
Article:
Aykroyd, RG orcid.org/0000-0003-3700-0816 and Aljohani, H (2020) A Bayesian approach to wavelet-based modelling of discontinuous functions applied to inverse problems. Communications in Statistics - Simulation and Computation, 49 (1). pp. 207-225. ISSN 0361-0918

https://doi.org/10.1080/03610918.2018.1484473

A Bayesian approach to wavelet-based modelling of discontinuous functions applied to inverse problems

Robert G. Aykroyd[1*] and Hassan Aljohani[2†]

[1]University of Leeds, UK and [2]Taif University, Saudi Arabia

Abstract

Inverse problems are examples of regression with more unknowns than the amount of information in the data and hence constraints are imposed through prior information. The proposed method defines the underlying function as a wavelet approximation which is related to the data through a convolution. The wavelets provide a sparse and multi-resolution solution which can capture local behaviour in an adaptive way. Varied prior models are considered along with level-specific prior parameter estimation. Archaeological stratigraphy data are considered where vertical earth cores are analysed producing clear piecewise constant function estimates.

Keywords: Archaeological stratigraphy, Elastic-net, Haar wavelet, Hierarchical models, Laplace distribution, Markov chain Monte Carlo, Sparsity.

1 Introduction

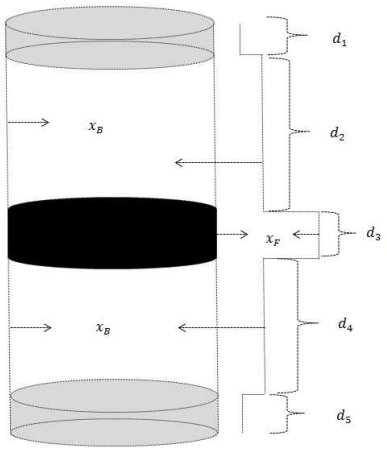
Many scientific investigations involve recording measurements which are only indirectly related to the quantity of interest. In some cases these will involve only a simple linear relationship leading to a calibration problem, but in more complicated cases this can be a convolution which necessitates a more challenging analysis. Because of the nature of many such inverse problems, deconvolution can introduce extra numerical issues and hence require a more thoughtful approach. Although a particular application is investigated here, the methodology provides a general framework for many other inverse problems in the applied sciences — a general introduction to inverse problems can be found in Ribés and Schmitt (2008) with a mathematical review in Stuart (2010).

The application of inverse estimation methods to be considered here is the analysis of core magnetic readings in geophysics and in particular archaeology — though it is very similar to applications in oil exploration. To investigate the earth's subsurface narrow

*Address for correspondence: Robert G. Aykroyd, Department of Statistics, University of Leeds, LS2 9JT, UK. E-mail: R.G.Aykroyd@leeds.ac.uk

†This work was carried out while at the University of Leeds, Leeds, UK.

cores are extracted and examined. In some cases analysis may only involve observation of colour, for example volcanic ash or charcoal from fires, but in other cases there is no visible evidence. In some cases there may be evidence based on magnetic properties, and a magnetometer passed along the core might reveal changes. The key quantity is then the magnetic susceptibility which measures the degree of magnetization of a material when placed in a magnetic field — a discussion of magnetic properties of material can be found in, for example, Le Borgne (1960). Figure 1 shows a diagram of a core with various key dimensions marked along with typical parameter values – these are values used in Aykroyd and Al-Gezeri (2014) in simulations to mimic real data. When the core is passed through the magnetic detector, data recording starts before the core enters the equipment and continues until after it is completely removed from the other side, hence the distances d_1 and d_5 correspond to an empty detector with susceptibility zero. As the core enters the detector the topsoil from the site is recorded with susceptibility x_B which represents a background susceptibility for a distance d_2. As the core passes through, the second part of the core, of length d_3 represents the archaeological feature, with susceptibility profile x_F. There is a second background part which is of length d_4 and again has susceptibility x_B. Finally, d_5 represents the last distance after the core has emerged, and has zero susceptibility before the data recording stops. Hence the first three cores have small extent whilst the fourth and fifth cores have large extent.



		Core				
		I	II	III	IV	V
Distances	d_1	80	80	80	80	80
	d_2	30	30	30	30	30
	d_3	70	70	80	160	150
	d_4	140	140	130	50	60
	d_5	80	80	80	80	80
	x_F	1.1	1.5	1.8	1.1	1.0
	x_B	0.2	0.2	0.2	0.2	0.2

Figure 1: Diagram of an extracted core and corresponding susceptibility profile along with typical parameter values derived from a real archaeological site.

In archaeology the equipment can be used to take measurements at fine resolution along such cores, but this means that nearby readings are necessarily highly correlated. To remove, or at least reduce, these effects a deconvolution step is needed, but in the presence of substantial noise simple approaches would not yield reliable results. The first methods to be applied were the, Fourier transform based, Wiener Filter and singular

value decomposition algorithms (see, for example, Press and Flannery, 1992), but the algorithms can be unstable, with smooth reconstructions and "Gibbs ringing" artifacts (Bracewell, 1986; Champeney, 1987). Statistical approaches to deconvolution problems began in the early 1980s (see the review paper by Besag, 1989, and references therein). Such approaches involved modelling local variability, in terms of first differences, using Gaussian and Laplace distributions which also introduce local smoothing. This works well for smoothly varying continuous functions but it is not helpful here as the scientist wishes to divide the core into distinct segments representing separate archaeological periods. In these situations the ideal answer would be a combination of constant values, representing stable conditions, with jumps which mark abrupt changes. Wavelet methods are often used elsewhere to describe functions which are smooth in parts but with significant discontinuities. The theme of this paper, therefore, is the use of wavelet methods for inverse problems where the aim is to produce piecewise constant output from noisy data produced through a convolution.

The paper is organised as follows: Section 2 gives background to wavelets and inverse problems, and defines notation to be used later. Section 3 describes the problem to be solved. Section 4 described the model within a Bayesian framework, with implementation issues discussed in Section 5. The analysis of real archaeological stratigraphy data is shown in Section 6, with overall conclusions in Section 7.

2 Background to wavelets and inverse problems

This section contains some of the basic ideas of wavelets and inverse problems with the reader directed to the following references for more information. In later sections, however, further details will be included if needed. For wavelets see for example Daubechies (1988); Donoho and Johnstone (1994); Nason (2008) and for inverse problems see for example Ribés and Schmitt (2008); Stuart (2010); Aykroyd (2015) with some related applications in Aykroyd et al. (2001); Aykroyd and Al-Gezeri (2014), and some work combining wavelets and inverse problems in Donoho (1995) and Abramovich and Silverman (1998).

Consider the problem where $\mathbf{f} = \{f_i : i = 1, \ldots, n\}$ is a vector of values of some unknown function at a set of n equally-spaced locations, and that $\mathbf{y} = \{y_i : i = 1, \ldots, n\}$ are observed data values recorded at the same locations. Further, it is assumed that $\mathbf{y}$ and $\mathbf{f}$ are related by

$$\mathbf{y} = \mathbf{f} + \epsilon,$$

where ϵ is a vector of random variables such that $\epsilon \sim N(0, \sigma^2 I_n)$. The aim is to find an estimate of $\mathbf{f}$ given the data $\mathbf{y}$.

Wavelets are a common choice for this type of non-parametric regression problem when

noise removal or a multi-resolution analysis is required. Let W be an orthogonal matrix holding an appropriate discrete wavelet basis, such as the Haar wavelet, as used here, or one from the general Daubechies families (Daubechies, 1988). The wavelet decomposition of the data $\mathbf{y}$ can be written as

$$\mathbf{d_y} = W\mathbf{y} = W(\mathbf{f} + \epsilon) = W\mathbf{f} + W\epsilon = \mathbf{d_f} + \eta$$

where $\mathbf{d_y}$ and $\mathbf{d_f}$ are vectors of the wavelet coefficients of $\mathbf{y}$ and $\mathbf{f}$ respectively. Also, by the orthogonality of W it can be seen that the errors $\eta \sim N(0, \sigma^2 I_n)$. This shows that noise in the measurements results in corresponding noise in the wavelet coefficients.

It is a common approach to say that fine level coefficients are the result of noise with the signal being represented in a small number of low-level coefficient values. The method of wavelet thresholding can then be used to set the small coefficients to zero, or shrinkage to move the coefficient values closer to zero. A set of modified coefficient values $\mathbf{d}^*_\mathbf{y}$ after thresholding or shrinkage can be used as an estimate of the wavelet coefficients of $\mathbf{f}$, that is $\widehat{\mathbf{d}_\mathbf{f}} = \mathbf{d}^*_\mathbf{y}$, with resulting estimate of $\mathbf{f}$ defined as

$$\widehat{\mathbf{f}} = W^T \mathbf{d}^*_\mathbf{y}.$$

This denoising method can also be given an interpretation in a Bayesian setting which is the approach followed later.

