = \frac{1}{N} \sum_{j=1}^{N} T_{-j} \mathcal{W}^{-1} \, \delta \left(\mathcal{W} T_j \mathbf{y} \right)$$

where $\mathbf{y}$ is the noisy observed signal, T_j is a translation operator indexed by shift j, with inverse translation T_{-j}, $\mathcal{W}$ is the wavelet transform operator, δ is the wavelet thresholding/shrinkage operator and N is the total number of possible translations.

This work was supported by the National Science Foundation, grant no. MIP–9701692 and the Army Research Office, grant no. DAAD19-99-1-0349.

Even though this procedure provides one of the best known signal denoising techniques, the treatment is still very crude, since it simply *globally* averages the different results together instead of attempting to combine the best (most reliable) features of each in a more constructive and necessarily *local* fashion. It may help to consider this from a different perspective. Instead of translating the input signal, we can inverse-translate the wavelet basis, so that each such translation provides a new wavelet basis. The above procedure is thus equivalently expressed as: apply each of these different bases to the input signal, process them, then take the average. Most often, each different wavelet basis accurately/parsimoniously captures only a few local features (*e.g.,* edges or singularities that align well with the given basis); most features are not aligned with just one orthogonal basis. Ideally, we would like to extract only those local features that are well-modeled in each basis and then "patch" them together. A formal process that performs this ambitious goal is the contribution of this paper.

To illustrate the idea, consider an image segmentation problem. An image may be viewed as consisting of several independent components, such as a table, a piece of grass, or different textures of some material. The aim of segmentation is to automatically segment the image into these different components. In our new framework, we assume that each "component" is best described by one of the possible wavelet basis systems, as shown in figure 1; each region is best described by a particular wavelet basis. That is, region $B1$ is modeled with wavelet basis 1, region $B2$ with basis 2, and so on.

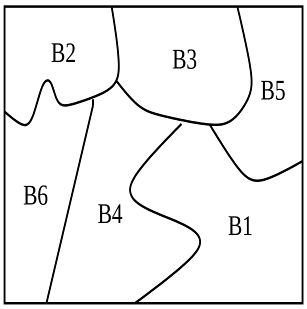

Figure 1: *Local image "component" representations in terms of different wavelet bases.*

The goal of our framework is two-fold: 1. **identify** components/regions of a signal or image that are best represented by a particular wavelet basis; 2. **fuse** the resulting collection of local component representations into a global one in a natural and coherent way. This type of problem can be formally posed and tackled using statistical model averaging techniques. The model averaging technique was first used by Laplace to account for model uncertainty. Lately different flavors of this technique have been formulated as Bayesian model averaging (BMA) methods. For a good tutorial, see [3]. In this paper we formulate a framework using BMA to effectively treat the signal modeling/processing problem mentioned above. We first present the formulation in Section 2, then address detailed approximation and implementation issues in Section 3. Experimental results demonstrate the advantages of our new framework. In the end, conclusions are drawn and avenues for future work are discussed in Section 4.

2. BMA FORMULATION

For the purpose of clarity we develop the formulation with signal denoising in mind; modifications and generalizations to related applications such as signal segmentation and fusion are not difficult to make. Assume an underlying signal

$$\mathbf{x} = (x_i)_{i\in\mathcal{I}} \in \mathbb{R}^N$$

and suppose that we observe this signal in noise:

$$\mathbf{y} = \mathbf{x} + \mathbf{n}$$

where $\mathbf{n}$ is usually white Gaussian noise. Furthermore, assume that we have a finite collection of statistically independent models $\{M_j\}_{j\in\mathcal{J}}$, $\mathcal{J} = \{1,\ldots,M\}$, each of which is parameterized by a coefficient vector $\boldsymbol{\theta}_j$. In the context of wavelet-based modeling, each model corresponds to a different wavelet basis expansion whose coefficients are random variables, and we take the coefficient vectors of different bases/models to be statistically independent of each other. Assuming each of the bases is unique, each model induces a *different* prior probability density for the signal denoted $p(\mathbf{x}|M_j)$.

