

Flexible Learning on the Sphere Via Adaptive Needlet Shrinkage and Selection

James G. Scott
McCombs School of Business
University of Texas at Austin
Austin, TX 78712
James.Scott@mcombs.utexas.edu

Original: June 2009
Revised: September 2009

Abstract

This paper introduces an approach for flexible, robust Bayesian modeling of structure in spherical data sets. Our method is based upon a recent construction called the needlet, which is a particular form of spherical wavelet with many favorable statistical and computational properties. We perform shrinkage and selection of needlet coefficients, focusing on two main alternatives: empirical-Bayes thresholding for selection, and the horseshoe prior for shrinkage. We study the performance of the proposed methodology both on simulated data and on a real data set involving the cosmic microwave background radiation. Horseshoe shrinkage of needlet coefficients is shown to yield the best overall performance against some common benchmarks.

1 Introduction

Wavelets are one of the most widely used tools in modern statistics. They have many useful properties that make them appropriate for multiscale analysis and signal processing, and they have also seen broad application in a variety of other contexts, from computer vision to nonparametric function estimation.

The goal of this paper is to explore a set of tools for robust Bayesian modeling on the sphere using needlets, which are a generalization of wavelets to the unit sphere. Spherical data sets arise in astrophysics, cell biology, ecology, geophysical science, medical imaging, and three-dimensional shape recognition. A particularly important application occurs in the analysis of data from NASA's Wilkinson Microwave Anisotropy Probe, whose goal is to investigate the character of the cosmic microwave background (CMB) radiation. Section 4.2 contains an application of Bayesian needlet modeling to a publicly available CMB data set.

Needlets, which were introduced to the mathematical community by Narcowich et al. (2006), have many of the same advantages over spherical harmonics that wavelets enjoy over conventional Fourier series. Like spherical harmonics, needlets have bounded support in the frequency domain. Unlike spherical harmonics, however, needlets also have highly localized support in the spatial domain, decaying quasi-exponentially fast away from their global maximum. As a result, they can easily and parsimoniously represent random fields over the sphere that exhibit sharp local peaks or valleys.

These mathematical features are shared by other forms of spherical wavelets. Needlets, however, have some uniquely advantageous statistical and computational properties. First, a

recent result from Baldi et al. (2009) shows that needlets separated by a fixed geodesic distance have coefficients that are asymptotically uncorrelated, and therefore independent under the assumption of Gaussianity, as resolution increases. This result—the only one of a similar character for any type of spherical wavelets—implies that needlets make an excellent choice of basis for statistical modeling on the sphere. The usual practice in wavelet shrinkage, after all, involves treating empirical wavelet coefficients as though they were observed data arising from a statistical error model, rather than treating the wavelet basis elements themselves as inputs to a regression problem. (See, for example, Clyde and George (2000).) This is sensible because wavelets, unlike needlets, are orthogonal. But the above result, while asymptotic in character, can be thought of a loose justification for approaching needlet shrinkage in much the same way—in essence, to place the likelihood in the multipole domain, rather than the spatial domain. This greatly simplifies matters computationally. In particular, it avoids the difficulty of working with the large matrices that would otherwise be needed in order to represent the needlet basis elements, which are not orthogonal.

A second useful feature of needlets is that the same batch of needlet functions appears in both the forward and reverse needlet transform. This computationally attractive property is a highly nontrivial result of the careful construction used to define needlets, and is not shared by other commonly used forms of spherical wavelets.

Further investigations of the theoretical properties of needlets can be found in Baldi et al. (2007) and Baldi et al. (2008). An application of needlets to CMB data analysis appears in Marinucci et al. (2008), while a Bayesian treatment of other kinds of spherical wavelets for shape recognition is in Faucheur et al. (2007).

This paper makes the following contributions to this very recent literature:

1. We introduce a novel Bayesian modeling approach for robust shrinkage and selection of needlet coefficients. This method is derived from the horseshoe prior of Carvalho et al. (2008), and differs markedly both from existing needlet methods based on thresholding, and from existing Bayesian methods for conventional wavelets.
2. We investigate the need to properly scale the empirical needlet coefficients before any shrinkage procedure is applied, which can dramatically affect performance.
3. We study the problem of sparsity in the multipole domain, and show how sparsity patterns can be accommodated using a highly adaptive approach for needlet selection.

2 Spherical Needlets

2.1 The mathematical construction

The following construction of needlets is due to Narcowich et al. (2006), and reflects the same normalization described by Baldi et al. (2009). Needlets are based on two complementary ideas familiar from conventional wavelets: the discretization of the sphere into successively finer meshes of basis elements, and the construction of a “window” operator whose convolution with a periodic function can yield spatial localization.

