Bayesian Analysis (2011) **6**, Number 2, pp. 307–328

Bayesian Estimation of Intensity Surfaces on the Sphere via Needlet Shrinkage and Selection

James G. Scott*

Abstract. This paper describes an approach for Bayesian modeling 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, and Bayesian local shrinkage rules. We study the performance of the proposed methodology both on simulated data and on two real data sets: one involving the cosmic microwave background radiation, and one involving the reconstruction of a global news intensity surface inferred from published Reuters articles in August, 1996. The fully Bayesian approach based on robust, sparse shrinkage priors seems to outperform other alternatives.

Keywords: needlets, shrinkage estimate, spherical wavelets

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.

*McCombs School of Business, University of Texas at Austin, Austin, TX, mailto:james.scott@mccombs.utexas.edu

 DOI:10.1214/11-BA611

These 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. (2009b) 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 implies that needlets make an excellent choice of basis for statistical estimation of intensity surfaces 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 as 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 assumption is highly nontrivial, since there is undoubtedly overlap in the needlet kernels themselves. But it 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. For further discussion of the issues of asymptotic uncorrelation and coefficient dependence structure in the context of needlet analysis, see Mayeli (2010) and Lan and Marinucci (2009).

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 surprisingly nontrivial to ensure. It results from the careful mathematical construction of a "window function" 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. (2009c) and Baldi et al. (2009a). 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 describe a Bayesian modeling approach for robust shrinkage and selection of needlet coefficients. This method is derived from the horseshoe prior of Carvalho et al. (2010), and differs both from existing needlet methods based on thresholding, and from existing Bayesian methods for conventional wavelets.

2. We propose two algorithms for fitting models of this form, which are of general utility in Bayes-type shrinkage rules.

3. We investigate the need to properly scale the empirical needlet coefficients before any shrinkage procedure is applied, which can dramatically affect performance.

4. We study the problem of sparsity in the multipole domain.

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. (2009b). 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 kth 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

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

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

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

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

where $\overline{Y}$ 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

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

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 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:

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

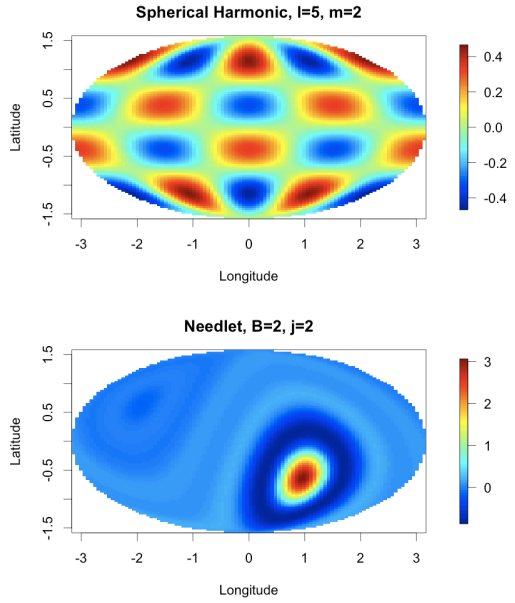

Figure 1: Top: spherical harmonic for $l = 5$, $m = 2$. Bottom: a needlet centered at $(\theta = \pi/3, \phi = -\pi/6)$ with $j = 2$ and $B = 2$. The sphere has been projected to $\mathbb{R}^2$ using the Mollweide projection.

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

$$f(x) = \sum_{j} \sum_{k=1}^{M_j} \beta_{jk} \; \psi_{jk}(x) \, . \tag{4}$$

The coefficients β_{jk} are given by

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

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 $l = \lfloor \delta^{j-1} \rfloor, \ldots, \lceil \delta^{j+1} \rceil$ 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 $\ell_{\max}$:

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

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 $\widehat{\beta}_{jk}$ are then computed by plugging the harmonic coefficients $\widehat{\alpha}_l^m$ into (5), yielding

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

The focus of this paper is on improving the straight cubature estimator through shrinkage and selection of the empirical needlet coefficients $\widehat{\beta}_{jk}$. 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,

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

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

2.3 Heteroskedasticity in the multipole domain likelihood

Statistical learning of needlet coefficients must 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. This factor is a cubature weight associated with each basis element, thereby ensuring that the total surface area of the sphere normalizes to 4π.