A model indicator function space $\mathcal{G}$ is defined as a set of mappings of the form

$$g : \mathcal{I} \times \mathcal{J} \to \{0,1\}$$

such that for each sample/pixel index i the elements $g(i,j)$, $j = 1,\ldots,M$, are zero for all but one value of j, say $j = j^*$, indicating that model M_{j^*} generated x_i. In other words, g can be viewed as a matrix whose i-th row $g(i,\cdot)$ is the model indicator function for x_i. We assume the entire signal $\mathbf{x}$, which we observe in noise, is actually a composite of the various wavelet-based models, with g indicating the mapping of this composition. For example, Figure 1 illustrates the "mask" generated by one particular g.

Now if we observe data $\mathbf{y}$ and wish to infer $\mathbf{x}$, a natural approach is to compute the posterior distribution of $\mathbf{x}$, which can be expressed as

$$p(\mathbf{x}|\mathbf{y}) = \sum_{g\in\mathcal{G}} p(\mathbf{x}|g,\mathbf{y})p(g|\mathbf{y}) \tag{1}$$

where by Bayes rule

$$p(g|\mathbf{y}) = \frac{p(\mathbf{y}|g)p(g)}{p(\mathbf{y})}$$

Then the maximum *a posterior* (MAP) estimator of $\mathbf{x}$ is given by

$$\begin{aligned}\widehat{\mathbf{x}} &= \operatorname*{argmax}_{\mathbf{x}}\ p(\mathbf{x}|\mathbf{y}) \\ &= \operatorname*{argmax}_{\mathbf{x}} \sum_{g\in\mathcal{G}} p(\mathbf{x}|g,\mathbf{y})p(\mathbf{y}|g)p(g) \end{aligned} \tag{2}$$

The physical interpretation of (2) is easy to see. For each model composition, dictated by g, we can infer the best estimate from the term $p(\mathbf{x}|g,\mathbf{y})$. Each of these will be weighted by the overall *marginal likelihood* of the composite model to the data as expressed by the term $p(\mathbf{y}|g)$. Calculating the MAP estimate is, unfortunately, far from easy, but we will develop approximate methods that are computationally feasible and lead to good results.

For any fixed $g \in \mathcal{G}$ and $j \in \mathcal{J}$, let

$$\mathbf{x}_{g,j} = \{x_i :\ i \in \mathcal{I},\ g(i,j) = 1\}$$

Then for each $g \in \mathcal{G}$, $\{\mathbf{x}_{g,j}\}_{j\in\mathcal{J}}$ is a partition of the elements of $\mathbf{x}$ corresponding to the regions indicated by g. In simple terms, $\mathbf{x}_{g,j}$ is simply the subset of samples/pixels that were generated by model M_j. Since all models are assumed to be mutually independent, we have

$$p(\mathbf{x}|g,\mathbf{y}) = \prod_{j\in\mathcal{J}} p(\mathbf{x}_{g,j}|g,\mathbf{y})$$

and additionally assuming white Gaussian observation noise noise

$$p(\mathbf{y}|g) = \prod_{j\in\mathcal{J}} p(\mathbf{y}_{g,j}|g)$$

Thus (2) can be written as

$$\widehat{\mathbf{x}} = \operatorname*{argmax}_{\mathbf{x}} \sum_{g\in\mathcal{G}} \prod_{j\in\mathcal{J}} p(\mathbf{x}_{g,j}|g,\mathbf{y})p(\mathbf{y}_{g,j}|g)p(g) \tag{3}$$

The prior distribution $p(g)$ plays an important in both the performance and computational tractability of the BMA procedure. The size of $\mathcal{G}$ may be huge — for example, with a length N signal and M models, $\mathcal{G}$ contains M^N elements. In signal or image denoising problems, we consider each model indicator function to be equally plausible *a priori*. Hence, we choose an uninformative, uniform prior for g (*i.e.*, $p(g) = const.$). More sophisticated priors may be advantageous for other problems such as segmentation, but we will not address such extensions here.

3. APPLICATION TO WAVELET-BASED SIGNAL DENOISING

In this section, we apply our framework to the wavelet-based signal denoising problem to show the advantages and implementation details of the framework set up in the previous section. We show that the usual global averaging method [1] is one special case in approximating the optimal solution. By more careful approximation choices we can obtain better estimators. For clarity of explanation we deal with 1-d signals with additive white Gaussian noise.