Let \mathbb{S}^2 denote the unit sphere, with coordinates indexed by longitude ϕ and latitude θ . Suppose we have pixelized the sphere using a mesh $\Xi_j = \{\xi_{jk}\}_{k=1}^{M_j}$, where ξ_{jk} is the k th pixel center at resolution level j . Associated with each point ξ_{jk} is a weight λ_{jk} , chosen so that functions f over the sphere can be integrated using the cubature formula

$$\int_{\mathbb{S}^2} f(x) dx \approx \sum_{k=1}^{M_j} \lambda_{jk} f(\xi_{jk}).$$

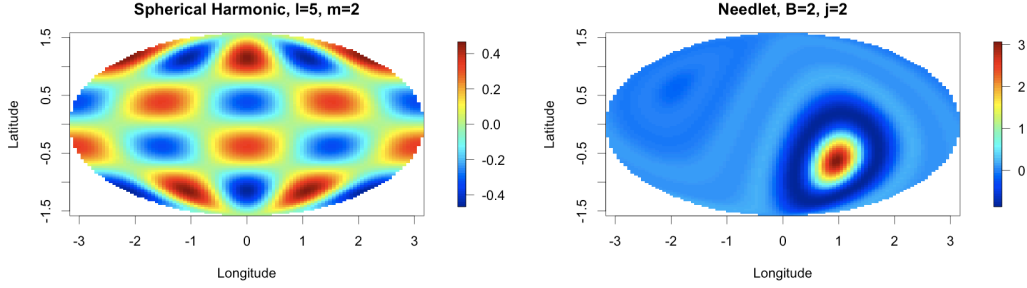


Figure 1: Left: spherical harmonic for $l = 5$, $m = 2$. Right: a needlet centered at $(\theta = \pi/3, \phi = -\pi/6)$ with $j = 2$ and $B = 2$. The sphere has been projected to \mathbb{R}^2 using the Mollweide projection.

In practice, the pixels are often chosen to have equal areas, in which case $\lambda_{jk} = 4\pi/M_j$ (recalling that the sphere has total Lebesgue measure 4π).

Let $\{Y_l^m(x) : l \geq 0, -l \leq m \leq l\}$ be the set of orthonormal spherical harmonics, and let α_l^m be their associated coefficients, so that any L^2 function on the sphere can be expanded as

$$f(x) = \sum_{l=0}^{\infty} \sum_{m=-l}^l \alpha_l^m Y_l^m(x) \quad (1)$$

$$\alpha_l^m = \int_{\mathbb{S}^2} f(x) \bar{Y}_l^m(x) dx, \quad (2)$$

where $\bar{\cdot}$ denotes complex conjugation.

The spherical needlet function centered at the cubature point ξ_{jk} is then defined, for some fixed bandwidth parameter $\delta > 1$, as

$$\psi_{jk}(x) = \sqrt{\lambda_{jk}} \sum_{l=\lfloor \delta^{j-1} \rfloor}^{\lceil \delta^{j+1} \rceil} b_\delta \left(\frac{l}{\delta^j} \right) \sum_{m=-l}^l \bar{Y}_l^m(x) Y_l^m(\xi_{jk}), \quad (3)$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote the floor and ceiling operators, respectively.

The function b_δ , meanwhile, is defined by a Littlewood–Paley decomposition. Let μ_δ be an even function that has continuous derivatives of all orders, that has support on $[-1, 1]$, that is nonincreasing away from zero, and that takes values on $[0, 1]$, with $\mu(x) = 1$ whenever $|x| \leq \delta^{-1}$. Then define:

$$b_\delta(x) = \sqrt{\mu(x/\delta) - \mu(x)}.$$

The map of interest, f , is then reconstructed using the needlet expansion as an alternative to the harmonic expansion in (1):

$$f(x) = \sum_j \sum_{k=1}^{M_j} \beta_{jk} \psi_{jk}(x). \quad (4)$$

The coefficients β_{jk} are given by

$$\beta_{jk} = \sqrt{\lambda_{jk}} \sum_{l=\lfloor \delta^{j-1} \rfloor}^{\lceil \delta^{j+1} \rceil} b_\delta \left(\frac{l}{\delta^j} \right) \sum_{m=-l}^l \alpha_l^m Y_l^m(\xi_{jk}). \quad (5)$$

This reconstruction formula appears simple on its face, and indeed is quite straightforward to implement in practice. It is, however, a profound consequence of the carefully chosen properties required of b_δ . Intuitively, b_δ operates as a “window” function, one that is convolved with the spherical harmonics across a bounded set of frequencies $l = \lfloor \delta^{j-1} \rfloor, \dots, \lceil \delta^{j+1} \rceil$ to produce a needlet. (The authors of Marinucci et al. (2008) show how an example of such a μ can be explicitly defined using elementary functions, though any μ satisfying the Littlewood–Paley construction will suffice.) Figure 1 shows examples of a single spherical harmonic and a single needlet projected onto the plane.

2.2 Constructing random fields using needlets

Suppose that we wish to estimate a random field f over the sphere on the basis of a noisy realization $y(x)$ for $x = \{x_i = (\theta_i, \phi_i)\}_{i=1}^N$. The needlet estimation procedure begins by reconstructing the set of empirical harmonic coefficients via cubature, up to some maximum resolution ℓ_{\max} :

$$\hat{\alpha}_l^m = \sum_{i=1}^N w_i y(x_i) \bar{Y}_l^m(x_i),$$

with w_i an appropriate cubature weight reflecting the surface area associated with pixel x_i . Once the meshes Ξ_j are chosen, the needlet coefficients $\hat{\beta}_{jk}$ are then computed by plugging the harmonic coefficients $\hat{\alpha}_l^m$ into (5), yielding

$$\hat{f}_Q = \sum_j \sum_{k=1}^{M_j} \hat{\beta}_{jk} \psi_{jk}.$$

The focus of this paper is on improving the straight cubature estimator through shrinkage and selection of the empirical needlet coefficients $\hat{\beta}_{jk}$. The quality of the resulting reconstruction can be measured by standard loss functions, either in the spatial domain or the needlet domain. This paper will use quadratic loss,

$$\ell^2(f, \hat{f}) = \sum_{i=1}^N \{f(x_i) - \hat{f}(x_i)\}^2 \quad \text{and} \quad \ell^2(\beta, \hat{\beta}) = \sum_j \sum_{k=1}^{M_j} (\beta_{jk} - \hat{\beta}_{jk})^2,$$

though other loss functions involving functions of the $\hat{\beta}_{jk}$'s, such as those for the angular power spectrum of f , are easy to use as well.