The overall aim in a general linear inverse problem is also to estimate an unknown function, $\mathbf{f}$, from a finite set of measurements, $\mathbf{y}$, but these quantities are related through some convolution equation such as

$$\mathbf{y} = H\,\mathbf{f} + \boldsymbol{\epsilon}$$

where H is a given matrix and $\boldsymbol{\epsilon}$ is some error vector. Following a regression approach an estimate of $\mathbf{f}$ might be found using the usual least-squares estimate

$$\widehat{\mathbf{f}}_{LS} = (H^T H)^{-1} H^T \mathbf{y}.$$

If, however, H is not of full rank, then the inversion cannot to completed. In fact, this is a characteristic of all ill-posed inverse problems where the Hadamard conditions state that a problem is well-posed if: a solution exists, the solution is unique and the solution changes continuously with the data

3 Data modelling

In archaeology is it now required to investigate a potential site using geophysical remote sensing methods before any physical excavation is started. One possible technique is

archaeological stratigraphy, where a narrow soil core is removed and examined for possible human activity. A core sample is obtained using a soil borer and although the strata in the core often show no variation in colour or texture, an analysis of the magnetic susceptibility can differentiate between the separate layers. Once collected, the core is passed through a detector coil, allowing readings of the susceptibility to be made along the length of the sample. The datasets, which will be analysed later, are shown in Figure 2 and, although the approximate positions of high activity regions can be identified, it would not be possible to clearly identify distinct segments.

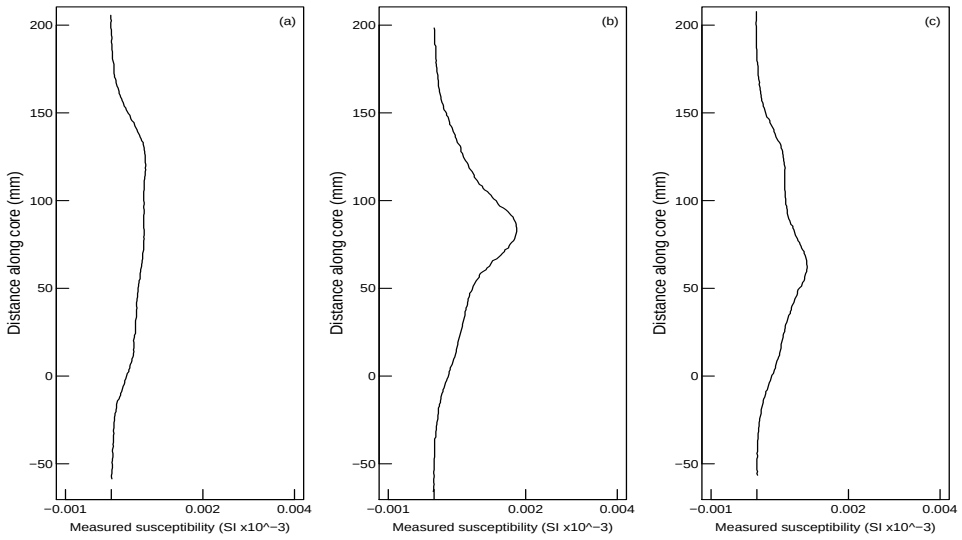


Figure 2: Measured magnetic data: (a) pyre core I, (b) pyre core II and (b) pyre core III.

Let the output readings be denoted $\mathbf{y} = \{y_i : i = 1, \ldots, n\}$ which are recorded at equally-spaced points along the core length. For estimation consider the core partitioned into elements along its length, with susceptibilities denoted by $\mathbf{f} = \{f_i : i = 1, \ldots, n\}$. Since the detector coil is sensitive to the susceptibility across an extended section of the sample, the reading indicates not the value at a sharply defined point, but a weighted average of the values over an extended range, hence

$$\mathbf{y} = H\mathbf{f} + \boldsymbol{\epsilon}$$

where the appropriate form of the spread, or transfer, function, H, is given in (Allum et al., 1999). The form of the above equation makes this a linear inverse problem.

In practice, the observed measurements are subject to error from various sources. It has been verified in calibration experiments (Allum et al., 1999) that a Gaussian error model, with zero mean and constant variance, accounts satisfactorily for the apparent errors. Further it is possible to use similar calibration experiments to get a good estimate of the noise variance if needed. Hence assuming an additive Gaussian error model the

conditional distribution of the data given the truth is

$$\mathbf{y}|\mathbf{f} \sim \mathtt{N}(H\mathbf{f}, \sigma^2 I_n)$$

with likelihood

$$\pi(\mathbf{y}|\mathbf{f}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - [H\mathbf{f}]_i)^2\right\}, \quad \mathbf{y}, \mathbf{f} \in \mathbb{R}^n; \sigma^2 > 0,$$

where $[H\mathbf{f}]_i$ denotes the ith element of the vector obtain after the product $H\mathbf{f}$ is calculated. It is not possible to reliably estimate $\mathbf{f}$ using the likelihood alone and so previous approaches have used Bayesian modelling with smoothing prior distributions directly on the unknown function. For example, a Gaussian process defined by, improper, density

$$\pi(\mathbf{f}) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{n-1}(f_{i+1} - f_i)^2\right\}, \quad \tau^2 > 0.$$

For application in archaeological stratigraphy see, for example, Allum et al. (1999). By its very nature, this type of prior description leads to smooth reconstructions which are in contradiction of the original aim to produce an estimate which retains the piecewise constant appearance required. Even carefully designed prior models only partially achieve this aim and hence an alternative approach, such as wavelet methods, is needed.

4 Bayesian modelling

4.1 General

The key ingredients in the Bayesian approach are the likelihood function and prior distribution, and hence the resulting posterior distribution. The likelihood is the conditional distribution of the data given the unknowns, denoted as $\pi(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of model parameters. The prior distribution, denoted $\pi(\boldsymbol{\theta})$, quantifies detailed expert knowledge or general beliefs about the unknowns–the choice of the exact form of this distribution is more subjective than is the choice of likelihood.

For estimation, evidence from the data and from prior beliefs are brought together by combining the likelihood and prior distribution, using Bayes's Theorem, to form the posterior distribution, defined as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})}$$

where $\pi(\mathbf{y})$ is a normalizing constant. Note that since this usually involves a high dimensional integral it will be unacceptably time-consuming to perform the calculation.

Fortunately, the normalising constant contains no information about the unknowns and hence can be dropped, giving the key statement

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

– that is "posterior" is proportional to "likelihood" times "prior". This distribution incorporates evidence from the data and knowledge from the prior distribution allowing an estimation process which balances the two types of information.

When there are multiple groups of parameters and prior parameters they will be assumed independent and modelled separately. Hence, if $\boldsymbol{\theta}$ is made-up of two sub-sets, say, with $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, then the above equation simply becomes

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2)$$

with other definitions and results following in an obvious fashion.

In the Bayesian setting, the posterior distribution is the basis for estimation and hence a point estimate can be found, for example, using the maximum a posteriori (MAP) estimate, the posterior median, or, as here, the posterior mean. In addition, the joint posterior distribution can be examined, for example, to construct marginal posterior distributions, or to calculate Bayesian credible intervals–some examples will be shown as part of the data analysis in Section 6.

4.2 Likelihood using wavelet coefficients

In stating the aim to express the unknown function, $\mathbf{f} = \{f_i : i = 1, \ldots, n\}$, in terms of wavelets, the estimation problem has changed to the need to estimate the wavelet coefficients, $\mathbf{d_f}$, of $\mathbf{f}$. Recalling that given $\mathbf{d_f}$ and W, then $\mathbf{f}$ can easily be recreated as

$$\mathbf{f} = W^T\mathbf{d_f}$$

and hence the corresponding form of the likelihood is

$$\pi(\mathbf{y}|\mathbf{d_f}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - [H\ W^T\mathbf{d_f}]_i\right)^2\right\}, \quad \mathbf{y}, \mathbf{d_f} \in \mathbb{R}^n; \sigma^2 > 0$$

where $\left[H\ W^T\mathbf{d_f}\right]_i$ denotes the ith element of the vector obtain after the product $H\ W^T\mathbf{d_f}$ is calculated. It is common to estimate the noise variance, σ^2, based on the finest level wavelet coefficients of the data (see, for example, Nason, 2010) as they are likely to only contain noise – this is the approach used in Section 6.

