Perfect simulation for wavelet thresholding with correlated coefficients

Graeme K. Ambler[*] and B. W. Silverman

March 2, 2004

Abstract

We introduce a new method of Bayesian wavelet shrinkage for reconstructing a signal when we observe a noisy version. Rather than making the usual assumption that the wavelet coefficients of the signal are independent, we assume that they are locally correlated in both location (time) and scale (frequency). This leads us to prefer a novel prior structure to which is, unfortunately, analytically intractable. We demonstrate that it is possible to draw exact, independent samples from the posterior distribution using Coupling From The Past, making a simulation-based approach possible.

1 Introduction

Consider the the standard nonparametric regression problem

$$y_i = g(t_i) + \varepsilon_i. \tag{1}$$

where we observe a noisy version of an unknown function g at regularly spaced intervals t_i. The noise, ε_i is assumed to be independent and Normally distributed with zero mean and variance σ^2.

The standard wavelet-based approach to this problem is based on two properties of the wavelet transform:

1. A large class of "well-behaved" functions can be sparsely represented in wavelet-space.

2. The wavelet transform transforms independent, identically distributed noise to independent, identically distributed wavelet coefficients.

These two properties combine to suggest that a good way to remove noise from a signal is to transform the signal into wavelet space, discard all of the small coefficients (i.e. threshold), and perform the inverse transform. Since the true (noiseless) signal had a sparse representation in wavelet space, the signal will be concentrated in a small number of large coefficients. The noise, on the other hand, will still be spread evenly among the

[*] *Address for correspondence:* Department of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, U.K. Email: graeme.ambler@bristol.ac.uk

coefficients, so by discarding the small coefficients we must have discarded mostly noise and will thus have found a better estimate of the true signal.

The problem then arises of what to choose as a threshold value. Many methods exist for choosing the value of the threshold.

SureShrink (Donoho and Johnstone 1995) is a method for soft thresholding which minimises Stein's unbiased estimate of risk (Stein 1981). A different threshold is chosen for each level of the transform. The authors prove SureShrink is near-optimal in the way that it adapts to the smoothness of the underlying function.

Cross-validation is a general method which has been used in a number of areas of statistics. The principle is to split the data set into two pieces, a test set and a training set. The training set is then used to to fit a model (in this case a function). The test set is then used to assess the performance of the method. Nason (1996) suggests splitting the data into the odd- and even-numbered observations (we shall call them the 'odds' and the 'evens'). First the evens are used to get an estimator for the function (using some threshold t) and the sum of squared errors(SSE) between the estimate and the odds is calculated. Secondly, the odds are used to get an estimator of the function using the same threshold t, and the SSE between the new estimate and the evens is calculated. Finally, the combined SSE is minimised numerically over values of t.

False discovery rates (Benjamini and Hochberg 1995) were originally introduced in the field of multiple hypothesis testing, and control the expected proportion of false-positives. Abramovich and Benjamini (1996) use this methodology to control the expected number of coefficients which are not thresholded but should have been.

BayesThresh (Abramovich et al. 1998) uses a Bayesian hierarchical model, assuming independent $N(0, \sigma^2)$ noise. They use a mixture of a point mass at 0 and a $N(0, \tau^2)$ density as their prior on the population wavelet coefficients. The marginal posterior median of the population wavelet coefficient is then used as their estimate of it. This gives a thresholding rule, since the point mass at 0 in the prior gives non-zero probability that the population wavelet coefficient will be zero.

We discuss an extension of BayesThresh in Section 2. An outline of the paper is as follows. In Section 1.1 we introduce coupling from the past, which we use to simulate from our posterior distribution. In Section 1.2 we discuss the area-interaction point process, and a discrete version which we use as a prior distribution. In Section 1.3 we discuss an extension of coupling from the past which will allow us to sample exactly from the posterior distribution of our model. As already mentioned, we introduce an extension of the method of Abramovich et al. in Section 2. In Section 3 we discuss a method for generating independent samples from our posterior distribution using a variation of the extension of coupling from the past introduced in Section 1.3. In Section 4 we present a simulation study to compare our method with the others introduced in this section. Section 5 presents some conclusions and discusses possible avenues for future work.

1.1 Coupling from the past

It has long been considered a failing of Markov chain Monte Carlo (MCMC) that one can rarely be absolutely sure that the Markov chain which is used for a given simulation has converged to its stationary distribution. This means that it is not possible to generate an unbiased sample[1]. It would be nice, therefore, to find a method for guaranteeing that the chain has reached equilibrium, and thus that the resulting sample will be unbiased. Coupling from the past (Propp and Wilson 1996) is a method for doing this.