2.3 Scaling of the needlet coefficients

Statistical learning of needlet coefficients must immediately confront the issue of scaling, which can dramatically affect the performance of any shrinkage procedure. Essentially, there is a factor of λ_{jk} that must appear in the product of the needlet coefficient and the needlet function in order for the reconstruction in (4) to be valid. But it is not immediately obvious how much of this factor to attribute to the function ψ_{jk} , and how much to the coefficient β_{jk} . A similar issue appears in wavelet shrinkage; see, for example, Vidakovic and Muller (1999). Here, however,

the scale is much harder to determine, because needlets are not constructed in the same “dilate and shift” manner as wavelets—an operation which creates a natural hierarchy of scales.

The authors of Marinucci et al. (2008) observe that the normalization of β_{jk} by $\sqrt{\lambda_{jk}}$ in (5) is the correct constant for reconstructing the properly normalized angular power spectrum of f . Unfortunately, it is not clear that this scale is also appropriate for performing statistical estimation and thresholding of the β_{jk} ’s. The issue is that the coefficients normalized according to (5) cannot be treated as though they are on the same scale. Nor is it appropriate to simply rescale each M_j -sized block of coefficients associated with resolution-level j to have unit variance. The spatially localized behavior of needlets, after all, means that the average needlet loading at level $j + 1$ should be smaller than at level j , even accounting for differences of scale introduced by (5). It is unclear whether the correct rate of this decay, however, is the simple $\sqrt{1/M_j}$ rate.

From a statistical-modeling point of view, a better normalization seems to be

$$\begin{aligned}\tilde{\beta}_{jk} &= (\lambda_{jk})^{-1/2} \eta_j^{-1} \hat{\beta}_{jk} \\ \eta_j &= \sum_{l=\lfloor \delta^{j-1} \rfloor}^{\lceil \delta^{j+1} \rceil} (2l + 1).\end{aligned}\tag{6}$$

The intuition here is the following. After the original normalization by $\sqrt{\lambda_{jk}}$ is undone, the factor η_j simply renormalizes by the number of random terms in the sum that contribute to $\hat{\beta}_{jk}$ at level j . These terms are on a unit scale due to the orthonormality of the Y_l^m ’s, and so each one represents, in some sense, an independent random contribution to $\hat{\beta}_{jk}$. This suggests that all terms be rescaled by η_j^{-1} rather than $\sqrt{\lambda_{jk}}$. The “natural” rate of decay in scale as a function of j will then be encoded by the window function b_δ .

It is an open question whether (6) gives the correct scaling in some more fundamental sense. We have found, however, that shrinkage and selection procedures tend to give much better results—often dramatically better—when they are applied to the rescaled $\tilde{\beta}_{jk}$ ’s rather than the original $\hat{\beta}_{jk}$ ’s. This is the scale, therefore, that will be adopted throughout the rest of the paper.

3 Statistical Modeling of Needlet Coefficients

3.1 Benchmark thresholding procedure

The existing literature on needlets focuses chiefly on estimating random fields using the empirical coefficients from the discrete wavelet transform (see the introduction for references to much of this literature). Yet there is a large body of complementary work on wavelets suggesting that shrinkage or thresholding of empirical coefficients can offer substantial gains in performance. We now demonstrate that the same is true of needlets. Moreover, the potential gains on realistic problems can often be dramatic, while the computational costs are quite low. This fact should be of great interest to practitioners who work with random fields on the sphere, such as the WMAP data considered in the next section.

Let $\tilde{\beta}_j$ be the vector of M_j needlet coefficients for level j , and stack these rescaled coefficients into a single p -dimensional column vector $\mathbf{z} = (\tilde{\beta}_1, \dots, \tilde{\beta}_{j_{\max}})'$. We will treat the vector \mathbf{z} as raw data observed with Gaussian error, $\mathbf{z} \sim N(\boldsymbol{\beta}, \sigma^2 I)$. This assumption of homoskedasticity in the multipole domain is only reasonable if careful attention is paid to the proper scaling of the empirical needlet coefficients.

As a benchmark procedure, we use the empirical-Bayes thresholding rule from Johnstone

and Silverman (2004). Their recommended model can be expressed as

$$\beta_i \sim w \cdot \text{DE}(\beta_i | 0, 1) + (1 - w) \cdot \delta_0,$$

a discrete mixture of a standard double-exponential prior and a point mass at zero, where the mixing weight $w \in [0, 1]$ is unknown and estimated by marginal maximum likelihood. The coefficients β_i are then estimated by the posterior median, which is a “soft” thresholding rule and will zero out any coefficients whose posterior probability being zero is greater than 50%.