But from a pure estimation standpoint, it is not obvious how much of this factor to attribute to the function ψ_{jk}, and how much to the coefficient β_{jk}, since only the product of these two quantities is statistically identifiable. A similar issue appears in wavelet shrinkage; see, for example, Vidakovic and Müller (1999). In needlet modeling, however, the scale is much harder to determine, because needlets are not constructed in the same "dilate and shift" manner as wavelets. This operation creates a natural hierarchy of scales, and therefore a natural set of variances associated with each term in the wavelet-domain likelihood.

It is therefore important to account for heteroskedasticity in the multipole domain (the β_{jk}'s), where the likelihood is defined. 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

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

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 $\widehat{\beta}_{jk}$ 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 $\widehat{\beta}_{jk}$. 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. This is the scale, therefore, that will be adopted throughout the rest of the paper, though it remains an open question whether there is any more fundamental sense in which (6) gives the correct scaling.

An important caveat here is that we are essentially describing a pre-processing step that ensures the assumption of a homoskedastic likelihood is statistically sensible. This pre-processing generates a set of rescaled needlet coefficients, to which a shrinkage procedure can then be applied. This re-scaling can then be reversed easily prior to applying the reverse needlet transform, which is needed to reconstruct the function. This is an important point, since the standard normalization of needlet coefficients ensures some important mathematical features of the resulting random field (most especially, finite variance). We do not propose that this fundamental scale of normalization be modified, except merely as an intermediate and easily undone step prior to the application of a thresholding or shinkage rule.

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 $\widetilde{\beta}_j$ be the vector of M_j needlet coefficients for level j, and stack these rescaled coefficients into a single p-dimensional column vector $\boldsymbol{z} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_{j_{\max}})'$. We will treat the vector $\boldsymbol{z}$ as raw data observed with Gaussian error, $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2 I)$. 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

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

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 of being zero is greater than 50%.

Johnstone and Silverman (2004) demonstrate that this highly adaptive estimator does very well at reconstructing sparse signals, proving that it 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. (2010).

Specifically, we reconstruct the underlying coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ using the posterior mean under the horseshoe prior. The horseshoe prior assumes that each β_i

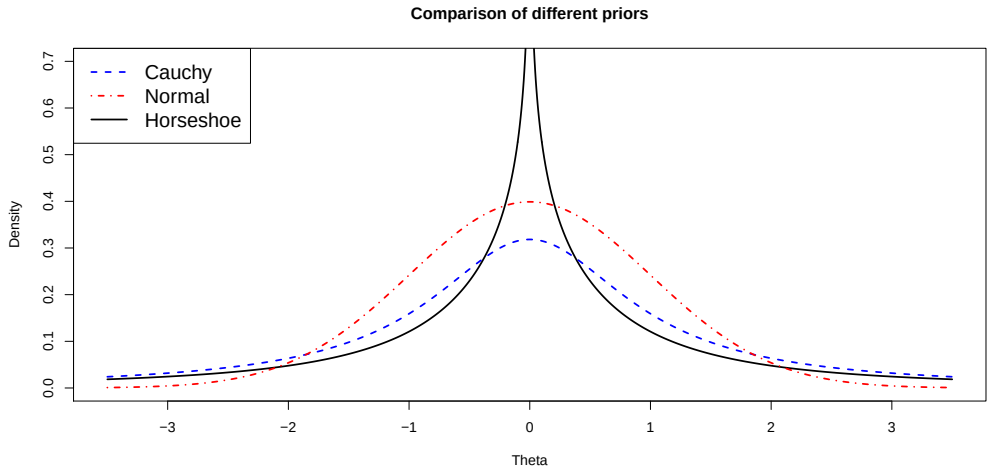

Figure 2: The horseshoe prior and two common priors: Normal and Cauchy.

has density $\pi_{HS}(\beta_i \mid \tau)$, with π_{HS} expressible as a scale mixture of normals:

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

where $\mathrm{C}^+(0,1)$ 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. (2010). 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 $\lambda_i^2 \sim \mathrm{Exp}(2)$, for instance, implies independent double-exponential priors for each β_i; inverse-gamma mixing, with $\lambda_i^2 \sim \mathrm{IG}(a,b)$, 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).