Here $\mathcal{I} = \{1, 2, \cdots, N\} = \mathcal{J}$ where N is the discrete signal length and clearly there are exactly N different signal (circular) translations, or equivalently N shifted wavelet basis systems.

3.1. Model Setup and Gaussian Approximation

Given a signal $\mathbf{x}$, let $\boldsymbol{\theta} = \mathcal{W}\mathbf{x}$ where $\mathcal{W}$ is the wavelet transform. It is well-known that the wavelet coefficients $\boldsymbol{\theta}$ tend to be uncorrelated and sparsely distributed, closely following a Gaussian mixture distribution [2]. Hence we model the wavelet coefficients as independent random variables and set the prior density of each wavelet coefficient θ_i to

$$\theta_i \sim \epsilon_i \mathcal{N}(0, \sigma_{i1}^2) + (1 - \epsilon_i)\mathcal{N}(0, \sigma_{i2}^2), \quad \sigma_{i1}^2 \ll \sigma_{i2}^2$$

where ϵ_i is a Bernoulli random variable with $p(\epsilon_i) = p_i^{\epsilon_i}(1 - p_i)^{1-\epsilon_i}$ and $\mathcal{N}$ represents Gaussian distribution. Then the wavelet coefficient of the noisy signal is distributed as

$$\omega_i \sim \epsilon_i \mathcal{N}(0, \sigma_{i1}^2 + \sigma^2) + (1 - \epsilon_i)\mathcal{N}(0, \sigma_{i2}^2 + \sigma^2)$$

where σ^2 is the noise variance.

The parameters $\{p_i, \sigma_{i1}^2, \sigma_{i2}^2\}_i$ can be robustly estimated directly from the observed data $\mathbf{y}$ using the EM algorithm [2]. The posterior distribution of θ_i is thus given as

$$\begin{aligned} p(\theta_1|\omega_i) &= p(\epsilon_i = 1|\omega_i)\mathcal{N}\left(\frac{\sigma_{i1}^2}{\sigma_{i1}^2 + \sigma^2}\omega_i, \frac{\sigma_{i1}^2\sigma^2}{\sigma_{i1}^2 + \sigma^2}\right) \\ &\quad + p(\epsilon_i = 0|\omega_i)\mathcal{N}\left(\frac{\sigma_{i2}^2}{\sigma_{i2}^2 + \sigma^2}\omega_i, \frac{\sigma_{i2}^2\sigma^2}{\sigma_{i2}^2 + \sigma^2}\right) \end{aligned}$$

Now each sample/pixel in the time/spatial domain will be a linear combination of a subset of these wavelet coefficients and hence will in general also have Gaussian mixture distribution of at least N components. However, under mild conditions it can be shown that $p(\epsilon_i = 1|\omega_i)$ is near 0 or 1 with probability 1. This implies that $p(\theta_i|\omega_i)$ is approximately Gaussian and hence each sample/pixel is a linear combination of Gaussian random variables. This result shows that the marginal density $p(x_i|g, \mathbf{y})$ is well approximated by a Gaussian density, a fact that will facilitate subsequent analysis. In Figure 2 the histograms for a block signal and a natural fruit image are shown to support this result.

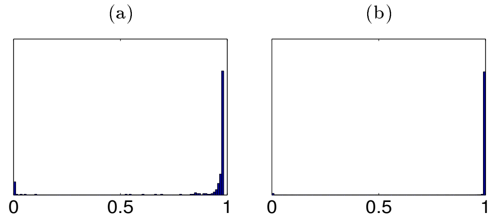

Figure 2: *Histogram of* $p(\epsilon_i = 1|\omega_i)$. *(a) Noisy block signal,* $\sigma^2 = 0.8$, *using Haar wavelet. (a) Noisy fruit image,* $\sigma^2 = 20$, *using Daubechies 6 wavelet.*

3.2. Suboptimal Solutions to MAP Criterion

In the most general case, the class $\mathcal{G}$ of generating functions in (2) contains N^N elements and thus makes the exhaustive summation unrealistic. This is a common problem encountered in BMA and the usual treatments are Occam's Window method[5] and Markov chain Monte Carlo model composition (MC^3) [6].