The authors of Johnstone and Silverman (2004) demonstrate that this highly adaptive estimator does very well at estimating sparse signals, proving that achieves “near minimaxity” across a wide range of sparsity classes. These theoretical results, coupled with the estimator’s impressive performance on a wide variety of real and realistic data sets, make it an excellent benchmark for the Bayesian needlet shrinkage procedure proposed here.

3.2 Shrinkage with the horseshoe prior

We now develop a Bayesian approach for estimating the needlet coefficients based on the horseshoe estimator of Carvalho et al. (2008).

Specifically, we reconstruct the underlying coefficients $\beta = (\beta_1, \dots, \beta_p)$ using the posterior mean under the horseshoe prior. The horseshoe prior assumes that each β_i has density $\pi_{HS}(\beta_i | \tau)$, with π_{HS} expressible as a scale mixture of normals:

$$\begin{aligned} (\beta_i | \lambda_i, \tau) &\sim \text{N}(0, \lambda_i^2 \tau^2 \sigma^2) \\ \lambda_i &\sim \text{C}^+(0, 1) \\ \tau &\sim \text{C}^+(0, 1), \end{aligned} \tag{7}$$

where $\text{C}^+(0, 1)$ is a half-Cauchy distribution.

Figure 2 plots the densities of the horseshoe, Cauchy and standard normal priors. The horseshoe density $\pi_{HS}(\beta_i | \tau)$ has no closed-form representation, but it obeys very tight upper and lower bounds that are expressible in terms of elementary functions, as detailed in Theorem 1 of Carvalho et al. (2008). Essentially, it behaves like $\pi_{HS}(\beta) \approx \log(1 + 1/\beta^2)$ (up to a constant). The distribution is absolutely continuous with respect to Lebesgue measure, while the density function has an infinitely tall spike at zero and heavy tails that decay like β^{-2} .

The horseshoe prior is in the well-studied family of multivariate scale mixtures of normals. Examples of this family are both common and quite familiar. Choosing $\lambda_i^2 \sim \text{Exp}(2)$, for instance, implies independent double-exponential priors for each β_i ; inverse-gamma mixing, with $\lambda_i^2 \sim \text{IG}(a, b)$, leads to Student-t priors. The former represents the underlying stochastic model for the LASSO of Tibshirani (1996), while the latter is associated with the relevance vector machine (RVM) of Tipping (2001).

This common framework allows us to compare different models by representing them in terms of their “shrinkage profiles.” Assume for now that $\sigma^2 = \tau^2 = 1$, and define $\kappa_i = 1/(1 + \lambda_i^2)$. Then κ_i is a random shrinkage coefficient, and can be thought of as the weight that the posterior mean for β_i places on 0 once the needlet coefficients \mathbf{z} have been observed:

$$E(\beta_i | z_i, \lambda_i^2) = \left(\frac{\lambda_i^2}{1 + \lambda_i^2} \right) z_i + \left(\frac{1}{1 + \lambda_i^2} \right) 0 = (1 - \kappa_i) z_i.$$

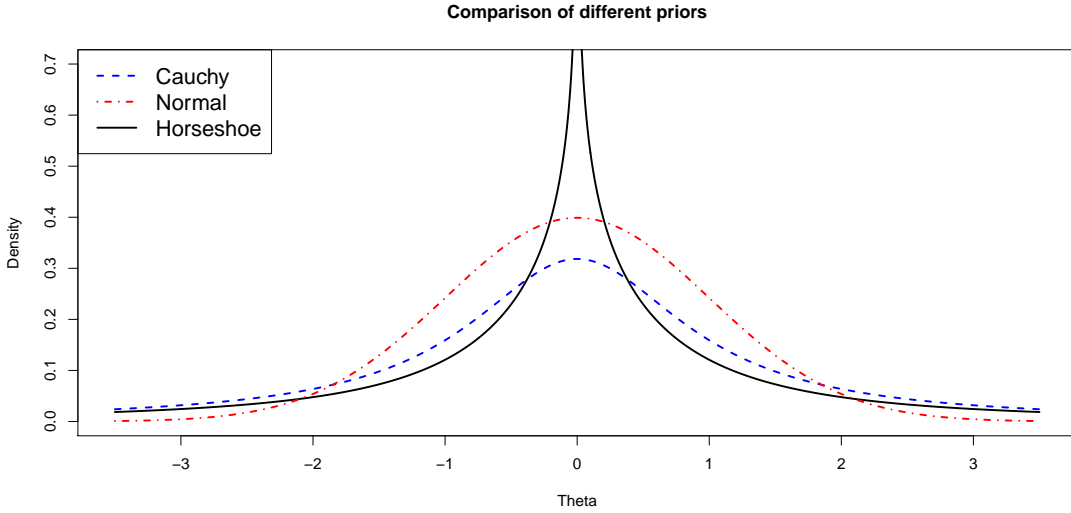


Figure 2: The horseshoe prior and two common priors: Normal and Cauchy.