The motivation behind coupling from the past (CFTP) is the following. Suppose that it is desirable to sample from the stationary distribution of an ergodic Markov chain $\{Z_t\}$ on some (finite) state space X with states $1, \ldots, n$. It is clear that if it were possible to go back an infinite amount in time, start the chain running (in state $Z_{-\infty}$) and then return to the present, the chain would (with probability 1) be in its stationary distribution when one returned to the present (i.e. $Z_0 \sim \pi$, where π is the stationary distribution of the chain). This is clearly not feasible in practice! Propp and Wilson (1996) showed that in fact we can achieve the same goal by going back a finite (random) amount of time only.

Consider a finite state space with n states, and that we set not one, but n chains $\{Z_t^{(1)}\}, \ldots, \{Z_t^{(n)}\}$ running at a fixed time $-M$ in the past, where $Z_{-M}^{(i)} = i$ for each chain $\{Z_t^{(i)}\}$. Now let all the chains be coupled Lindvall (1992) so that if $Z_s^{(i)} = Z_s^{(j)}$ at any time s then $Z_t^{(i)} = Z_t^{(j)} \quad \forall t \geq s$. Then if all the chains ended up in the same state j at time zero (i.e. $Z_0^{(i)} = j \quad \forall i \in X$), we would know that whichever state a chain passing from time minus infinity to zero was in at time $-M$, the chain would end up in state j at time zero. Thus j must be a sample from the stationary distribution of the Markov chain in question.

1.2 The Area-interaction point process

The area-interaction point process (Baddeley and van Lieshout 1995) is a spatial point process capable of producing both moderately clustered and moderately ordered patterns dependent on the value of its clustering parameter. It was introduced primarily to fill a gap left by the Strauss point process (Strauss 1975), which can only produce ordered point patterns (Kelly and Ripley 1976). The general area-interaction process is defined as follows.

Let χ be some locally compact complete metric space and $\mathfrak{R}^f$ be the space of all possible configurations of points in χ. Let ν be a finite Borel regular measure on χ and $Z : \chi \rightarrow \mathcal{K}$ be a myopically continuous function (where $\mathcal{K}$ is as usual the class of all compact subsets of χ). Then the probability density of the general area-interaction process is given by

$$p(X) = \alpha \lambda^{N(X)} \gamma^{-\nu(U(X))} \tag{2}$$

with respect to the unit rate Poisson process, where $N(X)$ is the number of points in configuration $X = \{x_1, \ldots, x_{N(X)}\} \in \mathfrak{R}^f$, α is a normalising constant and $U(X) = \bigcup_{i=1}^{N(X)} Z(x_i)$. In Section 2 we define the particular special case of this point process that we use.

Kendall (1998) extended CFTP to cover simulation of the area-interaction process discussed in Section 1.2, which has an infinite state space. The method makes use of

[1] A *biased sample* is one whose distribution is different from the equilibrium distribution of the Markov chain used to generate it, so that the estimate of any quantity depending on the equilibrium distribution may be biased.

a monotone coupling and stochastic domination. A coupling is monotone if whenever $Z_t^{(i)} \geq Z_t^{(j)}$ then $Z_{t+k}^{(i)} \geq Z_{t+k}^{(j)}$ $\forall k > 0$. Given a monotone coupling and unique minimum and maximum elements we need only simulate Markov chains starting in the maximum and minimum states and check that these two have coalesced at time 0, since chains starting in all other states would be sandwiched between these two. As there is no natural maximum element for the area-interaction process, Kendall used a Poisson process which stochastically dominates the area interaction process of interest to generate a maximum process. More recently, Ambler (2002) extended these techniques to more general classes of point processes. We describe this technique in the following section.

1.3 Perfect simulation of spatial point processes

For simplicity we restrict our discussion to spatial point processes on the unit square $[0,1] \times [0,1] \subseteq \mathbb{R}^2$. Suppose that we wish to sample from such a spatial point process with density

$$p(X) = \alpha\lambda^{N(X)} \prod_{i=1}^{m} f_i(X),$$

where α, $\lambda \in (0,\infty)$ and $f_i : \mathfrak{R}^f \to \mathbb{R}$ are positive valued bounded monotonic functions. Ambler (2002) shows that this is possible using the following algorithm.

Let D be a two-dimensional Poisson process with rate equal to

$$\lambda \prod_{i=1}^{m} \max_{X,\{x\}} \left(\frac{f_i(X \cup \{x\})}{f_i(X)} \right), \tag{3}$$

evolving over time according to a birth-death process with birth rate equal to (3) and unit death rate. Let $D(T)$ be the configuration of points in process D at time T. For simplicity of notation, constrain this function to be right-continuous, so that if there is a birth in D at time T then $D(T)$ is the configuration which existed in D immediately prior to the birth.