Notice that even when the number of summation terms is manageable, jointly estimating $\widehat{\mathbf{x}}$ is often still extremely difficult. So as the first step toward reasonable suboptimal solution to (2), we instead estimate each pixel x_i individually by marginalizing out all other pixels. This process leads to the following marginal MAP estimator for each $x_i,\ i = 1, 2, \cdots, N$:

$$\widehat{x}_i = \underset{x_i}{\operatorname{argmax}} \sum_{g \in \mathcal{G}} p(x_i|g, \mathbf{y})p(\mathbf{y}|g)p(g) \tag{4}$$

In this equation $p(x_i|g, \mathbf{y})$ is determined as $T_{-j}\mathcal{W}^{-1}\boldsymbol{\theta}_j$ where $g(i, j) = 1$ and T_{-j} is the translation operator of step $-j$ and $\boldsymbol{\theta}_j$ is the set of wavelet coefficients in model j. It's then easy to see that for any $g, g' \in \mathcal{G}$, if $g(i, j) = 1 = g'(i, j)$, then $p(x_i|g, \mathbf{y}) = p(x_i|g', \mathbf{y})$. Let's denote this common density as $p(x_i|j, \mathbf{y})$. Thus (4) can be simplified to

$$\widehat{x}_i = \underset{x_i}{\operatorname{argmax}} \sum_{j=1}^{N} p(x_i|j, \mathbf{y})c_{i,j} \tag{5}$$

where

$$c_{i,j} = \sum_{g \in \mathcal{G}, g(i,j)=1} p(\mathbf{y}|g)p(g) \tag{6}$$

Now to find a reasonable approximation of $c_{i,j}$, we first take notice that $p(g)$ is assumed to be uniformly distributed. Secondly, the number of terms in the summation in (6) is N^{N-1} (since we have N models). However, if we further assume that the overall *marginal likelihood* of the data $\mathbf{y}$ given g is not too sensitive to the change of a single pixel, then each $c_{i,j},\ j = 1, \ldots, N$ will be roughly the same, namely, $1/N$. With this approximation, the optimization simplifies to

$$\widehat{x}_i = \underset{x_i}{\operatorname{argmax}} \frac{1}{N} \sum_{j=1}^{N} p(x_i|j, \mathbf{y}) \tag{7}$$

The right side of (6) involves the sum (mixture) of N Gaussian densities, the solution to it in general has no closed form. A natural choice is to use a Monte Carlo technique or Newton's method to find the optimal solution for this problem. Instead, in this paper we propose two simple approximation methods. The first leads to the well-known TI denoising scheme, while the second leads to a new approach that provides better solutions.

Usually the collection of the means for each $p(x_i|j, \mathbf{y})$ is relatively clustered, since each mean value estimates the same sample/pixel value of the signal. Also note that each such mean value also maximizes $p(x_i|j, \mathbf{y})$. Thus, in this scenario when we have N estimators $\widehat{x}_{i,j}$, each given by the mean of the corresponding density $p(x_i|j, \mathbf{y})$. It is reasonable to use the average of these solutions to approximate the optimal solution. This leads directly to the TI denoising method of [1, 4].

An alternative to the above approximation is to adopt a joint marginal MAP estimation procedure to estimate x_i and the correct model for this sample.

$$(\widehat{x}_i, \widehat{j}) = \underset{x_i, j}{\operatorname{argmax}}\, p(x_i, j|\mathbf{y}) \tag{8}$$

This criterion simply selects the mean of the component of highest probability (highest peak) as the final estimator, since each $p(x_i, j|\mathbf{y})$ is (approximately) Gaussian. Specifically, we select

$$\widehat{x}_i = \mu_{i,j^*} \tag{9}$$

where $\mu_{i,j}$ and $\sigma^2_{i,j}$ are the mean and variance determined by the Gaussian density function $p(x_i|j, \mathbf{y})$ and $\sigma^2_{i,j^*} \leq \sigma^2_{i,j}$, for all $j = 1, 2, \cdots, N$. The overall computational complexity of this algorithm is $O(N^2)$. In general, with M models and a length N signal the complexity is $O(MN)$.

