

# SEQUENTIAL BAYESIAN WAVELET SHRINKAGE

Mark Coates\*

Rice University  
Electrical and Computer Engineering  
6100 S. Main, Houston, TX 77005

Arnaud Doucet

University of Cambridge  
Department of Engineering  
Cambridge, UK, CB2 1PZ.

## ABSTRACT

We propose a wavelet model that incorporates coefficient correlation and is expressed in state-space form, allowing the development and application of sequential estimation algorithms for wavelet denoising. We detail a sequential simulation-based estimation algorithm based on particle filters [6]. This algorithm allows Bayesian wavelet denoising to be performed on-line, enabling the processing of large data-sets, and it is intrinsically parallelizable. Our experiments indicate that the algorithm performance is comparable to most Bayesian framework batch-based algorithms.

## 1. INTRODUCTION

Wavelet shrinkage methods have proved to be a very effective means of performing non-parametric regression (or signal denoising). The original proposal by Donoho and Johnstone (see [5]) involved the straightforward application of soft or hard thresholding to the wavelet coefficients. More recently, wavelet shrinkage has been considered within a Bayesian framework [1, 2, 3], where a prior distribution is imposed on the wavelet coefficients of the unobserved process. The Bayesian approach provides scope for a more explicit (and accurate) modelling of the wavelet coefficients. The Bayesian methods in [1, 2, 3] assume independence between coefficients; the approaches in [4, 8] incorporate correlation in the coefficient modelling.

The Bayesian methods perform significantly better (both subjectively and in a mean squared error sense) than the thresholding approaches. Unfortunately, the methods are also substantially more computationally expensive. The estimation of the unobserved coefficients is performed *en bloc*, and most commonly involves expectation-maximisation (EM) or simulation-based algorithms [6]). When another block of observations (and hence noisy coefficients) become available, a completely new estimation task must be addressed. In this paper, we propose a wavelet-based model that incorporates coefficient correlation, but express it in a

state-space form. The state-space form allows us to perform the estimation exercise in a sequential fashion, enabling the on-line processing of very large data sets.

The coefficient estimation cannot be solved analytically, so we develop a sequential simulation-based algorithm founded on so-called particle filters [6]. Particle filters outperform classical suboptimum methods such as the extended Kalman filter and the Gaussian sum filter which rely on analytical approximations [6]. The algorithm we outline is intrinsically parallelizable, and experiments indicate that its performance is comparable to Bayesian framework batch-based algorithms.

### 1.1. The Wavelet Shrinkage problem

The regression model can be expressed as

$$y_t = x_t + \epsilon_t \quad (1)$$

where  $x_t$  is the underlying unknown process and  $\epsilon_t$  are independent  $\mathcal{N}(0, \sigma^2)$  random errors, representing additive white noise. By applying the discrete wavelet transform, the regression model can be expressed in the wavelet domain:

$$d_{j,k} = \beta_{j,k} + \varepsilon_{j,k} \quad (2)$$

where  $d_{j,k}$ ,  $\beta_{j,k}$  and  $\varepsilon_{j,k}$  are, respectively, the wavelet coefficients of  $y_t$ ,  $x_t$  and  $\epsilon_t$ . As the wavelet transform is an orthogonal transformation, the noise coefficients  $\varepsilon_{j,k}$  remain mutually independent and distributed according to  $\mathcal{N}(0, \sigma^2)$ . Wavelet shrinkage proceeds by estimating the unobserved  $\beta_{j,k}$ , and then transforming the coefficients back to the original domain by applying the inverse discrete wavelet transform to obtain an estimate of  $x_t$ .