Now let U be a birth-death process which is started from an initial configuration equal to that of D at some time in the past, and L be a birth-death process which is started from an initial configuration equal to a thinned version of D, where points are accepted with probability

$$\prod_{i=1}^{m} \min_{X,\{x\}} \left(\frac{f_i\left(X \cup \{x\}\right)}{f_i(X)} \right) \Big/ \max_{X,\{x\}} \left(\frac{f_i\left(X \cup \{x\}\right)}{f_i(X)} \right)$$

The processes U and L evolve through time as follows. If a point $\{u\}$ is born in D at time T then $\{u\}$ is also born in U at time T with probability

$$\prod_{i=1}^{m} \max \left\{ \frac{f_i[U(T) \cup \{u\}]}{f_i[U(T)]}, \frac{f_i[L(T) \cup \{u\}]}{f_i[L(T)]} \right\} \Big/ \max_{X,\{x\}} \left(\frac{f_i\left(X \cup \{x\}\right)}{f_i(X)} \right) \tag{4}$$

The point $\{u\}$ is born in L at time time T with probability

$$\prod_{i=1}^{m} \min \left\{ \frac{f_i[U(T) \cup \{u\}]}{f_i[U(T)]}, \frac{f_i[L(T) \cup \{u\}]}{f_i[L(T)]} \right\} \Big/ \max_{X,\{x\}} \left(\frac{f_i\left(X \cup \{x\}\right)}{f_i(X)} \right) \tag{5}$$

If a point dies in D then if it existed in U or L then it dies there also.

Finally, generate D backwards in time from zero to some time $-T$ and start U and L there. Now run them forward to time zero. If $U(0) = L(0)$ then the configuration $U(0)$ (or equivalently $L(0)$) is a sample from the required spatial point process. If not, we must generate D further back in time and try again, keeping the probabilities used for acceptance/rejection used in the first round.

We make use of a slightly modified version of this technique in Section 3 to sample from our posterior distribution.

2 An Extension of Bayesian Wavelet Thresholding

We describe a novel thresholding procedure which uses a discrete area-interaction process to model the correlation between neighbouring coefficients in the wavelet transform.

The principle behind our method is to model the discrete wavelet transform as a marked lattice process. The 'lattice' is the natural binary tree which is commonly used to represent the coefficients. A discretized area-interaction process is used as a prior on the distribution of non-zero coefficients. We also make use of the extra information gained by allowing multiple points to exist at a single location, using the number of points as a shrinkage factor. This is different from Abramovich et al. (1998), where the implicit assumption was that the configuration was Binomial (i.e. a totally random configuration of non-zero coefficients). The reason for thinking that the discretized area-interaction process would make a better prior is that the wavelet transform provides time-frequency localisation. This means that the effect of, for example, a discontinuity in the signal or in one of the first few derivatives of the signal will produce significant coefficients of the wavelet transform of the signal only in the coefficients close to the location at which the discontinuity occurs. This fact means that the wavelet transform will have most of its coefficients clustered around a few locations, thus leading to a clustered rather than uniformly random distribution of coefficients. This can be seen clearly in Figure 1, which shows the discrete wavelet transform of several common test functions represented in the natural binary tree configuration.

More formally, we begin by allowing for the presence of noise by assuming that the true wavelet coefficients are corrupted by Gaussian noise with zero mean and some variance σ^2. This gives the following likelihood:

$$\widehat{d_{jk}}|d_{jk} \sim N(d_{jk}, \sigma^2),$$

where $\widehat{d_{jk}}$ is the value of the noisy wavelet coefficient (the data) and d_{jk} is the value of the true coefficient. We then place a prior on the value of the wavelet coefficients:

$$d_{jk}|\mathbf{J} \sim N(0, \tau^2 J_{jk}), \tag{6}$$

where τ^2 is a constant and J_{jk} is the number of points at location (j, k) of a certain lattice process J which exists on the natural binary tree commonly used to represent wavelet coefficients. Thus the more points at a given location, the larger the variance of the prior on d_{jk}, resulting in a higher probability of large values of d_{jk}. Finally, we place a hyperprior on this lattice process:

$$P(\mathbf{J}) = \alpha\lambda^{N(\mathbf{J})}\gamma^{-m(U(\mathbf{J}))} \tag{7}$$