4.3 A single prior model for wavelet coefficients

Now attention falls on appropriate choices for the prior distributions, $\pi(\mathbf{d_f})$, on the wavelet coefficients, $\mathbf{d_f}$, or on a subset of the coefficients. An obvious choice is the Gaussian distribution with density

$$\pi(\mathbf{d_f}|\kappa) = \left(\frac{\kappa}{\pi}\right)^{n/2} \exp\left\{-\kappa \sum_{j=0}^{J} d_j^2\right\}, \quad \mathbf{d_f} \in \mathbb{R}^n;\ \kappa > 0$$

where $J = \log_2(n)$. Gaussian distributions have been found lacking as priors for wavelet coefficients across a wide range of problems (Johnstone and Silverman, 2005). Further discussion of prior distributions choice can be found in Berger and Pericchi (2001). For sparsity, however, the Laplace distribution might be a better choice with density function

$$\pi(\mathbf{d_f}|\kappa) = \left(\frac{\kappa}{2}\right)^{n} \exp\left\{-\kappa \sum_{j=0}^{J} |d_j|\right\}, \quad \mathbf{d_f} \in \mathbb{R}^n;\ \kappa > 0.$$

As a compromise between these two, a density based on the elastic-net function (Hastie et al., 2009) might be suitable. The corresponding density function can be shown to be given by

$$\pi(\mathbf{d_f}|\kappa,\gamma) = \frac{1}{(Z(\kappa,\gamma))^n} \exp\left\{-\kappa(\ \gamma \sum_{j=0}^{J} d_j^2 + (1-\gamma) \sum_{j=0}^{J} |d_j|\right\},$$
$$\mathbf{d_f} \in \mathbb{R}^n; \kappa > 0, 0 \le \gamma \le 1,$$

where the normalizing constant, derived in the Appendix, is given by

$$Z(\kappa,\gamma) = \begin{cases} 2/\kappa & \gamma = 0, \\ \frac{2\sqrt{\pi}}{\sqrt{\kappa\gamma}} \exp\left(\frac{\kappa(1-\gamma)^2}{4\gamma}\right)\left(1 - \Phi\left(\frac{\kappa(1-\gamma)}{\sqrt{2\kappa\gamma}}\right)\right), & 0 < \gamma < 1, \\ \sqrt{\pi/\kappa} & \gamma = 1. \end{cases}$$

Clearly, for the limit values of γ, this reduces to the Gaussian case ($\gamma = 1$) and the Laplace case ($\gamma = 0$).

Note that each of these distributions has introduced additional parameters, κ and γ, which will also be modelled. given that γ can only take values within the range $[0, 1]$, the beta distribution is a sensible choice of prior model, $\gamma \sim \texttt{Beta}(a, b)$, with density

$$\pi(\gamma) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \gamma^{a-1}(1-\gamma)^{b-1}, \quad 0 \le \gamma \le 1; a, b > 0.$$

The parameters, a and b, can be fixed based on knowledge or information from separate calibration experiments. In particular, an expert might provide a mean and variance leading to corresponding values for a and b. Here, however, values $a = b = 1$ have

been used which correspond to a uniform prior recognising that there is no true prior preference. Finally, to favour large values of κ and hence promote sparsity and shrinkage in the wavelet coefficients, the following distribution with, improper, density is used

$$\pi(\kappa) \propto \exp\bigg(- c^2/\kappa \bigg), \quad \kappa > 0; c > 0.$$

Although, there may be external information, it is more likely that the value of the hyper-parameter would be fixed after initial experiments, and here $c = 1000$ has been used.

Before moving on, at the start of this section it was mentioned that the prior on the wavelets coefficients might only act on a subset of the coefficients. In particular, the coarsest level might be left unaffected as these are most likely to contain mainly signal information – here the coarsest two levels are not subject to shrinkage.

4.4 Multiple prior models for wavelet coefficients

As an extension, the various parameters are now allowed to be grouped by wavelet resolution level with the obvious extensions to the definitions given in the previous section. For the wavelet coefficients, $\mathbf{d_f}$ the Gaussian prior density function becomes

$$\pi(\mathbf{d_f}|\boldsymbol{\kappa}) = \left(\frac{\kappa_j}{\pi}\right)^{(2^j-1)/2} \exp\left\{-\kappa_j \sum_{l=0}^{2^j-1} \mathrm{d}_{\mathrm{f}l,j}^2\right\}, \quad \mathbf{d_f} \in \mathbb{R}^n; \boldsymbol{\kappa} > \mathbf{0},$$

where $\boldsymbol{\kappa} = (\kappa_j : j = 0, \ldots, J-1)$ with $J = \log_2(n)$ and $\mathbf{d_f}^j$ are the level j wavelet coefficients. The Laplace prior density function becomes

$$\pi(\mathbf{d_f}|\boldsymbol{\kappa}) = \left(\frac{\kappa_j}{2}\right)^{2^j-1} \exp\left\{-\sum_{l=0}^{2^j-1} \kappa_j |\mathrm{d}_{\mathrm{f}l,j}|\right\}, \quad \mathbf{d_f} \in \mathbb{R}^n; \boldsymbol{\kappa} > \mathbf{0},$$

and for *elastic-net* based model, the density function is

$$\pi(\mathbf{d_f}|\boldsymbol{\kappa}, \boldsymbol{\gamma}) = \left(\frac{1}{Z(\kappa_j, \gamma_j)}\right)^{2^j-1} \exp\bigg\{ - \kappa_j \sum_{l=0}^{2^j-1} \Big(\gamma_j \mathrm{d}_{\mathrm{f}j,l}^2 + (1-\gamma_j)|\mathrm{d}_{\mathrm{f}j,l}|\Big)\bigg\},$$
$$\mathbf{d_f}_j \subset \mathbb{R}^{2^{j-1}}; \kappa_j > 0, 0 < \gamma_j < 1,$$

where

$$Z(\kappa_j, \gamma_j) = \begin{cases} 2/\kappa_j & \gamma_j = 0, \\ \frac{2\sqrt{\pi}}{\sqrt{\kappa_j \gamma_j}} \exp\left(\frac{\kappa_j(1-\gamma_j)^2}{4\gamma_j}\right)\left(1 - \Phi\left(\frac{\kappa_j(1-\gamma_j)}{\sqrt{2\kappa_j\gamma_j}}\right)\right), & 0 < \gamma_j < 1, \\ \sqrt{\pi/\kappa_j} & \gamma_j = 1, \end{cases}$$

for $j = 0, \ldots, J-1$. Then, with common hyper-prior densities

$$\pi(\gamma_j) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\, \gamma_j^{a-1}(1-\gamma_j)^{b-1}, \quad 0 \le \gamma_j \le 1, j = 0, \ldots, J-1; a, b > 0.$$

and

$$\pi(\kappa_j) \propto \exp\Big(-c/\kappa_j \Big), \quad \kappa_j > 0, j = 0, \ldots, J-1; c > 0$$

The same hyper-prior parameters, a, b and c, are used as in the non-level dependent densities. As for the single wavelet coefficient prior, again here the coarsest two levels are not subject to shrinkage.

5 Numerical methods

A standard Metropolis–Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) is used to produce approximate samples from the posterior distribution by simulating a Markov chain. The use of such methods for parameter estimation, and more general density exploration, through the Markov chain Monte Carlo (MCMC) approach, is widespread – a review can be found in Robert and Casella (2011), then for theoretical details see Gamerman and Lopes (2006); Lui (2001); Brooks et al. (2011), for general practical examples see the collection by Gilks et al. (1995) and for examples in archaeology see Aykroyd et al. (2001); Aykroyd and Al-Gezeri (2014).

Based on the various model definitions in the previous section the parameter vector will simply be referred to as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ which will represent $(\mathbf{d_f}, \kappa)$, $(\mathbf{d_f}, \boldsymbol{\kappa})$, $(\mathbf{d_f}, \kappa, \gamma)$, or $(\mathbf{d_f}, \boldsymbol{\kappa}, \boldsymbol{\gamma})$ as appropriate, with p simply counting the total number of parameters.

The Markov chain can start at any feasible point in the parameter space, let this arbitrary value be denoted $\boldsymbol{\theta}^0$. From this starting value a discrete time Markov chain is simulated to produce values $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^K$. The algorithm will now be defined, and is also summarised in Figure 1.

Consider the Markov chain transition from state $\boldsymbol{\theta}^{k-1}$ at time $k-1$ to state $\boldsymbol{\theta}^k$ at time k. One of the simplest schemes, which works well for many applications, is based on separate single variable updates based on a random walk. That is, at each step only the value of a single variable is proposed and further that the proposal is a perturbation of the current value. Suppose that a new value for θ_i is being proposed, then $\theta_i^k = \theta_i^{k-1} + \epsilon$ and an obvious choice is $\epsilon \sim N(0, \tau^2)$. This proposal is accepted with probability

$$\min\left\{1, \frac{\pi(\boldsymbol{\theta}^k|\mathbf{y})}{\pi(\boldsymbol{\theta}^{k-1}|\mathbf{y})}\right\}$$

otherwise the value is reset with $\theta_i^k = \theta_i^{k-1}$.