### 1.2. Sequential Monte Carlo/Particle Filters

This section provides a brief introduction to sequential Monte Carlo methods; a more detailed review may be found in [6]. The aim of Monte Carlo simulation is to

---

\* This work was supported by Texas Instruments.

provide a method of approximating probability densities and integrals involving those densities. In the perfect sampling case, this is achieved by drawing  $N$  i.i.d. random samples  $\{\mathbf{x}_{0:n}^{(i)}; i = 1, \dots, N\}$  directly from the posterior distribution  $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ , so that the distribution can be empirically estimated as:

$$\widehat{P}(d\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) = \frac{1}{N} \sum_{i=1}^N \delta_{\mathbf{x}_{0:n}^{(i)}}(d\mathbf{x}_{0:n})$$

Expectations such as

$$I(f_n) = \int f_n(x_{0:n})p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) d\mathbf{x}_{0:n}$$

can be estimated as:

$$\overline{I}_N(f_n) = \int f_n(x_{0:n})\widehat{P}(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})d\mathbf{x}_{0:n} = \frac{1}{N} \sum_{i=1}^N f_n(\mathbf{x}_{0:n}^{(i)})$$

In most cases, however, it is usually impossible to draw directly from the posterior distribution. An alternative solution consists of using the importance sampling method. In this method, an ‘‘importance’’ distribution  $\pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$  is chosen; this distribution should be easy to sample from, and its support must include that of  $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ . Then one can write:

$$I(f_n) = \frac{\mathbb{E}_{\pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})}(f_n(x_{0:n})w(\mathbf{x}_{0:n}))}{\mathbb{E}_{\pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})}(w(\mathbf{x}_{0:n}))}$$

where  $w(\mathbf{x}_{0:n}) = p(\mathbf{y}_{0:n}|\mathbf{x}_{0:n})p(\mathbf{x}_{0:n})/\pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ . The Bayesian importance sampling estimate is then:

$$\widehat{I}_N^*(f_n) = \sum_{i=1}^N f_n(\mathbf{x}_{0:n}^{(i)}) \widetilde{w}_n^{(i)}$$

where an unnormalised importance weight  $w_n^{(i)} = p(\mathbf{y}_{0:n}|\mathbf{x}_{0:n}^{(i)})p(\mathbf{x}_{0:n}^{(i)})/\pi(\mathbf{x}_{0:n}^{(i)}|\mathbf{y}_{0:n})$ , and the normalised importance weights are equal to:

$$\widetilde{w}_n^{(i)} = \frac{w_n^{(i)}}{\sum_{j=1}^N w_n^{(j)}}$$

What is of particular importance in this method is that convergence results [6] imply that the importance sampling method can be interpreted not only as an integration technique, but also as a method for sampling from  $P(d\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ . The distribution can thus be empirically estimated as:

$$\widehat{P}_N(d\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) = \sum_{i=1}^N \widetilde{w}_n^{(i)} \delta_{\mathbf{x}_{0:n}^{(i)}}(d\mathbf{x}_{0:n})$$

In the sequential framework, we wish to obtain at time  $n$  an estimate of the distribution  $p(\mathbf{x}_{0:n}|\mathbf{y}_{0:n})$ . We then want to be able to estimate  $p(\mathbf{x}_{0:n+1}|\mathbf{y}_{0:n+1})$  but we don’t want to redo all the work involved in forming the first estimate. If we are considering the importance sampling method, this implies that at time  $n + 1$ , we want to form the trajectory (sample)  $\mathbf{x}_{0:n+1}^{(i)}$  without modifying the previous trajectory  $\mathbf{x}_{0:n}^{(i)}$ . This is possible if we only consider importance distributions that satisfy:

$$\begin{aligned} \pi(\mathbf{x}_{0:n}|\mathbf{y}_{0:n}) &= \pi(\mathbf{x}_{0:n-1}|\mathbf{y}_{0:n-1})\pi(\mathbf{x}_n|\mathbf{x}_{0:n-1}, y_{0:n}) \\ &= \pi(\mathbf{x}_0|\mathbf{y}_0) \prod_{k=1}^n \pi(\mathbf{x}_{0:k}|\mathbf{x}_{0:k-1}, y_{0:k}) \quad (3) \end{aligned}$$

At the time instant  $n + 1$ , we sample  $x_{n+1}^{(i)}$  from  $\pi(\mathbf{x}_{n+1}|\mathbf{x}_{0:n}^{(i)}, y_{0:n+1})$  and append it to  $x_{0:n}^{(i)}$  to form  $x_{0:n+1}^{(i)}$ ; we then update the associated importance weight  $w_n^{(i)}$ . This procedure is called sequential importance sampling (SIS).