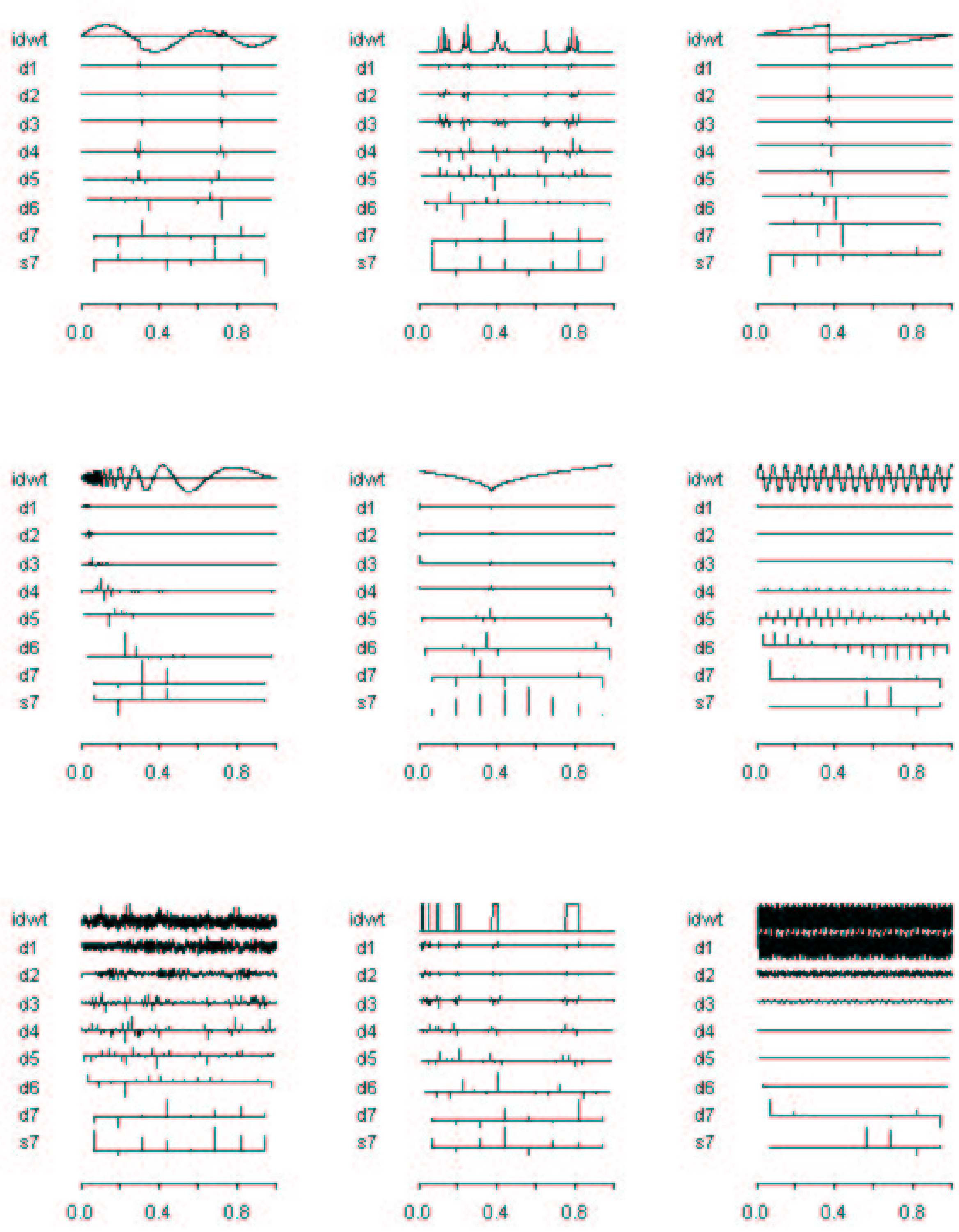

Figure 1: Examples of the discrete wavelet transform of some test functions. There is clear evidence of clustering in most of the graphs. The original functions are shown above their discrete wavelet transform each time.

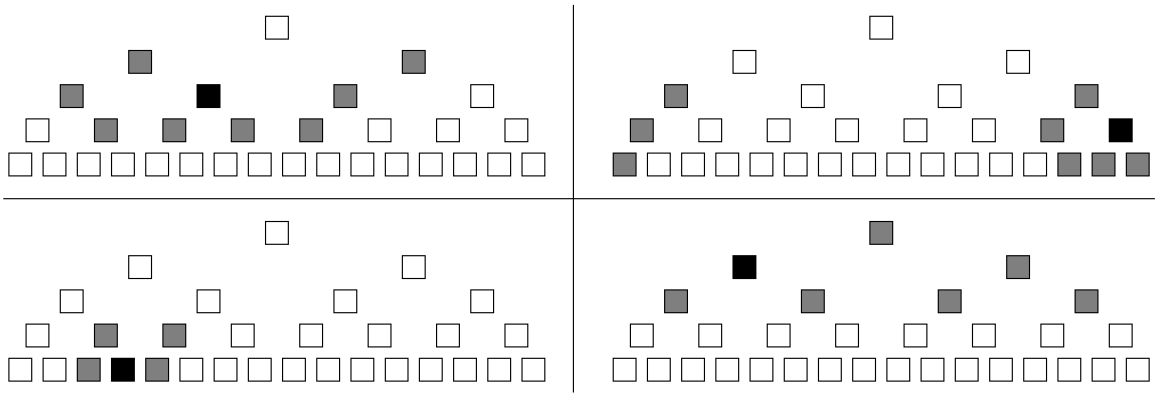

Figure 2: The four plots give examples of what we used as $Z(\cdot)$ for four different example locations showing how we dealt with boundaries. Grey boxes are $Z(x) \setminus \{x\}$ for each example location x, while x itself is shown as black.