Since $\kappa_i \in [0, 1]$, this is clearly finite, and so

$$\begin{aligned}
 E(\beta_i | y) &= \int_0^1 (1 - \kappa_i) z_i \pi(\kappa_i | \mathbf{z}) d\kappa_i \\
 &= \{1 - E(\kappa_i | \mathbf{z})\} z_i.
 \end{aligned} \tag{8}$$

By applying this transformation and inspecting the priors on κ_i implied by different choices for $\pi(\lambda_i)$, we can better appreciate how these models attempt to discern between signal and noise. Choosing $\lambda_i \sim C^+(0, 1)$ implies $\kappa_i \sim \text{Be}(1/2, 1/2)$, a density that is symmetric and unbounded at both 0 and 1. This horseshoe-shaped shrinkage profile “expects” to see two things *a priori*: strong signals ($\kappa \approx 0$, no shrinkage), and zeros ($\kappa \approx 1$, total shrinkage).

No other commonly used proper shrinkage prior shares these features. The double-exponential prior tends to a fixed constant near $\kappa = 1$, and disappears entirely near $\kappa = 0$. The Student- t prior and the Strawderman–Berger “pseudo-Cauchy” prior (see Johnstone and Silverman (2004)) are both unbounded near $\kappa = 0$, reflecting their heavy tails. But both are bounded near $\kappa = 1$, limiting these priors’ ability to squelch noise back to zero.

One approach that resembles the horseshoe is the normal–Jeffreys prior used by Figueiredo (2003) and Bae and Mallick (2004), where each local variance term has the improper Jeffreys prior, $\pi(\lambda_i) \propto 1/\lambda_i$. The normal–Jeffreys mixture is the improper limit of a proper Beta(ϵ, ϵ) prior for κ_i as $\epsilon \rightarrow 0$, and therefore also produces a horseshoe-like shape for $\pi(\kappa_i)$. But this prior leads to an improper joint posterior for β , meaning that the posterior mean—the Bayes estimator under quadratic loss—is undefined. It also does not allow adaptivity through the global parameter τ , since it is explicitly constructed to be free of hyperparameters. This additional aspect of “global shrinkage” distinguishes the horseshoe estimator from the approaches described in Tipping (2001) and Figueiredo (2003).

The authors in Carvalho et al. (2009) study the horseshoe prior in traditional problems of regression and function estimation, and find that it has a number of advantages over common alternatives:

- It is highly adaptive to different patterns of sparsity. Similar concerns are identified by

Scott and Berger (2006); these concerns about multiplicity arise when basis elements are tested indiscriminately, and the fact that they are handled automatically through data-based adaptation of τ is a big advantage.

- It is tail-robust, in the sense that large deviations from zero will remain unshrunk regardless of how small τ is estimated to be by the data.
- It is highly computationally efficient, since the prior admits closed-form expressions for posterior moments when τ is fixed.
- It performs very similarly to the gold standard of Bayesian model averaging over different β_i begin zero. It does so, however, while avoiding the computational difficulties associated with calculating marginal likelihoods and exploring an enormous discrete model space.
- It is proper, and therefore ensures a proper posterior.

3.3 Model fitting

Two computational strategies for fitting this model are available: importance sampling and a hybrid slice-sampling/Gibbs sampling approach. The first approach has the advantage that posterior moments can be computed without having to worry about potential MCMC convergence issues. The second has the advantage that it generates draws from the full posterior distribution of all model parameters, and so allows straightforward assessments of uncertainty with respect to features of the underlying random field. In practice, the proposed slice-sampler works very well, and this is the algorithm used to compute the results of the next section. Throughout, we use Jeffreys's prior for the sampling variance σ^2 .

3.3.1 Importance sampling

Given the global scale parameter τ , the posterior moments for β_i under the horseshoe prior can be expressed in terms of hypergeometric functions. After a change of variables to $\kappa_i = 1/(1+\lambda_i^2)$ and some straightforward algebra, the posterior can be verified as

$$\mathbb{E}(\beta_i | z_i, \tau) = \left\{ 1 - \frac{2 \Phi_1(1/2, 1, 5/2, z_i^2/2\sigma^2, 1 - 1/\tau^2)}{3 \Phi_1(1/2, 1, 3/2, z_i^2/2\sigma^2, 1 - 1/\tau^2)} \right\} z_i,$$

where Φ_1 is the degenerate hypergeometric function of two variables (Gradshteyn and Ryzhik, 1965, 9.261). And by the law of total variance,

$$\begin{aligned} \text{Var}(\beta_i | z_i, \tau) &= \mathbb{E}\{\text{Var}(\theta_i | z_i, \lambda_i^2, \tau)\} + \text{Var}\{\mathbb{E}(\theta_i | z_i, \lambda_i^2, \tau)\} \\ &= \sigma^2 \left\{ 1 - \frac{2 \Phi_1(1/2, 1, 5/2, z_i^2/2\sigma^2, 1 - 1/\tau^2)}{3 \Phi_1(1/2, 1, 3/2, z_i^2/2\sigma^2, 1 - 1/\tau^2)} \right\} \\ &\quad + z_i^2 \frac{8 \Phi_1(1/2, 1, 7/2, z_i^2/2\sigma^2, 1 - 1/\tau^2)}{15 \Phi_1(1/2, 1, 3/2, z_i^2/2\sigma^2, 1 - 1/\tau^2)}, \end{aligned} \tag{9}$$

will all other posterior moments for θ_i following similar expressions. See Gordy (1998) for details of computations involving the Φ_1 function.