At each time instant, the computational complexity of the SIS algorithm is  $\mathcal{O}(N)$ . Although  $N$  is large in practice, the algorithm is parallelizable, i.e., each particle can be treated as a separate entity. The particle values only have to be congregated when estimates and posterior probabilities are required.

Unfortunately, if the importance function is amenable to sequential sampling (of the form in (3)), then the unconditional variance of the weights, calculated by considering the  $y_{0:k}$  as random variables, increases (stochastically) with time. What this means is that as time progresses, more and more of the trajectories are trapped in parts of the state-space with very low posterior probability and thus have minimal weight. The few trajectories remaining in the highly probable parts of the subspace are heavily weighted, but do not provide a good approximation to the distribution.

Resampling is a technique designed to alleviate the degeneracy of the SIS algorithm. The technique involves periodically including an extra step in the SIS algorithm, wherein a new set of streams are sampled from the existing streams according to the discrete probability distribution described by the weights  $\widetilde{w}_k^{(i)}$ . The weights of the new streams are set to  $1/N$ . This increases the number of good streams, improving the estimation of future states. The resampling procedure has some drawbacks. There is a loss of diversity, as many of the new chains are identical. Resampling also hampers the parallelisation of the SIS algorithm.

## 2. THE WAVELET MODEL

### 2.1. Mixture models for the marginal densities

The energy of the discrete wavelet transforms of the majority of real-world signals is concentrated in a few large coefficients. The marginal density of each coefficient is typically described by a heavy-tailed non-Gaussian pdf with a high peak at zero. The densities can be reasonably well modelled using Gaussian mixture models [1, 2, 4]. The adoption of a two-component model (with zero-mean Gaussians) is often sufficiently accurate for estimation. In such a model, a discrete, hidden binary state variable  $m_{j,k}$  is assigned to the wavelet coefficient at scale  $j$  and time-index  $k$ . The mixture model is then comprised of a zero-mean Gaussian of small variance to represent the high peak in the marginal density corresponding to the coefficients that have small magnitude (the majority) and a zero-mean, large variance Gaussian to represent the heavy tails. In the limiting case, the small variance Gaussian can be replaced by a Dirac delta-function at 0.

Denoting the mixing probabilities as  $p_{0,j,k} = p(m_{j,k} = 0)$ , and  $p_{1,j,k} = p(m_{j,k} = 1)$ , the marginal densities can be expressed for the two component model (model *M1*) as:

$$M1 : p(\beta_{j,k}) = p_{0,j,k} \mathcal{N}(0, \sigma_{1,j,k}^2) + p_{1,j,k} \mathcal{N}(0, \sigma_{2,j,k}^2)$$

where  $\beta_{j,k}$  is the value of the wavelet coefficient, and  $\sigma_{1,j,k}^2$  and  $\sigma_{2,j,k}^2$  are the variances of the mixing densities. Similarly, for the case where the small variance Gaussian has been collapsed to a Dirac delta, we have the model:

$$M2 : p(\beta_{j,k}) = p_{0,j,k} \delta(\beta_{j,k}) + p_{1,j,k} \mathcal{N}(0, \sigma_{2,j,k}^2)$$

### 2.2. The joint density — a state-space model

The simplest approach to modelling the joint density of the set of wavelet coefficients is to assume independence between the marginal mixtures; this leads to the independent mixture (IM) model of [2] (see also [1, 4, 3]). In the wavelet transforms of many signals, however, there is correlation between coefficients [4]. This tends to be localised, potentially extending both across scale (clustering) and within scale (persistence). A number of wavelet domain models incorporate correlation [4, 8].

Crouse *et al.* propose a wavelet-domain hidden Markov model (HMM) framework in [4] that is tailored towards a more partial correlation structure, exploiting the commonly observed locality. Within this framework, the hidden states of the mixtures modelling the individual wavelet coefficients have a Markov dependency. A simple but effective structure is the Hidden