Although, the derivation is complicated, the statement and implementation of the algorithm is usually straightforward. The only consideration left is the choice of proposal variance. When choosing a value for τ^2, it is important to realise that both low and high values lead to long transient periods and highly correlated samples and hence unreliable

Set an initial value for $\boldsymbol{\theta}$, call this $\boldsymbol{\theta}^0$
Repeat the following steps for $k = 1, \ldots, K$
  Repeat the following steps for $i = 1, \ldots, p$
    Generate ϵ from a Gaussian distribution $N(0, \tau^2)$
    Generate a propose new value $\theta_i^* = \theta_i^k + \epsilon$
    Evaluate

$$\alpha = \alpha(\boldsymbol{\theta}^k|\boldsymbol{\theta}^*) = \min\left\{1, \frac{\pi(\theta_1^k, \ldots, \theta_{i-1}^k, \theta_i^*, \ \theta_{i+1}^{k-1}, \ldots, \theta_m^{k-1}|\mathbf{y})}{\pi(\theta_1^k, \ldots, \theta_{i-1}^k, \theta_i^{k-1}\, \theta_{i+1}^{k-1}, \ldots, \theta_m^{k-1}|\mathbf{y})}\right\}$$

    Generate u from a uniform distribution, $U(0, 1)$
    If $\alpha > u$, accept and set $\theta_i^k = \theta_i^*$, else $\theta_i^k = \theta_i^{k-1}$
  End repeat
End repeat
Discard initial values and use remainder to make inference.

Algorithm 1: A single-variable random walk MCMC algorithm.

estimation. A reasonable proposal variance can be chosen adaptively during the early burn-in period, and it has been proven theoretically that for a wide variety of high dimensional problems an acceptance rate of 23.4% (Gelman et al., 1997) is optimal. Further, to get good mixing and hence low autocorrelation it may be necessary to include separate proposal variance for groups of parameters, or even for individual parameters.

If the algorithm is designed carefully, then as the iterations progress the current parameter set does not depend on the starting values, and the subsequent values can be treated as a correlated sample from the posterior distribution. Key issues then become how to judge when this initial transient behaviour has ended, and the chain is in equilibrium, and how many iterations to perform to have a sufficiently large sample for reliable estimation. Further, it is wise to also check Markov chain paths and to calculate sample autocorrelation functions. For good estimation the paths should look "unstructured" and the autocorrelation functions be close to zero for all except small lags. A variety of more formal convergence diagnostics are available, see for example Raftery and Lewis (1995), Cowles and Carlin (1996) and Geyer (2011). Once the sample has been generated from the posterior distribution, a number of possible estimators are available. Let $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots, \boldsymbol{\theta}^N$ be the MCMC sample collected after equilibrium of the Markov chain has been declared then the pixel-wise posterior mean and variances can be estimated by the sample mean and variance. Also, sample percentiles can be used to estimate confidence bounds. Various options will be needed in the next section.

6 Application to real data

In this section the results of a real data analysis using the models described earlier are shown. The cores were extracted from 'the Park', Guiting Power in Gloucestershire, which is a late iron-age farmstead, but were part of a modern investigation. The experiment consisted of burning to the ground a wooden funeral pyre containing the corpse of a sheep and covering the burnt area with top-soil. Five cores were removed from the pyre region of the site, four from the main area of burning and one from the periphery — for more detail see Allum et al. (1999); Aykroyd and Al-Gezeri (2014). The analysis for the core from the periphery (Core I) and two of the four cores from the main area (Cores II and III) will be described in detail.

In turn each of the three prior distributions has been used, and also with single prior distribution and with multiple, level-specific, prior distributions for the wavelet parameters — giving six reconstructions for each dataset. The aim is to produce a piecewise constant susceptibility profile and hence the Haar wavelet is used. The summary output will consist of posterior estimates of the underlying magnetic susceptibility function, with credible intervals, and posterior distributions of wavelet coefficients which allow the easy identification of substantial wavelet coefficients.

For estimation, the Algorithm 1 was run with a burn in of 1000 iterations, and then a main run of 10000 iterations. Wavelet coefficients were initialised at zero and prior parameters at values corresponding to their prior mean. For this problem, it was found that efficiency can be increased by using individual proposal variances for each parameter. Large initial values of the proposal variances were set, but these were adjusted automatically every 10 iterations during the burn-in phase to achieve an acceptable acceptance rate. To reduce autocorrelation the main run was thinned by taking every 10 iterations to produce a working sample of 1000. Various monitoring plots and statistics were considered to confirm that the burn-in period was sufficient to consider the remaining iterations for estimation, and that the main run was sufficiently large for good estimation.

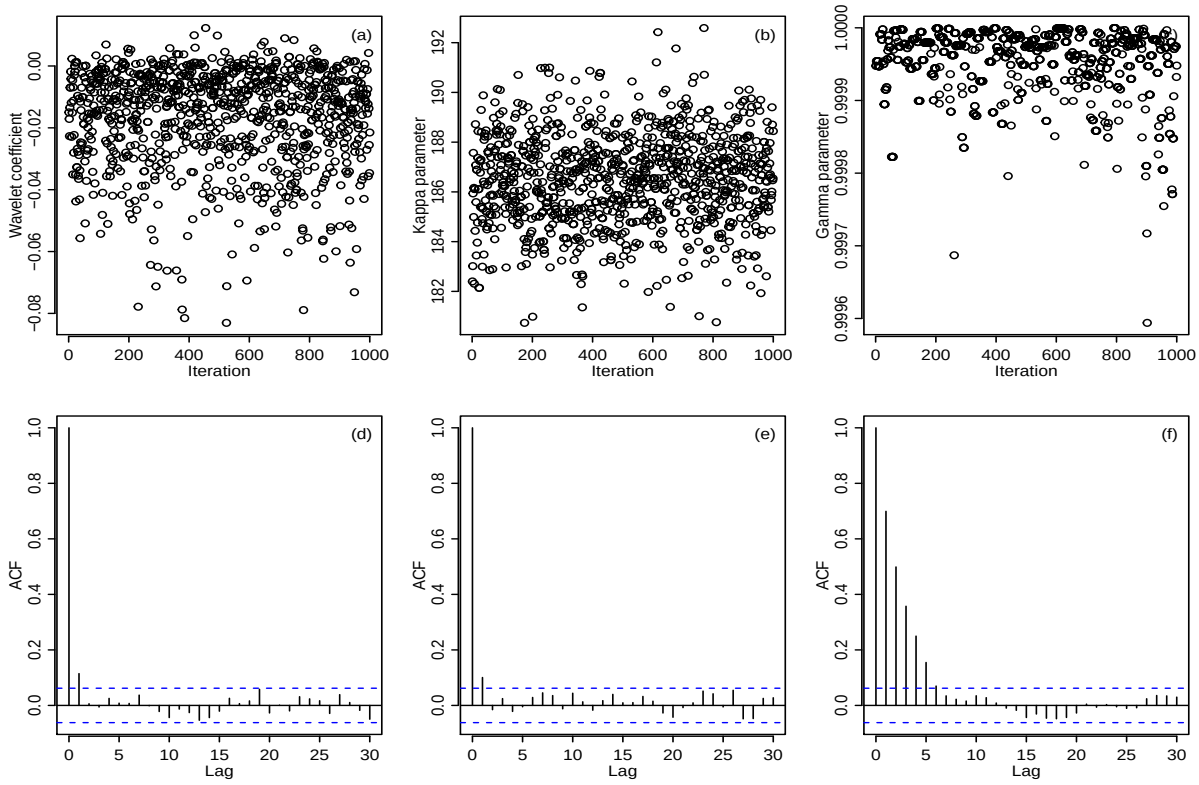


Figure 3: Examples of monitoring output from elastic-net single prior model: (a)-(c) trace plots for selected parameters and (d)-(f) corresponding autocorrelation functions.

Figure 3 shows examples of monitoring statistics: a course resolution wavelet coefficient from resolution level 2 in (a), κ in (b) and γ in (c) with corresponding autocorrelation functions shown in (d)-(f). These show good mixing with low correlation indicating that the algorithm is efficient. It is noticeable, however, from (c) and (f), that estimation of γ is more difficult than the other parameters. Sample size calculations, based on time series principles (see, for example, Sokal, 1989), suggest that about 500 iterations are required — which is well within the size of the working sample. The working sample of size 1000 can now be used for parameter estimation and other inference. For example, using the posterior mean or median as point estimates and 95% credible intervals calculated based on posterior variance or posterior percentiles.

Figure 4 summarises the estimation of the wavelet coefficients using a novel augmented wavelet tableaux — for details of the standard version see, for example, Donoho and Johnstone (1994) or Silverman (1999). For ease of interpretation only the coarsest five resolution levels are shown and a common scale is used throughout to allow comparison. The resolution level is shown on the vertical scale, with 0 being the coarsest. At level 0 there is a single wavelets coefficient, at level 1 there are two etc. The horizontal scale shows the location of the wavelet as a proportion of the total length. The single level 0 coefficient multiplies a wavelet which spans the full width, each of the two level 1

coefficients multiplies a wavelet which spans only half the full interval etc. The full wavelet approximation is then given by the summation of all these contributions.