The computation of posterior means and variances of the needlet coefficients under the horseshoe prior can therefore be reduced to a simple one-dimensional integral. Importance sampling is a natural approach. This requires one final ingredient, namely the marginal density

of the data \mathbf{z} given τ . Luckily this is readily computed:

$$p(\mathbf{z} | \tau) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{z_i^2}{2\sigma^2}\right) \frac{\text{Be}(1, 1/2)}{\text{Be}(1/2, 1/2)} \frac{\Phi_1(1/2, 1, 3/2, z_i^2/2\sigma^2, 1 - 1/\tau^2)}{\Phi_1(1/2, 1, 1, 0, 1 - 1/\tau^2)}. \quad (10)$$

After making the transformation $\xi = \log \tau$ to remove the domain restriction, the marginal posterior for ξ is

$$p(\mathbf{z}) = \int p(\mathbf{z} | \xi) \frac{2e^\xi}{\pi(1 + e^{2\xi})} d\xi,$$

recalling that a half-Cauchy prior has been assumed for τ .

Importance sampling proceeds by first computing $\hat{\xi}$ and \hat{s}_ξ^2 , the posterior mode and inverse second derivative at the mode, using a numerical optimization routine. Then a Student- t distribution with 4 degrees of freedom, centered at $\hat{\xi}$ and with scale parameter $a\hat{s}_\xi$, is used to generate proposals $\xi_{m=1}, \dots, \xi_{m=T}$. Here a is a tuning parameter used to control the scale of the proposal distribution and ensure a set of healthy importance weights. Values of $a \approx 3$ seem to work well in practice.

Posterior moments are then estimated as

$$\hat{h} \approx \frac{1}{T} \sum_{m=1}^T h(\xi_m) \cdot \frac{p(\mathbf{z} | \xi_m) \pi(\xi_m)}{t_3(\xi_m | \hat{\xi}, \hat{s}_\xi)},$$

where h is the posterior quantity of interest, such as a mean or a variance.

This approach is very appealing, because the same set of importance samples can be used for computing the posterior means and variances for all empirical needlet coefficients in parallel. This greatly streamlines the computation. One difficulty that sometimes arises is that, for extreme values of $\xi = \log \tau$, the Φ_1 function may become slow to evaluate. The issue seems to be particularly acute when τ is very close to zero. This difficulty can be alleviated, however, using the approximations to hypergeometric functions to be found in Butler and Wood (2002). These seem to give reasonable answers in the situations investigated so far, but may break down in more extreme circumstances.

We have sketched out the approach for the case of unknown τ , but typically σ^2 (which represents “noise variance” in the multipole domain) is also unknown. The method given above is easily modified to incorporate a bivariate importance function in ξ and $\phi = \log \sigma^2$. A multivariate- t or other similarly heavy-tailed density will work well here, with the inverse Hessian matrix at the mode specifying the covariance structure.

3.3.2 Markov-chain Monte Carlo

As a second option, Markov-chain Monte Carlo may be used to generate draws from the full joint posterior distribution of all model parameters. Simple Gibbs updates are available for the global variance components σ^2 and τ^2 , and are discussed in Gelman (2006). Also, it is clear that $(\beta_i | \tau, \lambda_i, z_i) \sim N(m, V)$, where

$$\begin{aligned} V &= \sigma^2 \left\{ \frac{\tau^2 \lambda_i^2}{1 + \tau^2 \lambda_i^2} \right\} \\ m &= Vz_i / \sigma^2 \end{aligned}$$

The chief difficulty is in efficiently sampling the local variance components λ_i^2 , given all other model parameters. We use the following slice-sampling approach, adapting an algorithm described by Damien et al. (1999). Define $\eta_i = 1/\lambda_i^2$, and define $\mu_i = \beta_i/(\sigma\tau)$. Then the

conditional posterior distribution of η_i , given all other model parameters, looks like

$$p(\eta_i \mid \tau, \sigma, \mu_i) \propto \exp \left\{ -\frac{\mu_i^2}{2} \eta_i \right\} \frac{1}{1 + \eta_i}.$$

Therefore, the following two steps are sufficient to sample λ_i :

1. Sample $(u_i \mid \eta_i)$ uniformly on the interval $(0, 1/(1 + \eta_i))$.
2. Sample $(\eta_i \mid \mu_i, u_i) \sim \text{Ex}(2/\mu_i^2)$ from an exponential density, truncated to have zero probability outside the interval $[0, (1 - u_i)/u_i]$.

Transforming back to the λ -scale will yield a draw from the desired conditional distribution. Ergodic averages of the draws for β_i are then used to estimate posterior means.

4 Examples

4.1 Simulated data

The above features make the horseshoe estimator an attractive choice for de-noising empirical needlet coefficients. We now describe a set of experiments that benchmark its performance on simulated data spanning a range of different sparsity patterns in the needlet domain.

First, we pixelized the sphere into 768 equal-area pixels with pixel centers x_i where observations y_i will be located. We then chose needlet meshes Ξ_j for $j = 0, \dots, 4$ of sizes $M_j = (12, 48, 192, 768, 3072)$, following NASA's standard hierarchical equal-area pixelization scheme for CMB data. The coefficients β_{jk} were simulated according to

$$\beta_{jk} \sim w \cdot t_{1.5} \left(\sqrt{9/M_j} \right) + (1 - w)\delta_0,$$

a sparse mixture of a point mass at zero and a Student- t density with 1.5 degrees of freedom and scale parameter $\sqrt{9/M_j}$, which puts the coefficients on the same scale as (5). The mixing ratio w encodes the signal density in the needlet domain. The target random field at points x_i was then calculated as $f(x_i) = \sum_j \sum_k \beta_{jk} \psi_{jk}(x_i)$, and $y_i = f(x_i) + \epsilon_i$ for $\epsilon_i \sim \text{N}(0, \sigma^2 = 9)$.