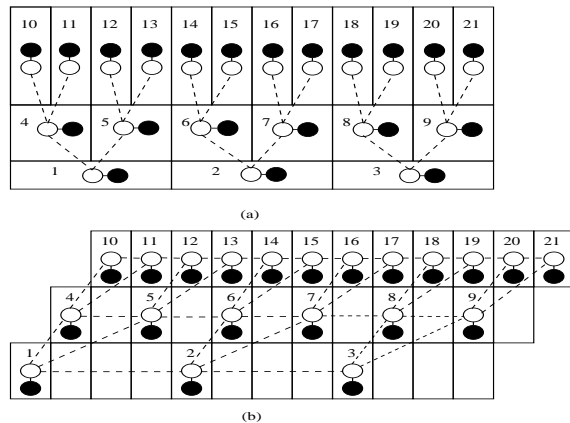


Figure 1: (a) The idealised wavelet transform tiling of the time-frequency plane. Each box contains a wavelet coefficient (dark circle) and its hidden state (white circle). The dashed lines indicate the state dependencies of the Hidden Markov Tree model [4]. (b) A mapping of the coefficients into a structure permitting a specification of the model in state-space form. The dashed lines indicate state dependencies in the model proposed in this paper.

Markov Tree (HMT). The dependencies in the HMT model are vertically *across scale*, as depicted by the dashed lines in Figure 1(a); the Markov dependency flows upwards through the tree-structure. In this paper, we adopt a similar model but express it in state-space form. We map the wavelet coefficients (and state variables) into a structure of the form shown in Figure 1(b). The structure preserves the region of influence of a wavelet coefficient at a particular scale but permits sequential analysis.

We consider the analysis of  $s$  scales of the wavelet transform, so that the mapped structure is comprised of  $s$  rows. The scale coefficients are considered relatively noise-free [4], so are neither included nor modified. We index each column in the structure by a time-index  $k$ . At any time  $k$ , wavelet coefficients are only observed for a subset of the analysed scales. An indicator variable  $I_{j,k}$  indicates the presence of an observation at scale  $j$  and time  $k$ . It is important to note that this indicator variable is known for all  $j$  and  $k$ , being simply a result of the mapping of the observed wavelet coefficients. The active set of coefficients at time  $k$  can then be defined as  $V_k \triangleq \{j : I_{j,k} = 1\}$ .

Denote the hidden binary state at scale  $j$  and index  $k$  as  $m_{j,k}$ . The vector of binary states at index  $k$ ,  $m_k \triangleq \{m_{j,k} : I_{j,k} = 1\}$ , has a varying dimension; the state space model is more readily specified if we define a vector  $r_k$  of fixed dimension  $s$  by iteratively augmenting

$m_k$ :

$$\begin{aligned} \mathbf{r}_k &= \{r_{0,k}, \dots, r_{s,k}\}, \quad \text{with } r_{j,1} = I_{j,1} m_{j,1} \\ r_{j,k} &= I_{j,k} m_{j,k} + (1 - I_{j,k}) m_{j,k-1} \quad \text{for } k > 0 \end{aligned}$$

This augmentation amounts merely to copying the binary state variable at scale  $j$  from the previous time index if no new observation exists at that scale in the current time index.

The hidden state dependency indicated in the structure of Figure 1(b) is reflected by  $p(\mathbf{r}_k | \mathbf{r}_{k-1})$ . The copying involved in constructing  $\mathbf{r}_k$  means that  $p(\mathbf{r}_k | \mathbf{r}_{k-1})$  varies with time. We can write  $p(\mathbf{r}_k | \mathbf{r}_{k-1})$  as:

$$\prod_{j \in \bar{V}_k} \delta_{r_{j,k-1}}(r_{j,k}) \prod_{j \in V_k} p(r_{j,k} | r_{j,k-1}, r_{j+1,k-1}) \quad (4)$$

Specifying the dependencies  $p(r_{j,k} | r_{j,k-1}, r_{j+1,k-1})$  for  $j = 1, \dots, s$  (which are presumed constant over time) therefore determines the transition probability at time  $k$ . In some cases, there is insufficient prior knowledge to specify these dependencies, and it becomes desirable to estimate them from the data using an initial block of data (see [4]).