The posterior sample for each wavelet coefficient is summarised using a block whose height is proportional to the posterior mean and whose width is inversely proportional to the posterior standard deviation — this means that the area of a block is a measure of significance of that coefficient. Those with a credible interval including zero are show in red. The top row summarises results using the models from a single prior distribution for the wavelet coefficients, and the bottom row the corresponding model but for multiple level-specific prior distributions. Those in the left-hand column use the Laplace distribution, the middle use the elastic-net based distribution, and those on the right use a Gaussian model. Note that in these graphs only the relative coefficient estimates and variability can be seen.

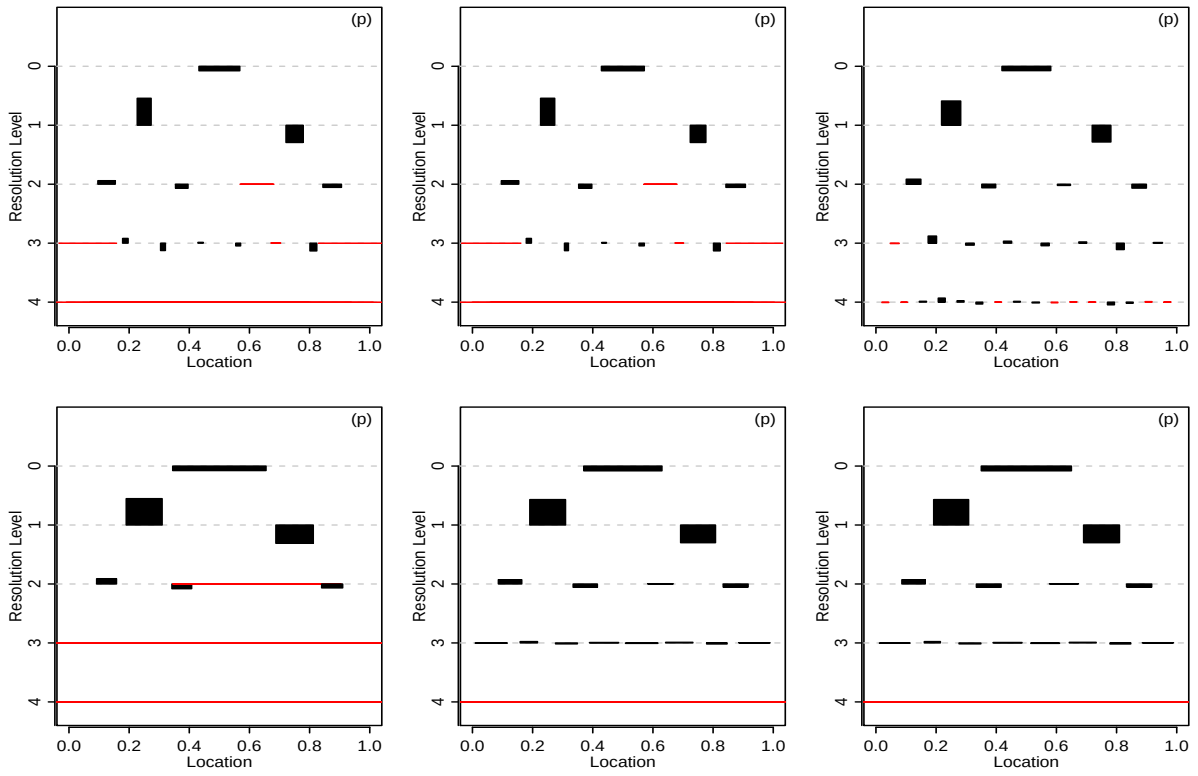


Figure 4: Pyre core I. Wavelet coefficient posterior distributions showing the five coarsest levels using the following prior: (a) single Laplace, (b) single elastic-net, (c) single Gaussian, (d) level-specific Laplace, (e) level-specific elastic-net, (f) level-specific Gaussian.

For all models the posterior variability is very small compared to the magnitude of the coefficients, but there is interesting detail. The significant wavelet coefficients are generally the same, although there are a greater number in the Gaussian case. It is very noticeable that when multiple level-specific prior models are used the variability is lower.

		Laplace	Elastic net	Gaussian
Pyre I	$\hat{\kappa}$	236.6	236.4	143.6
		(231.2,241.9)	(231.3,241.4)	(141.28,146.02)
	$\hat{\gamma}$	-	0.9999	-
			(0.9995,1.000)	
Pyre II	$\hat{\kappa}$	186.5	186.6	83.7
		(183,190.4)	(182.9,189.9)	(82.6,84.9)
	$\hat{\gamma}$	-	1.000	-
			(0.9998,1.000)	
Pyre III	$\hat{\kappa}$	215.3	215.6	124.3
		(210.9,220.0)	(211.3,220.1)	(122.4,126.3)
	$\hat{\gamma}$	-	0.9999	-
			(0.9996,1.000)	

Table 1: Parameter estimates (with 95% credible intervals) for the single prior model.

Estimates of the prior parameters are shown numerically for the single prior models in Table 1 and graphically in Figure 5 for the multiple level-specific model prior parameters.

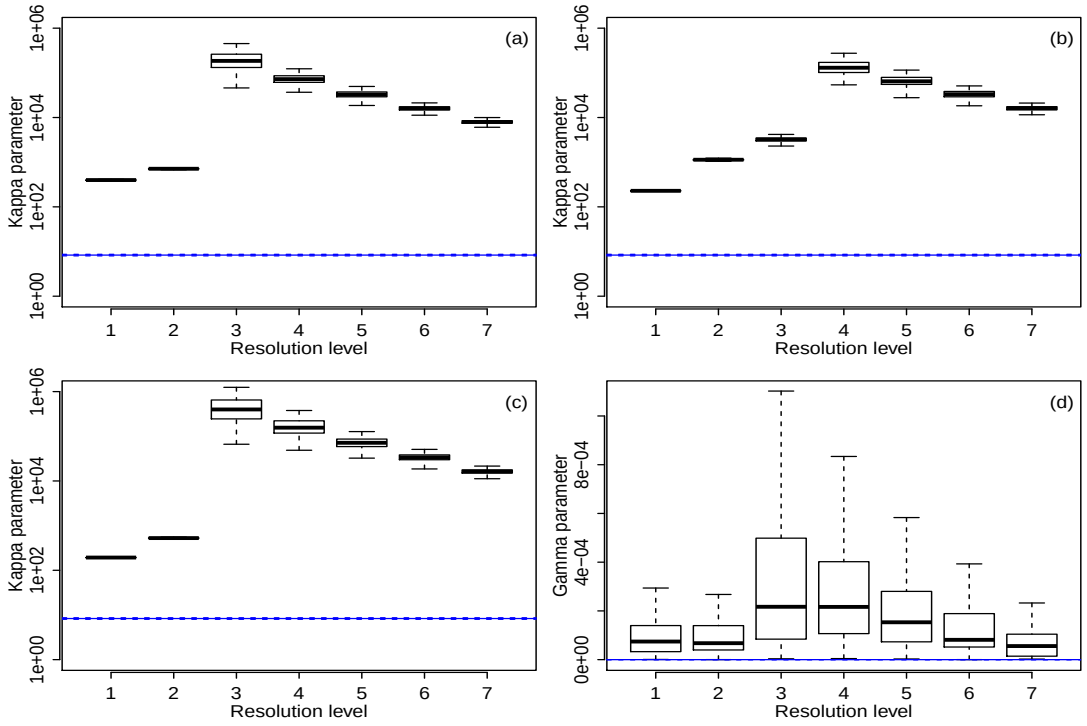


Figure 5: Pyre core I. Boxplots of multiple level-specific parameters: (a) $\hat{\boldsymbol{\kappa}}$ for the Laplace, (b) $\hat{\boldsymbol{\kappa}}$ for the elastic-net, (c) $\hat{\boldsymbol{\kappa}}$ for the Gaussian, (d) $\hat{\boldsymbol{\gamma}}$ for the elastic-net.

In Figure 5, panel (a) shows $\boldsymbol{\kappa}$ for the Laplace, (b) the elastic-net and (c) Gaussian

prior, with (d) showing $\boldsymbol{\gamma}$ for the elastic-net – recalling that $\boldsymbol{\gamma} = 1$ for the Laplace and $\boldsymbol{\gamma} = 0$ for the Gaussian. A horizontal blue line, surrounded by blue dotted lines, show the posterior mean and 95% credible interval for the corresponding parameter in the single prior model. These are representative of all cases, and in fact there is a largely similar pattern for all three prior models. There are lower values of κ for course resolutions (Levels 2, 3 and 4) and high values for other levels with a downward trend moving from Level 6 to Level 8. The estimates of $\boldsymbol{\gamma}$ are all very small, but all significantly away from the value in the corresponding parameter of the single prior model. Most interestingly, all estimates for $\boldsymbol{\gamma}$ are close to zero, that is they lead to a prior close to the Gaussian.