3.3. Experimental Results

In this section some experiments in 1-d signal denoising are carried out to demonstrate the potential of the new method. All experiments use the Haar wavelet basis. In Figure 4 two piecewise constant signals are shown. The signal in column (a) has a single discontinuity at a dyadic position, well-aligned with the standard Haar wavelet basis. The signal in column (b) has a discontinuity at a non-dyadic position, poorly aligned with the standard Haar basis. Since Haar basis fits signal (a) perfectly, but not signal (b), we see drastic performance difference (row 3). By comparison, the TI model produces relatively stable but less satisfactory results (row 4). The new method developed in this paper gives the best result for both signals (row 5). In Figure 3 we use the standard block test signal for comparison and again the new method outperforms the TI method.

4. CONCLUSIONS

In this paper we developed a novel wavelet-based signal modeling and estimation framework based on BMA. The major contribution is a new denoising method that performs significantly better than one the best known signal denoising method – the TI method [1, 4], which is also the result of a very crude approximation to the optimal solution. Although we focused denoising applications here, the BMA framework is also applicable to other problems such as signal/image segmentation, classification and data fusion. We plan to investigate these extensions in future work.

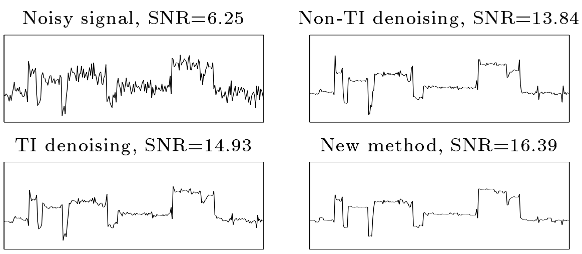

Figure 3: *Denoising comparison on block signal.*

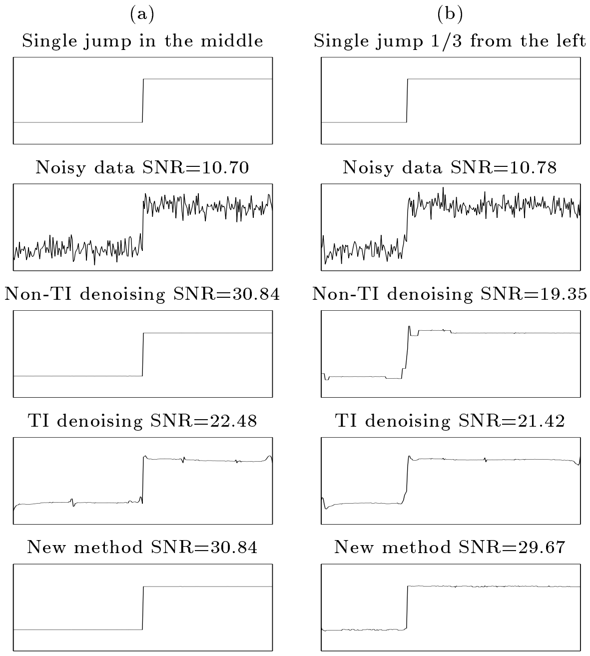

Figure 4: *Denoising comparison with single step function.*

5. REFERENCES

[1] R. Coifman and D. Donoho, "Translation invariant denoising," in *Lecture Notes in Statistics: Wavelets and Statistics*, New York: Springer-Verlag, pp. 125-150, 1995.

[2] M. Crouse, R. Nowak and R. Baraniuk, Wavelet-Based Statistical Signal Processing Using Hidden Markov Models, *IEEE Trans. Signal Processing*, vol. 46, pp.886–902, 1998

[3] J. A. Hoeting, D. Madigan and A. E. Raftery, "Bayesian Model Averaging: A Tutorial", Technical Report 9814, Dept. of Statistics, Colorado State University.

[4] M. Lang and H. Guo and J. E. Odegard and C. S. Burrus and R. O. Wells, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Processing Letters*, vol. 3, no. 1, pp. 10-12, 1996.

[5] D. Madigan and A. E. Raftery, "Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam's Window." *J. American Statistical Association*, vol. 89 pp. 1535-1546, 1994.

[6] D. Madigan and J. York, "Bayesian Graphical Models for Discrete Data" *International Statistical Review*, vol. 63 pp. 215-232, 1995.