The simulated observations $y(x_i)$ were then used to reconstruct f using five alternatives:

1. The straight harmonic estimator in (1) using cubature-based estimates $\hat{\alpha}_l^m$.
2. The straight needlet estimator using the $\hat{\alpha}_l^m$'s plugged into (4).
3. The empirical-Bayes thresholding procedure described in Johnstone and Silverman (2004), which uses a mixture of a Laplace prior and a point mass at zero to model the rescaled coefficients $\tilde{\beta}_{jk}$.
4. The horseshoe estimator on the $\tilde{\beta}_{jk}$'s, as described in the previous section.
5. The horseshoe estimator as above, but with the result thresholded to zero if the posterior mean of the shrinkage coefficient κ_i is larger than 0.5 (which is highly suggestive of noise).

Procedure 3, the empirical-Bayes thresholding estimator, makes for a state-of-the-art benchmark. While this procedure has not previously been applied in the needlet literature, it has been used with great success for shrinkage and selection of conventional wavelets. Indeed, a similar procedure was shown by Johnstone and Silverman (2005) to have many of the same properties of the horseshoe estimator for wavelet denoising, namely robustness and adaptivity to a wide range of sparsity patterns.

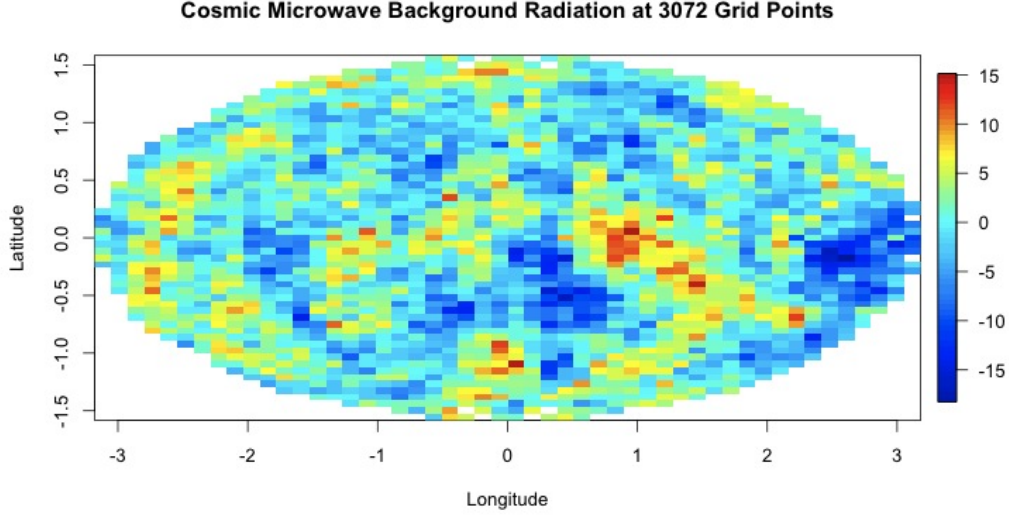


Figure 3: CMB temperature data (Mollweide projection).

Table 1: Simulated data. Sum of squared errors in reconstructing the needlet coefficients, $\ell^2(\beta, \hat{\beta})$, and average squared error per pixel in reconstructing the true random field, $\ell^2(f, \hat{f})/N$, for the five different procedures. The different values of w reflect the different sparsity patterns studied.

Procedure	$w = 0.25$		$w = 0.50$		$w = 0.75$	
	$\ell^2(\beta, \hat{\beta})$	$\ell^2(f, \hat{f})/N$	$\ell^2(\beta, \hat{\beta})$	$\ell^2(f, \hat{f})/N$	$\ell^2(\beta, \hat{\beta})$	$\ell^2(f, \hat{f})/N$
Straight needlet estimator	629	6.7	1682	7.4	1723	7.3
E-Bayes threshold	545	1.7	1596	3.0	1635	3.9
Horseshoe estimator	555	1.9	1610	3.1	1650	3.6
Thresholded horseshoe	563	1.7	1768	3.9	1661	3.6
Harmonic estimator	—	21.0	—	26.0	—	26.2

Table 2: Real data. Sum of squared errors in reconstructing the needlet coefficients, $\ell^2(\beta, \hat{\beta})$, and average squared error per pixel in reconstructing the true CMB temperature map, $\ell^2(f, \hat{f})/N$, for the four different procedures. The two values of σ reflect the two signal-to-noise ratios studied.

Procedure	$\sigma = 2$		$\sigma = 4$	
	$\ell^2(\beta, \hat{\beta})$	$\ell^2(f, \hat{f})/N$	$\ell^2(\beta, \hat{\beta})$	$\ell^2(f, \hat{f})/N$
Straight needlet estimator	40	3.4	157	12.1
E-Bayes threshold	64	5.0	91	6.7
Horseshoe estimator	31	2.8	76	6.0
Harmonic estimator	—	11.0	—	37.9

Table 1 shows these results on 100 simulated data sets for each of three different sparsity patterns.