We specify a state-space model by concatenating (over all  $(j, k)$  pairs in the union of active sets) the following relations (expressed at time  $k$  for any scale  $j$  in the active set  $V_k$ ):

$$\beta_{j,k} = b(r_{j,k}) v_{j,k} \quad (5)$$

$$d_{j,k} = \beta_{j,k} + n_{j,k} \quad (6)$$

Here,  $w_{j,k}$  is the wavelet coefficient of the underlying process  $x(t)$  at scale  $j$ , time index  $k$ ;  $d_{j,k}$  is the corresponding observed coefficient.  $n_{j,k}$  and  $v_{j,k}$  are mutually independent and distributed as  $\mathcal{N}(0, 1)$  and  $\mathcal{N}(0, \sigma^2)$ , respectively.  $b(r_{j,k})$  is determined for the different coefficient marginal density models as:

$$M1 : b(r_{j,k}) = r_{j,k} \sigma_{2,j,k} \quad (7)$$

$$M2 : b(r_{j,k}) = (1 - r_{j,k}) \sigma_{1,j,k} + r_{j,k} \sigma_{2,j,k} \quad (8)$$

### 3. THE SEQUENTIAL ESTIMATION ALGORITHM

In this section, we describe a sequential simulation based estimation algorithm for performing wavelet shrinkage under the model presented in the previous section. We restrict ourselves to the marginal density model M1 for brevity; the extension to model M2 is straightforward.

#### 3.1. Estimation objectives

Given at time  $k$  the observations  $\mathbf{d}_{1:k}$ , the aim of sequential wavelet shrinkage is to estimate the coefficients  $\beta_k$  of the hidden process  $x_t$ . Using the notation  $\beta_{1:k} \triangleq \{\beta_1, \dots, \beta_k\}$ , we can write the conditional posterior of the coefficients as:

$$p(\beta_{1:k} | \mathbf{d}_{1:k}, \mathbf{r}_{1:k}) = \prod_{j,k} p(\beta_{j,k} | d_{j,k}, r_{j,k}) \quad (9)$$

with  $p(\beta_{j,k} | d_{j,k}, r_{j,k} = 1) \sim \mathcal{N}(\frac{\sigma_{2,j,k}^2}{\sigma_{2,j,k}^2 + \sigma^2} d_{j,k}, \frac{\sigma_{2,j,k}^2}{\sigma_{2,j,k}^2 + \sigma^2})$ .

Assuming that the model parameters  $\sigma^2$ ,  $\sigma_2^2$  and  $p(\mathbf{r}_k | \mathbf{r}_{k-1})$  are exactly known, all Bayesian inference for the state-space model relies on the joint posterior  $p(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k})$  where  $p(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k}) = p(\beta_{1:k} | \mathbf{d}_{1:k}, \mathbf{r}_{1:k}) p(\mathbf{r}_{1:k} | \mathbf{d}_{1:k})$ . If the model parameters are unknown, block-based empirical estimation methods [4, 3, 7] are applied using the first portion of the available observations.

Given  $\mathbf{r}_{1:k}$ ,  $p(\beta_{1:k} | \mathbf{d}_{1:k}, \mathbf{r}_{1:k})$  is straightforwardly determined from (9).  $p(\mathbf{r}_{1:k} | \mathbf{d}_{1:k})$  could be computed exactly but this discrete distribution has  $s^k$  values and thus some approximations have to be made as time  $k$  increases. In this paper, we are interested in obtaining the filtering distribution  $p(\mathbf{r}_k, \beta_k | \mathbf{d}_{1:k})$ , and in particular the MMSE estimate of  $\beta_k$ , given by  $\hat{\beta}_{k|k} \triangleq \mathbb{E}_{p(\beta_k | \mathbf{d}_{1:k})} \{\beta_k\}$ . We perform an inverse discrete wavelet transform to sequentially estimate the unobserved process  $x_t$ .

#### 3.2. Monte Carlo simulation for estimation

If we were able to sample  $N$  i.i.d. random samples, called particles,  $(\mathbf{r}_{1:k}^{(i)}, \beta_{1:k}^{(i)})$ , according to  $p(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k})$ , then we could form empirical estimates of  $p(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k})$  and hence of  $\hat{\beta}_{k|k}$ . Unfortunately, it is impossible to sample efficiently from  $p(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k})$  at any time  $k$ . A solution to estimate  $p(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k})$  and  $\hat{\beta}_{k|k}$  consists of using the Bayesian importance sampling technique described in Section 1.2. The straightforward estimate is then:

$$\tilde{\beta}_{N,k|k} = \frac{\sum_{i=1}^N \beta_k^{(i)} w(\mathbf{r}_{1:k}^{(i)}, \beta_{1:k}^{(i)})}{\sum_{i=1}^N w(\mathbf{r}_{1:k}^{(i)}, \beta_{1:k}^{(i)})}$$

with  $w(\mathbf{r}_{1:k}) = p(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k}) / \pi(\mathbf{r}_{1:k}, \beta_{1:k} | \mathbf{d}_{1:k})$ .