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 combinations of β_i being in or out of the model. 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.

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 $\epsilon \to 0$, and therefore also produces a horseshoe-like shape for $\pi(\kappa_i)$. But this prior leads to an improper joint posterior for $\boldsymbol{\beta}$, meaning that the posterior mean—the Bayes estimator under quadratric 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).

3.3 The score function and overshrinkage of exceptional observations

This section summarizes the theoretical argument as to why priors of the form described above exhibit the desirable property of "Bayesian robustness." We recall the following theorem from Carvalho et al. (2010).

Theorem 3. *Let $p(|y - \beta|)$ be the likelihood, and suppose that $p(\beta)$ is a mean-zero scale mixture of normals: $(\beta \mid \lambda) \sim N(0, \lambda^2)$, with λ having proper prior $p(\lambda)$. Assume further that the likelihood and $p(\beta)$ are such that the marginal density $m(y) < \infty$ for all y. Define the following three pseudo-densities, which may be improper:*

$$\begin{aligned} m^\star(y) &= \int_{\mathbb{R}} p(|y-\beta|)\; p^\star(\beta)\; d\beta\,, \\ p^\star(\beta) &= \int_{\mathbb{R}^+} p(\beta \mid \lambda)\; p^\star(\lambda)\; d\lambda\,, \\ p^\star(\lambda) &= \lambda^2 p(\lambda)\,. \end{aligned}$$

Then

$$\begin{aligned} E(\beta \mid y) &= \frac{m^\star(y)}{m(y)} \frac{d}{dy} \log m^\star(y) \\ &= \frac{1}{m(y)} \frac{d}{dy} m^\star(y) \,. \end{aligned} \tag{8}$$

Versions of this representation theorem appear in Masreliez (1975), Polson (1991), and Pericchi and Smith (1992). Theorem 3 relaxes a specific regularity condition having to do with the boundedness of $p(\beta)$, and extends the usual result to situations where $p(\beta)$ is a scale mixture of normals with proper mixing density and finite marginal $m(y)$.

The theorem characterizes the behavior of an estimator in the presence of large signals. Specifically, it says that we can achieve inherent Bayesian robustness by choosing a prior for β such that the derivative of the log predictive density is bounded as a function of y. Ideally, of course, this bound should converge to 0 for large $|y|$, and will lead to $\mathrm{E}(\beta \mid y) \approx y$ for large $|y|$. This will avoid the overshrinkage of exceptional observations, such as those that might arise from the highly localized character of the functions which needlet expansions are designed to describe.

The following result demonstrates that the horseshoe prior is a member of a broader class of priors with redescending score functions, and is therefore tail robust. It is a broader result than Theorem 3 of Carvalho et al. (2010), in that it explicitly characterizes the tail weight of the marginal density in terms of the tail weight of the prior for the variance λ^2.

Theorem 4. *Suppose that* $(y \mid \beta) \sim N(\beta, 1)$*, and that the prior for* β *is a scale mixture:* $\pi(\beta) = \int N(\beta \mid 0, \lambda^2)\pi(\lambda^2)d\lambda^2$*. Suppose that* $\pi(\lambda^2)$ *is of asymptotic order* $(\lambda^2)^{a-1}$ *as* $\lambda^2 \to \infty$*. Then as* $y \to \infty$*, the predictive density* $m(y) = \int N(y \mid \beta, 1)\pi(\beta)d\beta)$ *is of asymptotic order* $(a-1)/y$*.*

Proof

Write the likelihood as $(y \mid z) \sim \mathrm{N}(0, z)$, where $z = 1 + \lambda^2$ with induced prior $\pi(z)$. If $\pi(\lambda^2)$ satisfies the tail condition of the theorem, then clearly so will $\pi(z)$:

$$\pi(z) = Cz^{a-1} \;\text{ as }\; z \to \infty \,,$$

for some constant C. The marginal likelihood is therefore also a scale mixture of normals,

$$m(y) = \int_1^\infty \frac{1}{\sqrt{2\pi z}} e^{-\frac{y^2}{2z}} \pi(z) \mathrm{d}z \,.$$

Following Theorem 6.1 of Barndorff-Nielsen et al. (1982), this expression's asymptotic order can be related to the tail weight of $\pi(z)$, and will simply be $|y|^{2a-1}$ as $|y| \to \infty$. The form of the score function then follows.