with respect to the unit rate independent auto-Poisson process (Cressie 1993), where $\mathbf{J} = (J_{jk})$ is the configuration. If we take a value of γ greater than one this gives a clustered configuration. Thus we would expect to see clusters of large values of d_{jk} if this were a reasonable model — which is exactly what we do see in Figure 1.

This is an extension of the model of Abramovich et al. (1998), who assume that the true wavelet coefficients are distributed as a mixture of a Normal distribution with zero mean and variance dependent on the level of the coefficient, and a point mass at zero as follows:

$$d_{jk} \sim \zeta_{jk} N(0, \tau_j^2) + (1 - \zeta_{jk})\delta(0),$$

where d_{jk} is the value of the kth coefficient at level j of the discrete wavelet transform and τ_j is a positive constant. Notice that (6) includes a point mass at zero when $J_{jk} = 0$ (i.e. when there are no points alive at that location). Abramovich et al. (1998) also assume that there is $N(0, \sigma^2)$ noise added to the true coefficients. This is equivalent to our likelihood $\widehat{d_{jk}} | d_{jk} \sim N(d_{jk}, \sigma^2)$.

Clearly a suitable interpretation of $U(\mathbf{J}) = \bigcup_{(j,k)} Z(J_{jk})$ in equation (7) is required. Organising the wavelet coefficients into the traditional binary tree layout, one possibility would be to use the parent, children and immediate sibling and cousin of a coefficient as $Z(x)$. Another would be to use a variation on this taking into account the length of support of the wavelet used. Figure 2 shows the scheme we used, which we feel captures the localisation of both time and frequency effects well.We decided to use the parent, the coefficient on the parent's level of the transform which is next-nearest to x, the two adjacent coefficients on the level of x, the two children and the coefficients adjacent to them, making a total of nine coefficients (including x itself). Figure 2 also shows how we dealt with boundaries: we assume that the signal we are examining is periodic, making it natural to have periodic boundary conditions horizontally. If $Z(x)$ overlaps with a vertical boundary we simply discard those parts which have no locations associated with them. The simple counting measure used has $m(K(x)) = 9$ unless x is in the bottom row or one of the top two rows.

The log posterior for our model is

$$\log f(\mathbf{J}|\widehat{\mathbf{d}}) = \log \alpha + N(\mathbf{J}) \log \lambda - m(U(\mathbf{J})) \log \gamma + \sum_{i=0}^{\infty} \sum_{J_{jk}=i} \log f_i(\widehat{d}_{jk}), \tag{8}$$

where $f_i(x)$ is the Gaussian density with zero mean and variance $i\tau^2$.

Clearly this is not an ordinary area-interaction process. We now describe an application of the method of Ambler (2002) which allows us to sample from this density.

3 Exact Sampling from the posterior

Although the expression in equation (8) may look like a rather complicated density it turns out that the powerful method of it can be simulated perfectly using a rather simple extension of the procedure for simulating the area-interaction process.

We make use of a spatial birth-death process like those introduced by Preston (1976) and used to sample exactly from spatial point processes in Section 1.3. Since equation (8) is a product of positive bounded functions from the configuration space into $\mathbb{R}$ the methods introduced in Section 1.3 may be applied. We modify the notation slightly in this section, using $\mathbf{J}^{max}$ and $\mathbf{J}^{min}$ to refer to the upper and lower processes respectively. These were called U and L in Section 1.3. We then use J_x^{max} and J_x^{min} to refer to the maximum and minimum processes at location x. Taking advantage of the lattice structure, the dominating process D which we use actually has a different rate at each location, rather than being constant as described in Section 1.3. See Ambler (2002) Chapter 5 for some other examples of this kind of dominating process. The rate is given by

$$\lambda_{jk}^{dom} = \lambda e^{\widehat{d}_{jk}^2 \tau^2 / 2\sigma^2(\tau^2+\sigma^2)} \tag{9}$$

at each location (j, k) on the lattice. The lower process is then started as a thinned version of D. Points are accepted with probability

$$P(x) = \gamma^{-M(\chi)} \left(\frac{\sigma^2}{\tau^2 + \sigma^2} \right)^{1/2} \times \exp\left(- \frac{\widehat{d}_x^2 \tau^2}{2\sigma^2(\tau^2+\sigma^2)} \right),$$

where $M(\chi) = \max_\chi(m(K(x)))$.