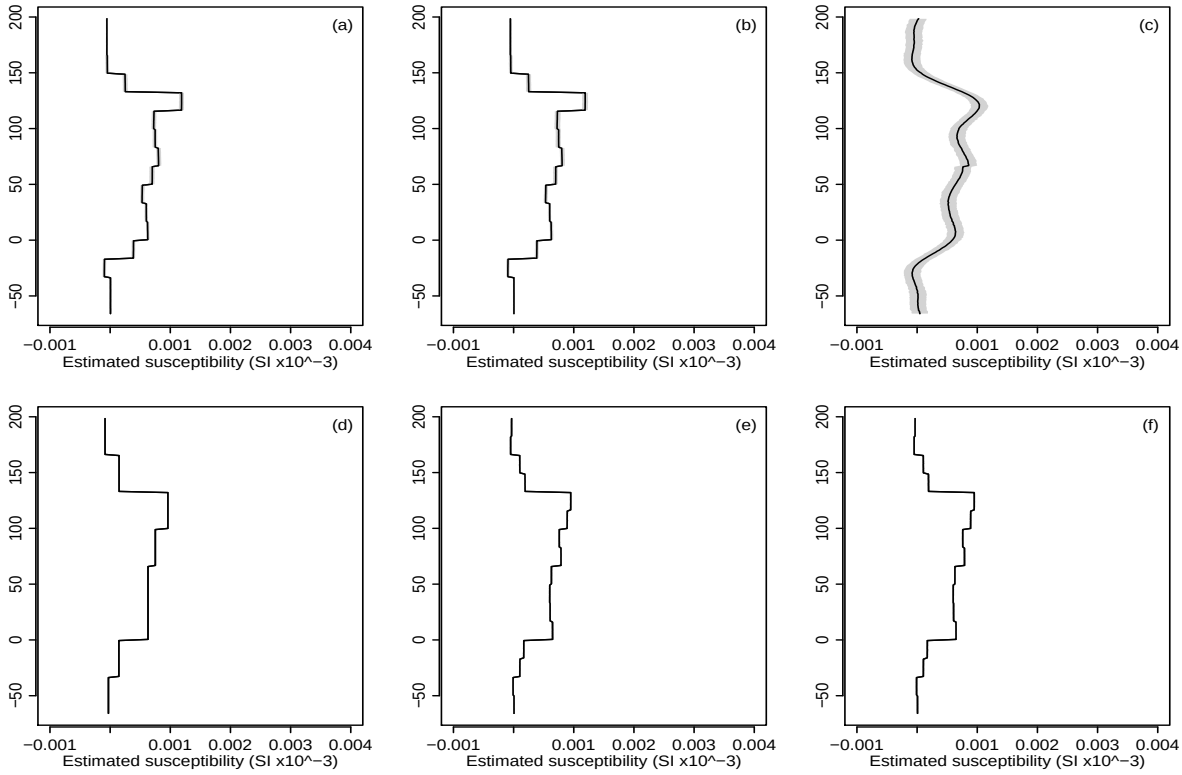


Figure 6: Pyre core I. Posterior mean (line) with point-wise 95% credible interval (grey) using the following prior: (a) single Laplace, (b) single elastic-net, (c) single Gaussian, (d) level-specific Laplace, (e) level-specific elastic-net, (f) level-specific Gaussian.

Figure 6 shows the magnetic susceptibility reconstructions based on the posterior median estimate where the hyper-parameters κ and γ are estimated using both single prior distributions (top row) and multiple level-specific prior distributions (bottom row) – this layout is as in earlier figure. The posterior estimate is shown as a line surrounded by a, very narrow, point-wise 95% credible interval shown in grey estimated using the sample percentiles. The most noticeable feature is the very poor performance of the single Gaussian wavelet prior model — the reconstruction is totally unacceptable as even the

known discontinuities at 0 and 150mm are masked by smoothing. The other combinations are acceptable, but there are clear benefits from using the multiple level-specific prior distributions for the wavelet coefficients as these lead to cleaner jumps and flatter tops in the reconstructions. Of the level-specific prior models, the Laplace reconstruction appears slightly better and uses fewer wavelet coefficients.

Figure 7 shows the wavelet tableaux corresponding to Core II with the same conclusions as from the Core I data. Any of the reconstruction, except the single Gaussian prior, produce a very sparse representation. The magnetic susceptibility profiles in Figure 8 clearly separate the core into three parts for the Laplace prior and the elastic-net prior model, whereas for the single prior Gaussian the reconstruction is unacceptable. Again, the slightly better is the multiple level-specific prior Laplace prior.

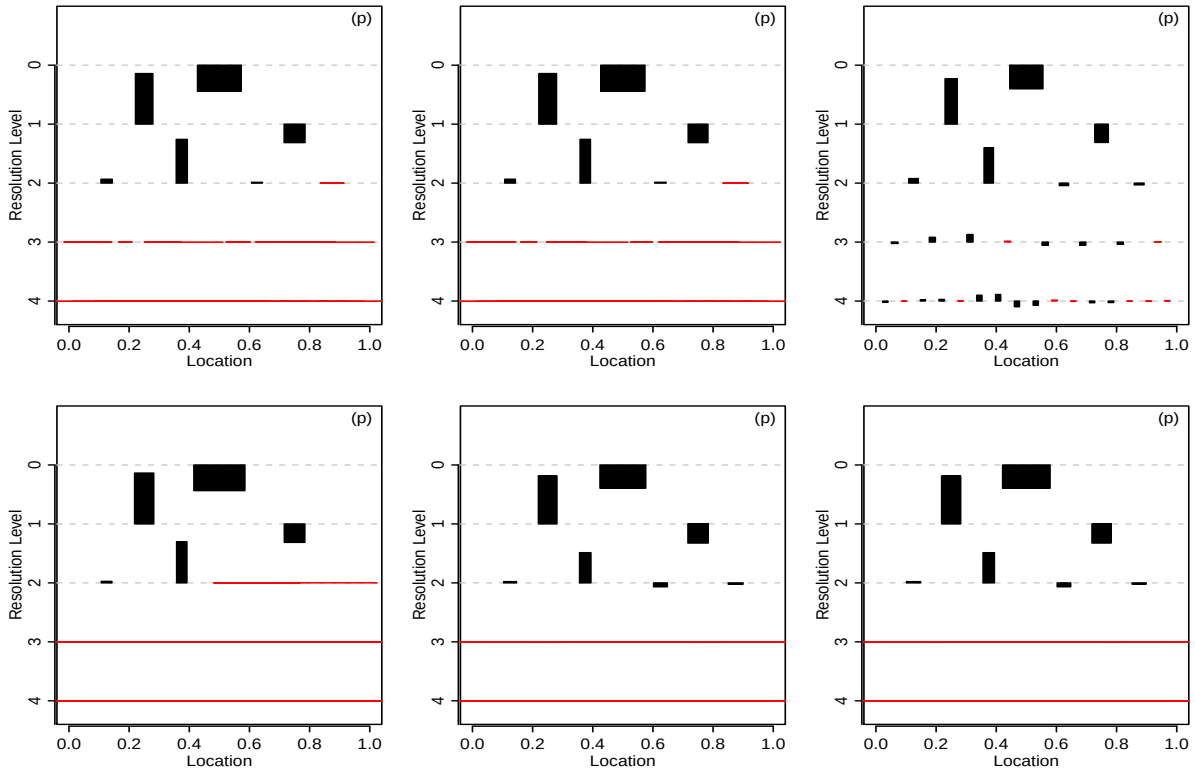


Figure 7: Core II. Wavelet coefficient posterior distributions showing the five coarsest levels using the following prior: (a) single Laplace, (b) single elastic-net, (c) single Gaussian, (d) level-specific Laplace, (e) level-specific elastic-net, (f) level-specific Gaussian.