The immediate corollary is that any scale-mixture prior where $p(\lambda^2)$ has polynomial tails leads to a redescending score function.

3.4 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, we have found that the proposed slice-sampler provides stable results with no readily apparent MCMC convergence problems, and this is the algorithm we have used to compute the results of the next section. Throughout, we use Jeffreys' prior for the sampling variance σ^2.

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 mean is

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

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

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

with 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 $\boldsymbol{z}$ given τ. Luckily this is readily computed:

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

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

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

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

Importance sampling proceeds by first computing $\hat{\xi}$ and $\hat{s}_\xi^2$, 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 $\xi_{m=1}, \ldots, \xi_{m=T}$. Here a is a tuning parameter used to control the scale of the proposal distribution. In repeated applications of the method, values of $a \approx 3$ generally provided sufficient coverage of the posterior distribution so that no small group of importance weights dominated the calculation.

Posterior moments are then estimated as

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

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

This approach is appealing in that 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).

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 $\phi = \log \sigma^2$. In this case, a multivariate-t or other similarly heavy-tailed density can be used as a proposal distribution for the importance sampler, with the inverse Hessian matrix at the mode specifying the covariance structure. This can be implemented using a package routine for numerical optimization that is capable of returning gradient information.

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 $(\beta_i \mid \tau, \lambda_i, z_i) \sim \mathrm{N}(m, V)$, where

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

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, is

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

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 $(\eta_i \mid \mu_i, u_i) \sim \text{Exp}(2/\mu_i^2)$ 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 Performance on benchmark 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. It should be emphasized that the goal here is not merely the reconstruction of isotropic random fields, which has been the focus in much of the previous literature. Rather, we are interested in the use of needlets as a set of basis functions for reconstructing any form of spatial intensity surface on the sphere, where only noisy observations are available.

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 $j = 0, \ldots, 4$ 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 $\epsilon_i \sim \text{N}(0, \sigma^2)$, $\sigma^2 = 9$.

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

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

	$w = 0.25$		$w = 0.50$		$w = 0.75$	
Procedure	$\ell^2(\beta, \widehat{\beta})$	$\ell^2(f, \widehat{f})/N$	$\ell^2(\beta, \widehat{\beta})$	$\ell^2(f, \widehat{f})/N$	$\ell^2(\beta, \widehat{\beta})$	$\ell^2(f, \widehat{f})/\mathrm{N}$
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

1. The straight harmonic estimator in (1) using cubature-based estimates $\widehat{\alpha}_l^m$.
2. The straight needlet estimator using the $\widehat{\alpha}_l^m$'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 $\widetilde{\beta}_{jk}$.
4. The horseshoe estimator on the $\widetilde{\beta}_{jk}$'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.

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

[1] http://lambda.gsfc.nasa.gov

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

	$\sigma = 2$		$\sigma = 4$	
Procedure	$\ell^2(\beta, \widehat{\beta})$	$\ell^2(f, \widehat{f})/N$	$\ell^2(\beta, \widehat{\beta})$	$\ell^2(f, \widehat{f})/\mathrm{N}$
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

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 $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, and setting $y_i = f_i + \epsilon_i$ for each grid point, $i = 1, \ldots, 3072$. 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 simulation 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 affect the needlet estimator much less drastically.

- Horseshoe shrinkage and selection of needlet coefficients offers further substantial improvements upon the straight needlet estimator, often by a factor of three of more. This happens 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 does much worse in the spatial domain, suggesting that the shrinkage procedures are better at reconstructing the coefficients with the most important contributions