Since we have the relationship  $p(\mathbf{r}_{1:k}, \beta_k | \mathbf{d}_{1:k}) = p(\mathbf{r}_{1:k} | \mathbf{d}_{1:k}) p(\beta_k | \mathbf{d}_{1:k}, \mathbf{r}_{1:k})$  where  $p(\beta_k | \mathbf{d}_{1:k}, \mathbf{r}_{1:k})$  is a Gaussian distribution, we obtain an approximation of  $p(\mathbf{r}_k, \beta_k | \mathbf{d}_{1:k})$  straightforwardly from an approximation of  $p(\mathbf{r}_{1:k} | \mathbf{d}_{1:k})$ . As  $\hat{\beta}_{k|k} =$

$\mathbb{E}_{p(\mathbf{r}_{1:k}|\mathbf{d}_{1:k})} [\mathbb{E}_{p(\beta_k|\mathbf{d}_{1:k},\mathbf{r}_{1:k})} (\beta_k)]$ , this means that we only have to draw  $N$  i.i.d. random samples  $(\mathbf{r}_{1:k}^{(i)})$  distributed according to a selected importance distribution  $\pi(\mathbf{r}_{1:k}|\mathbf{d}_{1:k})$ . An asymptotically consistent Monte Carlo estimate of  $\hat{\beta}_{k|k}$  is then:

$$\bar{\beta}_{N,k|k} = \frac{\sum_{i=1}^N \mathbb{E}_{p(\beta_k|\mathbf{d}_{1:k},\mathbf{r}_{1:k}^{(i)})} (\beta_k) w(\mathbf{r}_{1:k}^{(i)})}{\sum_{i=1}^N w(\mathbf{r}_{1:k}^{(i)})} \quad (10)$$

where  $w(\mathbf{r}_{1:k}) = p(\mathbf{r}_{1:k}|\mathbf{d}_{1:k})/\pi(\mathbf{r}_{1:k}|\mathbf{d}_{1:k})$ . The formation of this latter estimate is an example of ‘Rao-Blackwellisation’; it has been demonstrated [6] that the variances of the importance weights and the numerator and denominator of Rao-Blackwellised estimates are smaller than those of corresponding direct estimates.

### 3.3. Sequential Bayesian Importance Sampling

If we are only interested in estimating the current set of coefficients at any time  $k$ , it makes sense to apply the sequential importance sampling algorithm. Assuming that we form a Rao-Blackwellised estimate as in the previous section, we are required to choose an importance distribution  $\pi(\mathbf{r}_k|\mathbf{r}_{0:k-1},\mathbf{d}_{0:k})$ . A simple choice, and the one adopted in this paper, is to use the prior distribution  $p(\mathbf{r}_k|\mathbf{r}_{k-1})$  as importance distribution. The importance weight is then  $p(d_k|\mathbf{d}_{1:k-1},\mathbf{r}_{1:k}) = \prod_{j \in V_k} p(d_{j,k}|r_{j,k})$ , where  $p(d_{j,k}|r_{j,k}) \sim \mathcal{N}(0, \sigma^2 + r_{j,k} \sigma_{2,j,k}^2)$ . We must also choose a resampling strategy to reduce the effects of the degeneracy phenomenon in sequential importance sampling. In this paper, we adopt the residual resampling procedure (see [6]). Given at time  $k-1$ ,  $N \in \mathbb{N}^*$  random samples  $\mathbf{r}_{1:k-1}^{(i)}$  distributed approximately according to  $p(\mathbf{r}_{1:k-1}|\mathbf{d}_{1:k-1})$ , the particle filter proceeds as follows at time  $k$ .