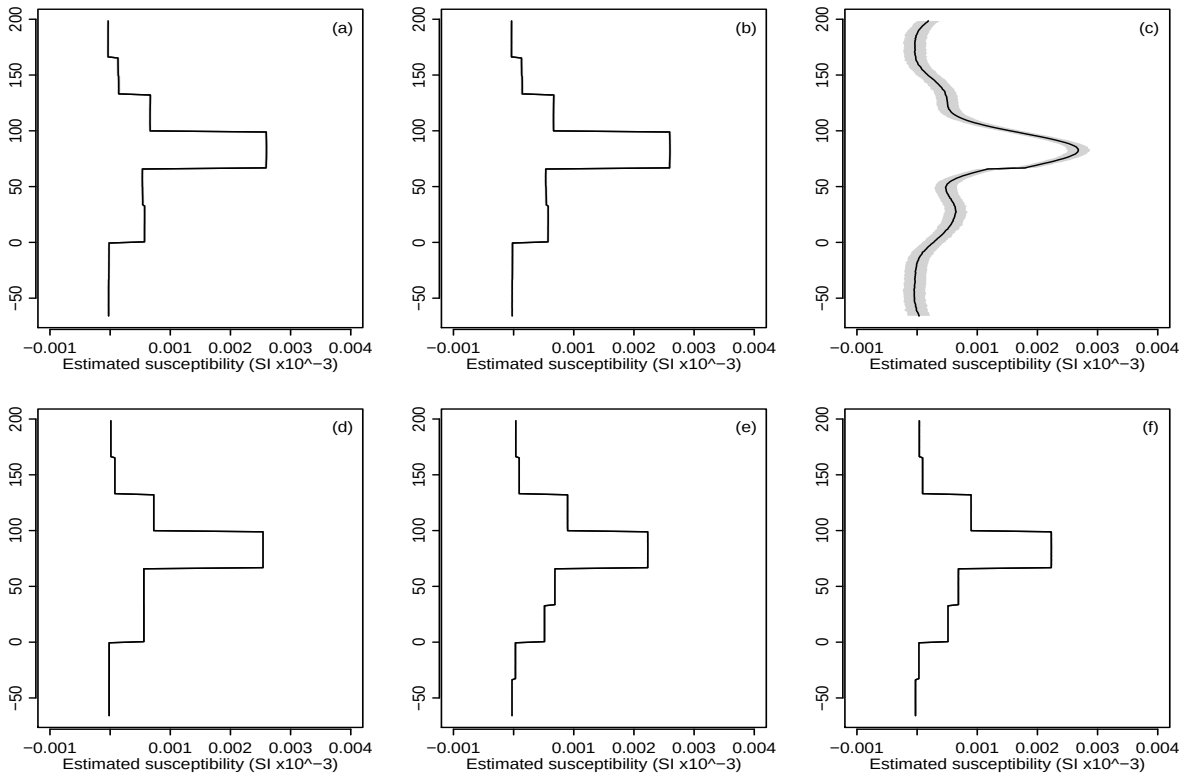


Figure 8: Core II. Posterior mean (line) with point-wise 95% credible interval (grey) using the following prior: (a) single Laplace, (b) single elastic-net, (c) single Gaussian, (d) level-specific Laplace, (e) level-specific elastic-net, (f) level-specific Gaussian.

Figure 9 show the reconstructions and wavelet tableaux corresponding to Core III with similar conclusions as from the previous core data. That is that the single Gaussian prior is unacceptable, and that the multiple level-specific models perform better than the single prior models. However, here the best reconstructions are with the multiple prior elastic-net and Gaussian models, obtained using only six wavelet coefficients.

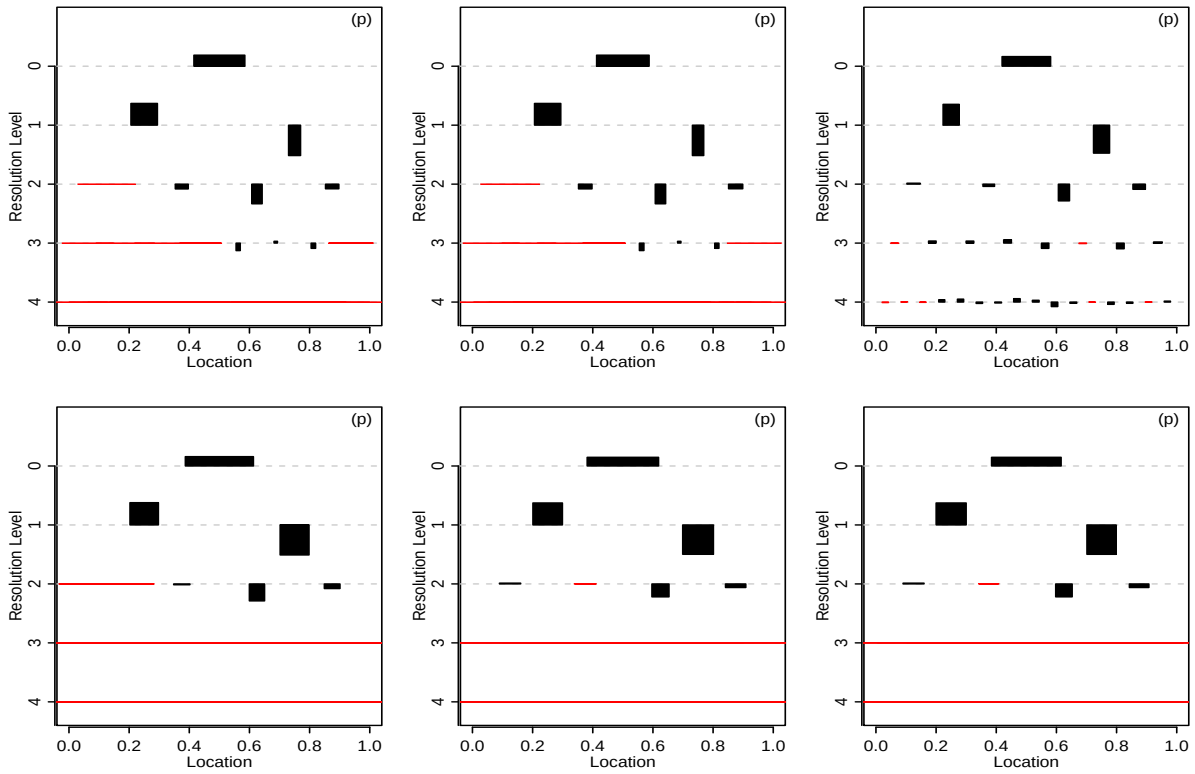


Figure 9: Core III. Wavelet coefficient posterior distributions showing the five coarsest levels using prior models: (a) single Laplace, (b) single elastic-net, (c) single Gaussian, (d) level-specific Laplace, (e) level-specific elastic-net, (f) level-specific Gaussian.

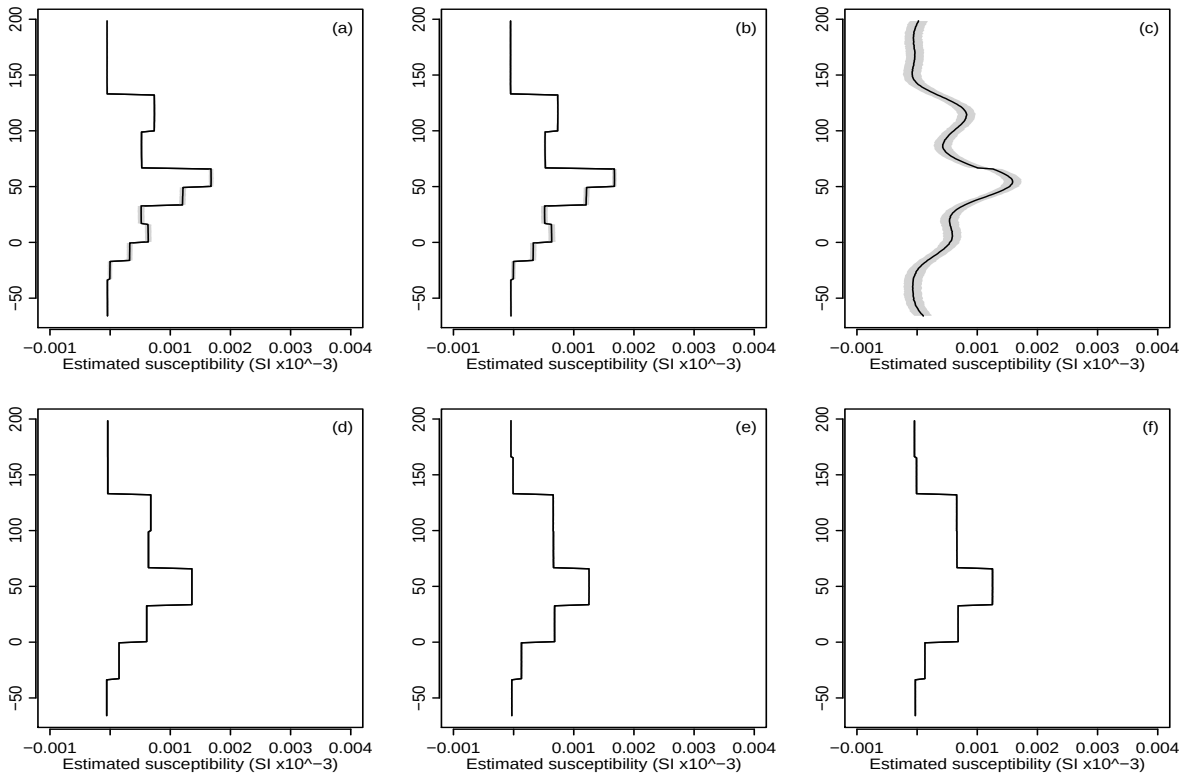


Figure 10: Pyre core III. Posterior mean (line) with point-wise 95% credible interval (grey) using the following prior: (a) single Laplace, (b) single elastic-net, (c) single Gaussian, (d) level-specific Laplace, (e) level-specific elastic-net, (f) level-specific Gaussian.