The upper and lower processes are then evolved through time, accepting points as described in Section 1.3 with probability

$$\gamma^{-m(K(x)\backslash U(\mathbf{J}^{max}))} \left(\frac{\tau^2 J_x^{max} + \sigma^2}{\tau^2(J_x^{max}+1)+\sigma^2} \right)^{1/2}$$
$$\times \exp\left(- \frac{\widehat{d}_x^2 \tau^2}{2} \frac{\tau^2 J_x^{min}(\tau^2(J_x^{min}+1)+2\sigma^2)}{\sigma^2(\tau^2+\sigma^2)(\tau^2 J_x^{min}+\sigma^2)(\tau^2(J_x^{min}+1)+\sigma^2)} \right)$$

for the upper process and

$$\gamma^{-m(K(x)\backslash U(\mathbf{J}^{max}))} \left(\frac{\tau^2 J_x^{min} + \sigma^2}{\tau^2(J_x^{min}+1)+\sigma^2} \right)^{1/2}$$
$$\times \exp\left(- \frac{\widehat{d}_x^2 \tau^2}{2} \frac{\tau^2 J_x^{max}(\tau^2(J_x^{max}+1)+2\sigma^2)}{\sigma^2(\tau^2+\sigma^2)(\tau^2 J_x^{max}+\sigma^2)(\tau^2(J_x^{max}+1)+\sigma^2)} \right)$$

for the lower process. The remainder of the algorithm carries over in the obvious way.

3.1 Using the Generated Samples

Although $\mathbf{d}$ was integrated out for simulation reasons in Section 2 it is, naturally, the quantity of interest. Having simulated realisations of $\mathbf{J}|\widehat{\mathbf{d}}$ we then generate $\mathbf{d}|\mathbf{J},\widehat{\mathbf{d}}$ for each realisation of $\mathbf{J}$ generated in the first step. The sample median of $\mathbf{d}|\mathbf{J},\widehat{\mathbf{d}}$ gives an estimate for $\mathbf{d}$, as required. The median is used instead of the mean as this gives a thresholding rule (defined by Abramovich et al. (1998) as a rule giving $p(d_{jk}|\widehat{\mathbf{d}}) > 0$).

We calculate $f(\mathbf{d}|\mathbf{J},\widehat{\mathbf{d}})$ using logarithms for ease of notation. Assuming that $J_{jk} \neq 0$ we find

$$\begin{aligned}
\log f(d_{jk}|\widehat{d}_{jk}, J_{jk} \neq 0) &= \log f(d_{jk}) + \log f(\widehat{d}_{jk}|d_{jk}, (j,k) \in \mathbf{J}) + C \\
&= \frac{-d_{jk}^2}{2\tau^2 J_{jk}} + \frac{-(\widehat{d}_{jk} - d_{jk})^2}{2\sigma^2} + C_1 \\
&= -\frac{(\sigma^2 + \tau^2 J_{jk})\left(d_{jk} - \frac{\tau^2 J_{jk}\widehat{d}_{jk}}{\sigma^2 + \tau^2 J_{jk}}\right)^2}{2\sigma^2\tau^2 J_{jk}} + C_2
\end{aligned}$$

where C, C_1 and C_2 are constants. Thus

$$f(d_{jk}|\widehat{d}_{jk}, J_{jk} \neq 0) \sim N\left(\frac{\tau^2 J_{jk}\widehat{d}_{jk}}{\sigma^2 + \tau^2 J_{jk}}, \frac{\sigma^2\tau^2 J_{jk}}{\sigma^2 + \tau^2 J_{jk}}\right).$$

When $J_{jk} = 0$ we clearly have $f(d_{jk}|J_{jk}, \widehat{d}_{jk}) = 0$.

4 Simulation Study

We now present a more careful simulation study of the performance of our estimator relative to several established Wavelet-based estimators. Similar to the study of Abramovich et al. (1998), we investigate the performance of our method on the four standard test functions of Donoho and Johnstone (1994, 1995), namely “Blocks”, “Bumps”, “Doppler” and “Heavisine”. These test functions are used because they exhibit different kinds behaviour typical of signals arising in a variety of applications.

The test functions were simulated at 256 points equally spaced on the unit interval. The test signals were centred and scaled so as to have mean value 0 and standard deviation 1. We then added independent $N(0, \sigma^2)$ noise to each of the functions, where σ was taken as 1/10, 1/7 and 1/3. The noise levels then correspond to root signal-to-noise ratios (RSNR) of 10, 7 and 3 respectively. We performed 25 replications. For our method, we simulated 25 independent draws from the posterior distribution of the d_{jk}’s and used the sample median as our estimate, as this gives a thresholding rule. For each of the runs, σ was set to the standard deviation of the noise we added, τ was set to 1.0, λ was set to 0.05 and γ was set to 3.0.