---

#### Sequential Wavelet Shrinkage

#### Sequential Importance Sampling step

1. At time  $k = 0$ :

- For  $i = 1, \dots, N$ , sample  $r_0^{(i)} \sim p(\mathbf{r}_0)$ , a specified initial distribution.
- For  $i = 1, \dots, N$ , evaluate the unnormalised importance weights:  $w_0^{(i)} = p(\mathbf{d}_0|\mathbf{r}_0^{(i)})$ .
- Normalise the importance weights:  $\tilde{w}_0^{(i)} = w_0^{(i)} / \sum_{j=1}^N w_0^{(j)}$ .

2. For times  $k > 0$ :

- For  $i = 1, \dots, N$ , sample  $\tilde{\mathbf{r}}_k^{(i)} \sim p(\mathbf{r}_k|\mathbf{r}_{k-1}^{(i)})$  and  $\tilde{\mathbf{r}}_{1:k}^{(i)} \triangleq (\mathbf{r}_{1:k-1}^{(i)}, \tilde{\mathbf{r}}_k^{(i)})$ .

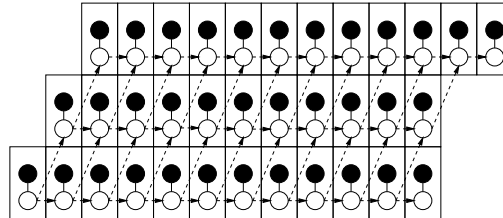


Figure 2: A scale-shifted, redundant tiling of the time-frequency plane corresponding to a redundant wavelet transform. The arrows indicate the dependencies between hidden states (clear circles) in the redundant state-space wavelet model.

- For  $i = 1, \dots, N$ , evaluate the unnormalised importance weights:  $w_k^{(i)} = p(\mathbf{d}_k|\tilde{\mathbf{r}}_k^{(i)})$ .
- Normalise the importance weights:  $\tilde{w}_k^{(i)} = w_k^{(i)} / \sum_{j=1}^N w_k^{(j)}$ .

#### Estimation step

- Estimate  $\hat{\beta}_{k|k}$  as  $\{\bar{\beta}_{j,k|k}; j \in V_k\}$  with

$$\bar{\beta}_{j,k|k} = \sum_{r_{j,k}^{(i)}=1} \tilde{w}_k^{(i)} \frac{\sigma_{2,j,k}^2}{\sigma_{2,j,k}^2 + \sigma^2} d_{j,k}$$

Alternatively calculate the posterior median estimate.

#### Selection step

- Multiply/Discard particles  $(\tilde{\mathbf{r}}_{1:k}^{(i)}; i = 1, \dots, N)$  with respect to high/low normalized importance weights  $\tilde{w}_k^{(i)}$  to obtain  $N$  particles  $(\mathbf{r}_{1:k}^{(i)}; i = 1, \dots, N)$ .

---

The computational complexity of this algorithm at each iteration is  $O(N)$ , and the algorithm is straightforwardly parallelizable.

## 4. TRANSLATION-INVARIANT DENOISING

Translation-invariant (TI) denoising procedures [7] perform much better than those based on a fixed origin transform. The TI algorithms essentially operate by averaging the denoised coefficients (or taking the median values) of a set of fixed origin algorithms.

A similar averaging could be adopted in the sequential framework by applying the algorithm to shifted data sets, but we would like to share some of the information determined in the inference for different shifts. The problem is that the wavelet coefficients observed for different shifts are not independent of one another, so that the relation  $p(\beta_{1:k}|\mathbf{d}_{1:k},\mathbf{r}_{1:k}) = \prod_{j,k} p(\beta_{j,k}|d_{j,k},r_{j,k})$  is no longer valid. Despite the violation of independence, we adopt an as-if-independent approach (similar to that adopted in [7]). We modify