7 Discussion

The aim of this work was to investigate the use of wavelet-based models for the estimation of piecewise constant functions in inverse problems. A general framework for dealing with such problems has been laid-out within which alternative model components can be considered in the future. The MCMC algorithm described provides a simple method to estimate wavelet coefficients, but also it allows functions of the parameters to be considered. Here, the underlying susceptibility profile is such a function of the wavelet coefficients, and hence can be estimated easily. Moreover, credible intervals and other measures of uncertainty can be considered, and novel graphical summaries can be produced.

Although the Laplace distribution continues to provide a good choice of wavelet coefficient prior model as in other applications, the use of the elastic-net based generalisation proposed here has been demonstrated to be a flexible alternative. More importantly, the proposed use of multiple prior distributions to model the wavelet coefficients at difference resolution levels separately is very beneficial. Further, once a move to multiple prior distributions has been made then there is little to choose between the exact form, such as

Laplace, Gaussian or elastic-net, leading to greater robustness — a similar observation was made for 2d estimation in Aykroyd (1998). It is likely that this combination will also be fruitful in other applications. Future work in this area is planned, which will consider recent extensions to wavelet coefficient prior modelling using mixture distributions (see for example, Johnstone and Silverman, 2005) and to prior distributions which use information from the neighbouring wavelet coefficients (Cai, 2008).

Inverse problems are widely encountered in the applied sciences and assumptions of piecewise constant, or at least piecewise smooth functions, are common. It is usual, however, to use prior distributions explicitly in terms of the function values themselves which usually lead to poor reconstruction — shrinkage type models move the estimates towards zero whilst smoothing prior models can destroy sharp discontinuities. Hence, the approach proposed here has the potential to have significant impact on a wide range of practical problems.

Acknowledgments

The authors thank the Editor and anonymous referees for their constructive comments which resulted in this improved version.

References

Abramovich, F. and B. Silverman (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika 85(1), pp.115–129*.

Allum, G., R. Aykroyd, and J. Haigh (1999). Empirical Bayes estimation for archaeological stratigraphy. *Journal of the Royal Statistical Society, Series C 48*, 1–14.

Aykroyd, R. G. (1998). Bayesian estimation for homogeneous and inhomogeneous Gaussian random fields. *IEEE Trans. PAMI 20*, 533–539.

Aykroyd, R. G. (2015). *Industrial tomography: Systems and applications*, Chapter Statistical image reconstruction, pp. 401–427.

Aykroyd, R. G. and S. M. Al-Gezeri (2014). 3D modelling and depth estimation in archaeological geophysics. *Chilean Journal of Statistics 5*, 19–35.

Aykroyd, R. G., J. G. B. Haigh, and G. T. Allum (2001). Bayesian methods applied to survey data from archaeological magnetometry. *Journal of the American Statistical Association 96*, 64–76.

Berger, J. O. and L. R. Pericchi (2001). *Objective Bayesian Methods for Model Selection: Introduction and Comparison*, Volume 38 of *Lecture Notes–Monograph Series*, pp. 135–207. Beachwood, OH: Institute of Mathematical Statistics.

Besag, J. (1989). Towards Bayesian image analysis. *Journal of Applied Statistics 16*, 395–407.

Bracewell, R. (1986). *The Fourier transform and its applications* (2nd edition ed.). McGraw and Hill.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC.

Cai, T. T. (2008). On information pooling, adaptability and superefficiency in nonparametric function estimation. *Journal of Multivariate Analysis 99*, 412–436.

Champeney, D. (1987). *A handbook of Fourier theorems*. Cambridge University Press.

Cowles, M. K. and B. P. Carlin (1996). Markov chain Monte Carlo convergence diagnostics:A comparative review. *Journal of the American Statistical Association 91*, 883–904.

Daubechies, I. (1988). Orthogonal bases of compactly suported wavelets. *Comm. Pure and Appl. Maths. 41*, 909–996.

Donoho, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet vaguelette-decomposition. *Applied and Computational Harmonic Analysis 2*, 101?126.

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

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference* (2nd ed.). Chapman & Hall/CRC Texts in Statistical Science.

Gelman, A., W. R. Gilks, and G. O. Roberts (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab. 7*, Ann. Appl. Probab.

Geyer, C. J. (2011). Introduction to Markov Chain Monte Carlo. In S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng (Eds.), *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

Gilks, W., S. Richardson, and D. Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.). Springer Series in Statistics.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains, and their applications. *Biometrika 57*, 97–109.

Johnstone, I. M. and B. W. Silverman (2005). Empirical bayes selection of wavelet thresholds. *Ann. Statist. 33*, 1700–1752.

Le Borgne, E. (1960). Influence du feu sur les proprietés magnétiques du sol. *Annales Geophysiqu 16*(159–195).

Lui, J. (2001). *Monte Carlo Strategies in Scientific Computing*. Berlin: Springer-Verlag.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chemical Physics 21*, 1087–1091.

Nason, G. (2010). *Wavelet methods in statistics with R. New York: Springer*.

Nason, G. P. (2008). *Wavelet Methods in Statistics with R*. New York: spr.

Press, W.H., T. S. V. W. and B. Flannery (1992). *Numerical Recipes in FORTRAN: the art of scientific computing* (2nd edition ed.). Cambridge University Press.

Raftery, A. and S. Lewis (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter (Eds.), *Practical Markov Chain Monte Carlo*. Chapman and Hall.

Ribés, A. and F. Schmitt (2008). Linear inverse problems in imaging. *IEEE Signal Processing Magazine 25*, 84 – 99.

Robert, C. and G. Casella (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science 26*, 102–115.

Silverman, B. W. (1999). Wavelets in statistics: beyond the standard assumptions. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences 357(1760), pp.2459–2473*.

Sokal, A. D. (1989). Monte Carlo methods in statistical mechanics: foundations and new algorithms. *Cours de Troisième Cycle de la Physique en Suisse Romande, Lausanne*.

Stuart, A. M. (2010). Inverse problems: A Bayesian perspective. *Acta Numerica 19*, 451–559.

Appendix: Normalization of the elastic-net density

The normalization constant is defined by the following integral

$$Z(\kappa,\gamma)=\int \exp\left\{-\kappa(\gamma {\mathrm{d_g}}^2+(1-\gamma)|\mathrm{d_g}|)\right\}d\mathrm{d_g}, \tag{1}$$

then suppose $A=\kappa\gamma$ and $B=\kappa(1-\gamma)$, giving

$$\begin{aligned}
Z(\kappa,\gamma) &= \int \exp\left\{-(A{\mathrm{d_g}}^2+B|\mathrm{d_g}|)\right\}d\mathrm{d_g} \\
&= \int_{-\infty}^{0} \exp\left\{-(A{\mathrm{d_g}}^2-B\mathrm{d_g})\right\}d\mathrm{d_g}+\int_{0}^{\infty} \exp\left\{-(A{\mathrm{d_g}}^2+B\mathrm{d_g})\right\}d\mathrm{d_g} \\
&= \int_{-\infty}^{0} \exp\left\{-A({\mathrm{d_g}}^2-\frac{B}{A}\mathrm{d_g})\right\}d\mathrm{d_g}+\int_{0}^{\infty} \exp\left\{-A({\mathrm{d_g}}^2+\frac{B}{A}\mathrm{d_g})\right\}d\mathrm{d_g} \\
&= \exp\left\{\frac{1}{4A}B^2\right\}\left\{\int_{-\infty}^{0} \exp\left\{-A(\mathrm{d_g}-\frac{B}{2A})^2\right\}d\mathrm{d_g}+\int_{0}^{\infty} \exp\left\{-A(\mathrm{d_g}+\frac{B}{2A})^2\right\}d\mathrm{d_g}\right\}.
\end{aligned} \tag{2}$$

Now, let $v=\sqrt{2A}(\mathrm{d_g}-\frac{B}{2A})$ and $u=\sqrt{2A}(\mathrm{d_g}+\frac{B}{2A})$, hence

$$\begin{aligned}
Z(\kappa,\gamma) &= \frac{1}{\sqrt{2A}}\exp\left\{A(\frac{B}{2A})^2\right\}\left(\int_{-\infty}^{-\sqrt{2A}\frac{B}{2A}} \exp\left\{-\frac{1}{2}v^2\right\}dv+\int_{\sqrt{2A}\frac{B}{2A}}^{\infty} \exp\left\{-\frac{1}{2}u^2\right\}du\right) \\
&= \frac{\sqrt{\pi}}{\sqrt{A}}\exp\left\{\frac{1}{4A}B^2\right\}\left(\Phi(-\sqrt{2A}\frac{B}{2A})+(1-\Phi(\sqrt{2A}\frac{B}{2A}))\right) \\
&= \frac{\sqrt{4\pi}}{\sqrt{A}}\exp\left\{\frac{1}{4A}B^2\right\}\left(1-\Phi(\sqrt{2A}\frac{B}{2A})\right).
\end{aligned} \tag{3}$$