4.2 Reconstruction of Noisy CMB Radiation Data

For a second experiment, we used publicly available temperature data on the cosmic microwave background radiation collected by NASA’s Wilkinson Microwave Anisotropy Probe.¹ The full data set maps temperature at over 3 million pixels covering the entire sky. For the purpose of testing different methods for needlet shrinkage, we constructed a reduced data set of 3072 equal-area pixels, each of which encodes the average temperature for an area comprising 1024 nearby pixels from the full data set. A heatmap of this data is in Figure 3.

We used this temperature map as the true f , and its corresponding discrete needlet transform as the true set of β_{jk} ’s. We then simulated 50 noisy data sets for two different signal-to-noise ratios. This was done by drawing $\epsilon_i \sim N(0, \sigma^2)$, and setting $y_i = f_i + \epsilon_i$ for each grid point, $i = 1, \dots, 3072$. The standard deviation of the data was about 5, so in our two experiments, we set $\sigma = 2$ and $\sigma = 4$.

We again benchmarked the horseshoe estimator against the harmonic estimator, the straight needlet estimator, and empirical-Bayes thresholding. Table 2 summarizes these results.

4.3 Summary of Results

From these results, the following conclusions about needlet shrinkage can be observed:

- All needlet-based estimators are a drastic improvement upon the harmonic estimator. The harmonic estimator performs so poorly, even worse than the MLE, because the finest-resolution harmonics cannot be reliably reconstructed from the data. These noisy harmonics appear to affect the needlet estimator much less.
- Horseshoe shrinkage and selection of needlet coefficients offers further substantial improvements upon the straight needlet estimator, often by a factor of three or more. This seems to be true regardless of the pattern of sparsity and signal-to-noise ratio. Even when the straight estimator performs almost as well in the needlet domain, it always does much worse in the spatial domain, suggesting that the shrinkage procedures are better at reconstructing the coefficients with the most important contributions to the topography of f .
- On simulated data, the horseshoe estimator quite closely matches the performance of empirical-Bayes thresholding based on point-mass mixture priors. On real data, however, the horseshoe does much better. For example, in the “low-noise” version of the CMB experiment, the empirical-Bayes procedure is even beaten by the straight needlet estimator, which is itself beaten by the horseshoe.
- The thresholded version of the horseshoe rarely beats the unthresholded version, even when the true signal has zeros. This is consistent with the intuition that the horseshoe estimator behaves like model averaging, which is typically better than selecting a single model.

References

K. Bae and B. Mallick. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18):3423–30, 2004.

¹<http://lambda.gsfc.nasa.gov>

- P. Baldi, G. Kerkyacharian, D. Marinucci, and D. Picard. Subsampling needlet coefficients on the sphere. arXiv:0706.4169v1 [math.ST], 2007.
- P. Baldi, G. Kerkyacharian, D. Marinucci, and D. Picard. Adaptive density estimation for directional data using needlets. arXiv:0807.5059v1 [math.ST], 2008.
- P. Baldi, G. Kerkyacharian, D. Marinucci, and D. Picard. Asymptotics for spherical needlets. *The Annals of Statistics*, 37(3):1150–71, 2009.
- R. Butler and A. Wood. Laplace approximations for hypergeometric functions with matrix argument. *The Annals of Statistics*, 30:1155–77, 2002.
- C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. Discussion Paper 2008-31, Duke University Department of Statistical Science, 2008.
- C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research W&CP*, 5(73–80), 2009.
- M. Clyde and E. I. George. Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, Series B (Methodology)*, 62(4):681–98, 2000.
- P. Damien, J. C. Wakefield, and S. G. Walker. Bayesian nonconjugate and hierarchical models by using auxiliary variables. *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 61:331–44, 1999.
- X. Faucheur, B. Vidakovic, and A. Tannenbaum. Bayesian spherical wavelet shrinkage: applications to shape analysis. In *Proceedings of SPIE Optics*, 2007.
- M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1150–9, 2003.
- A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, 1(3): 515–33, 2006.
- M. B. Gordy. A generalization of generalized beta distributions. Technical report, Board of Governors of the Federal Reserve System, Finance and Economics Discussion Series, 1998.
- I. Gradshteyn and I. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, 1965.
- I. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical-Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.
- I. M. Johnstone and B. W. Silverman. Empirical Bayes selection of wavelet thresholds. *Ann. Statist.*, 33:1700–1752, 2005.
- D. Marinucci, D. Pietrobon, A. Balbi, P. Baldi, P. Cabella, G. Kerkyacharian, P. Natoli, D. Picard, and N. Vittorio. Spherical needlets for cosmic microwave background data analysis. *Monthly Notices of the Royal Astronomical Society*, 8(2):539–45, 2008.
- F. Narcowich, P. Petrushev, and D. Ward. Localized tight frames on spheres. *SIAM J. Math. Anal.*, 38:574–94, 2006.
- J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc. B.*, 58(1):267–88, 1996.
- M. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–44, 2001.
- B. Vidakovic and P. Muller. *Bayesian Inference in Wavelet Based Models*, chapter An introduction to wavelets. Springer–Verlag, 1999.