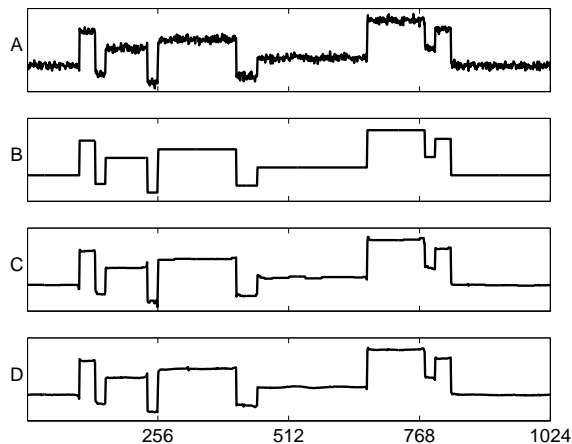


Figure 3: Example results of sequential denoising. A: Noisy blocks signal ( $\sigma_n^2 = 1$ ). B: Original signal. C: Sequential orthogonal denoising result. D: Sequential translation-invariant denoising result.

the dependency structure of our wavelet tree, so that direct Markov state dependencies occur between different shifted transforms rather than within a single shifted transform (see Figure 2). Figure 2 depicts a redundant tiling of the time–frequency plane, formed by interleaving all the sets of shifted transforms. Neighbouring states in the interwoven structure are related. We now apply the same algorithm as outlined in the preceding section. In this structure, one coefficient can be evaluated at each scale as each data sample is read, so the algorithm takes on a truly sequential flavour in the time domain.

### 5. RESULTS

We tested the orthogonal and redundant sequential shrinkage algorithms on Donoho and Johnstone’s standard length-1024 test signals [5]. Table 1 compares the results to other shrinkage approaches. Figure 3 depicts the performance of the algorithm for the Blocks signal.

For the results presented in the test, we used the tree structures of Figure 1(b) in the orthogonal algorithm, and Figure 2 in the redundant algorithm. The hyperparameters of the model were estimated using a block-based empirical Bayes approach. We estimated the variances by training the HMT model [4] on the first 128 samples of the first test signals. This procedure also provided a guide for the transition probabilities across scale; training a Hidden Markov chain model [4] (where dependencies only exist within scale) provided an estimate for transition probabilities within scale. Experiments indicated that the denoising performance was fairly robust to variation in the values of these parameters.

Method	Mean-squared error			
	Bumps	Blocks	Doppler	Heavisine
Sure [5]	0.683	0.222	0.228	0.095
Bayes [2]	0.350	0.099	0.165	0.087
IM [4]	0.335	0.105	0.170	0.080
HMT [4]	0.268	0.079	0.132	0.081
Seq. Orth	0.312	0.102	0.173	0.083
Seq. Red	0.232	0.066	0.102	0.069

Table 1: Denoising results for Donoho and Johnstone’s length-1024 test signals [5]. Noise variance  $\sigma_n^2 = 1$ .

### 6. CONCLUSIONS

We have proposed a sequential estimation algorithm based on particle filters to perform wavelet denoising. The algorithm is inherently parallelizable and allows Bayesian denoising to be performed on-line for models assuming complex correlation structures. The performance is comparable to the best block-based denoising techniques. The framework also provides scope for performing denoising in non-Gaussian noise environments.

### REFERENCES

- [1] F. Abramovich, T. Sapatinas, and B. Silverman. Wavelet thresholding via a Bayesian approach. *J. Royal Statistical Soc. B*, 60:725–749, 1998.
- [2] H. Chipman, E. Kolaczyk, and R. McCulloch. Adaptive Bayesian wavelet shrinkage. *J. Amer. Stat. Assoc.*, 92:1413–1421, 1997.
- [3] M. Clyde, G. Parmigiani, and B. Vidakovic. Multiple shrinkage and subset selection in wavelets. *Biometrika*, 85:391–402, 1998.
- [4] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-based signal processing using hidden Markov models. *IEEE Trans. Sig. Proc.*, 46:886–902, 1998.
- [5] D. Donoho and I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Stat. Assoc.*, 90:1200–1224, 1995.
- [6] A. Doucet. On sequential simulation-based methods for Bayesian filtering. Technical Report CUED/F-INFENG/TR.310, Cambridge University Engineering Department, 1998.
- [7] I.M. Johnstone and B.W. Silverman. Empirical Bayes approaches to mixture problems and wavelet regression. Technical report, Stanford University, 1998.
- [8] M. Vannucci and F. Corradi. Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *J. Royal Stat. Soc., Series B*, 61.