The values of parameters σ and τ were set to the true values of the standard deviation of the noise and the signal, respectively. In practice it will be necessary to develop some method for estimating these values. The value of λ was chosen to be 0.05 because it was felt that not many of the coefficients would be significant. The value of γ was chosen based on small trials for the heavisine and jumpsine datasets (not shown).

Method	RSNR	AMSEs for the following test functions:			
		Blocks	Bumps	Doppler	Heavisine
LatticeBayesThresh	10	0.0025	0.0084	0.0049	0.0032
	7	0.0056	0.0185	0.0087	0.0052
	3	0.0534	0.1023	0.0448	0.0149
BayesThresh	10	0.0344	0.1651	0.0167	0.0035
	7	0.0414	0.1716	0.0225	0.0057
	3	0.0860	0.2015	0.0448	0.0140
Cross-validation	10	0.0055	0.0392	0.0112	0.0030
	7	0.0096	0.0441	0.0135	0.0054
	3	0.0452	0.0914	0.0375	0.0057
SureShrink	10	0.0049	0.0131	0.0054	0.0065
	7	0.0098	0.0253	0.0099	0.0093
	3	0.0482	0.0973	0.0399	0.0147
False discovery rate	10	0.0159	0.0449	0.0144	0.0064
	7	0.0294	0.0758	0.0253	0.0093
	3	0.1230	0.2324	0.0861	0.0148

Table 1: Average mean-square errors for our estimator (labelled LatticeBayesThresh), ordinary BayesThresh, cross-validation, SureShrink and false discovery rate estimators for four test functions for two values of the root signal-to-noise ratio. Averages are based on 25 replicates.

We compare our method with several established wavelet-based estimators for reconstructing noisy signals: ordinary BayesThresh (Abramovich et al. 1998), SureShrink (Donoho and Johnstone 1994), cross-validation (Nason 1996) and the false discovery rate (Abramovich and Benjamini 1996). For test signals "Bumps", "Doppler" and "Heavisine" we used Daubechies least asymmetric wavelet of order 10 (Daubechies 1992). For "Blocks" we used the Haar wavelet, as the original signal was piecewise constant. The analysis was carried out using the freely available *R* statistical package. The WaveThresh package (Nason 1993) was used to perform the discrete wavelet transform and also to compute the BayesThresh, SureShrink, cross-validation and false discovery rate estimators.

The goodness of fit of each estimator was measured by its average mean-square error (AMSE) over the 25 replications. Table 1 presents the results. It is clear that our estimator performs extremely well with respect to the other estimators when the signal-to-noise ratio is large, but struggles when there is a small signal-to-noise ratio. This may be due to the fact that it was necessary to make some approximations in constructing the sampler. These are discussed in the appendix.

5 Conclusions and future work

We have introduced a procedure for Bayesian wavelet thresholding which uses the naturally clustered nature of the wavelet transform when deciding how much weight to give coefficient values. We have demonstrated that this procedure performs well compared to existing methods, though the implementation suffered from some problems which made exact computation infeasible. The performance seems good for moderate and low noise levels, though it was a little disappointing for higher noise levels.

One possible area for future work would be to replace equation (6) with

$$d_{jk}|\mathbf{J} \sim N(0, \tau^2 (J_{jk})^z),$$

where z would be a further parameter. This would modify the number of points which are likely to be alive at any given location and thus also modify the tail behaviour of the prior. The idea behind this suggestion is that when we know that the behaviour of the data is either heavy or light tailed, we could adjust z to compensate. This could possibly also help speed up convergence by reducing the number of points at locations with large values of d_{jk}. As inclusion of this extra parameter requires only minor modifications, the implementation discussed actually includes this option. The results presented in Section 4 were generated by simply setting $z = 1$.

A second possible area for future work would be to develop some automatic methods for choosing the parameter values, perhaps using the method of maximum pseudo-likelihood (Besag 1974; Besag 1975; Besag 1977).

Software implementing the work described in this paper is available on request from the first author.

References

Abramovich, F. and Y. Benjamini (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis 22*, 351–361.

Abramovich, F., T. Sapatinas, and B. W. Silverman (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B 60*, 725–749.

Ambler, G. K. (2002). *Dominated Coupling from the Past and Some Extensions of the Area-Interaction Process*. Ph. D. thesis, Department of Mathematics, University of Bristol.

Baddeley, A. J. and M. N. M. van Lieshout (1995). Area-interaction point processes. *Annals of the Institute for Statistical Mathematics 47*, 601–619.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B 57*, 289–300.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician 24*, 179–195.

Besag, J. (1977). Some methods of statistical analysis for spatial data. *Bulletin of the International Statistical Institute 47*, 77–92.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Philadelphia, Pennsylvania: SIAM.

Donoho, D. L. and I. M. Johnstone (1994). Ideal spatial adaption by wavelet shrinkage. *Biometrika 81*, 425–455.

Donoho, D. L. and I. M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*, 1200–1224.

Kelly, F. P. and B. D. Ripley (1976). A note on Strauss's model for clustering. *Biometrika 63*, 357–360.

Kendall, W. S. (1998). Perfect simulation for the area-interaction point process. In L. Accardi and C. C. Heyde (Eds.), *Probability Towards 2000*, pp. 218–234. Springer.

Lindvall, T. (1992). *Lectures on the Coupling Method.* New York: John Wiley & Sons.

Nason, G. P. (1993). The `WaveThresh` package: wavelet transform and thresholding software for s-plus and r. Available from Statlib.

Nason, G. P. (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society, Series B 58*, 463–479.

Preston, C. J. (1976). Spatial birth-and-death processes. *Bulletin of the Institute of International Statistics 46*, 371–391.

Propp, J. G. and D. B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms 9*, 223–252.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics 9*, 1135–1151.

Strauss, D. J. (1975). A model for clustering. *Biometrika 62*, 467–475.

A Dealing with large and small rates

When attempting to implement the algorithm in Section 3 we encountered problems due to extremely high birth rates in the dominating process. Recall from Equation 9 that if the maximum data value d_{jk} is twenty times larger in magnitude than the standard deviation of the noise (a not uncommon event for reasonable noise levels) then we have

$$\begin{aligned}\lambda_{dom} &= \lambda e^{400\sigma^2\tau^2/2\sigma^2(\tau^2+\sigma^2)} \\ &= \lambda e^{200\tau^2/(\tau^2+\sigma^2)}.\end{aligned}$$

Now unless τ is significantly smaller than σ, this will result in enormous birth rates. We are clearly not going to be able to simulate this efficiently.

To get around this problem we reasoned that the chances of there being no live points at a location whose data value is large (resulting in a value of λ_{dom} larger than e^4) is sufficiently small that for the purposes of calculating $m((x \oplus G) \setminus (Y(-M,u) \oplus G))$ for nearby locations it could be assumed that the number of points alive was strictly positive. This allows us to simulate the process accurately for the locations of interest and provide a reasonable level of discrimination in a more reasonable time frame.

Unfortunately, the problems do not stop there. Recall from Section 3.1 that

$$d_{jk} | J_{jk}, \widehat{d}_{jk} \sim N\left(\frac{\tau^2 J_{jk}\widehat{d}_{jk}}{\sigma^2 + \tau^2 J_{jk}}, \frac{\sigma^2\tau^2 J_{jk}}{\sigma^2 + \tau^2 J_{jk}}\right)$$

so that we need values of J_{jk} for each location (j,k) in the configuration. Unfortunately, we no longer know the value of J_{jk} for those locations which have large values of d_{jk}.

To get around this problem we first notice that

$$\frac{\tau^2 J_{jk}\widehat{d}_{jk}}{\sigma^2 + \tau^2 J_{jk}} \xrightarrow[J_{jk}\to\infty]{} \widehat{d}_{jk}$$

monotonically from below, and that

$$\frac{\tau^2 J_{jk}\sigma^2}{\sigma^2+\tau^2 J_{jk}} \xrightarrow[J_{jk}\to\infty]{} \sigma^2,$$

also monotonically from below. Since σ is typically small, convergence is very fast indeed. Taking $\tau = \sigma$ as an example we see that even when $J_{jk} = 5$ we have

$$\frac{\tau^2 J_{jk}\widehat{d}_{jk}}{\sigma^2+\tau^2 J_{jk}} = \frac{5}{6}\widehat{d}_{jk}$$

and

$$\frac{\tau^2 J_{jk}\sigma^2}{\sigma^2+\tau^2 J_{jk}} = \frac{5}{6}\sigma^2.$$

We see that we are already within $\frac{1}{6}$ of the limit. Convergence is even faster for larger values of τ.

We also recall that the dominating process gives an upper bound for the value of J_{jk} at every location. Thus a good estimate for d_{jk} would be gained by taking the value of J_{jk} in the dominating process for those points where we do not know the exact value. This is a good solution but is unnecessary in some cases, as sometimes the value of λ_{dom} is so large that there is little advantage in using this value. Thus for exceptionally large values of λ_{dom} we simply use $N(\widehat{d}_{jk}, \sigma^2)$ numbers as our estimate of d_{jk}.