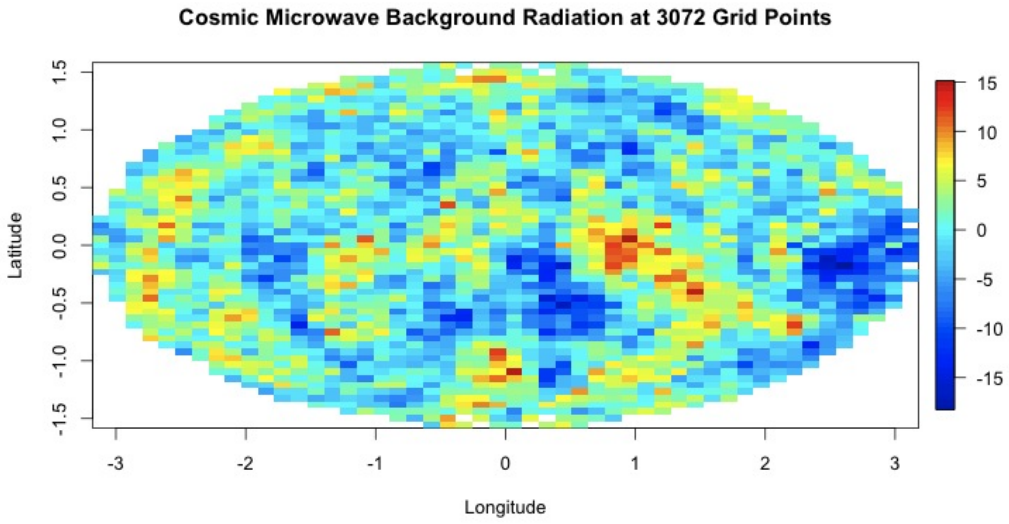

Figure 3: CMB temperature data (Mollweide projection).

to the topography of f.

- On simulated random fields, the horseshoe estimator quite closely matches the performance of empirical-Bayes thresholding based on point-mass mixture priors. When noise is added to a real field, however, the horseshoe does much better at reconstructing the ground truth. 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.

One point worth emphasizing is that, in principle, the discrete needlet transform and the spherical harmonic expansion should be identical. We would indeed see such behavior in the limiting case of infinite resolution—that is, if we were able to fit the function as a convergent weighted sum of all spherical harmonics ${Y_l}^m$, $l \in \mathbb{N}$. Any differences arise only due to truncation errors. But such errors are impossible to avoid in practice: one must inevitably choose a maximum resolution for the purpose of fitting a particular data set, and extra resolution is computationally expensive.

4.4 Estimating a news intensity surface over the globe

As a final illustration of the proposed methodology, we estimated a set of needlet coefficients that depict a news intensity surface over the globe. The data set is referred to as TR-CoNLL, and is a standard reference corpus in natural language processing. TR-CoNLL contains all 946 news articles filed anywhere in the world by the Reuters news-wire service for the period spanning August 22–31, 1996. From these 946 articles, 6,980 place-names have been annotated using the toponym-resolution algorithms of Leidner (2008). Toponym resolution is the task of mapping a set of ambiguous place names to the actual physical coordinates of the places they refer to—for example, determining automatically from context whether a reference to London refers to Ontario or England.

The spatial distribution of place-names mentioned by news articles in a given week can be thought of as a global news intensity surface, depicting where important and newsworthy events were happening. For example, here is an excerpt from one of the news stories, dated 24 August 1996 and filed in Islamabad, describing devastating floods in Pakistan:

> At least 30 people have been killed and about 100 injured in the flood-hit Pakistani city of Lahore, newspapers reported on Saturday. They said 461 mm (18 inches) of rain had drenched the Punjab provincial capital in 36 hours, turning streets into rivers, knocking out power, water and telephone services, disrupting air and rail traffic, and sweeping away houses and cars.

Both "Lahore" and "Punjab" appear as entries in the data set, associated with specific latitude/longitude coordinates.

Figure 4 shows the raw data, where each dot represents a single disambiguated toponym. A group of 20 news stories all describing the same location—suggesting a spate of interesting news—would show up as a single dot, obfuscating the importance of that area. But by estimating a news intensity surface, we allow nearby dots to mutually amplify the visual importance of a given region.

Figure 5 shows the results of fitting the harmonic estimator and the needlet shrinkage estimator proposed in this paper. As the figures show, the harmonic estimator exhibits a highly spurious non-locality, in the form of mild undulations in the estimated field. (This is represented by the alternating blue and white regions, and can most easily be seen in the middle of the oceans.) This behavior arises from the spatially non-local character of spherical harmonics, and makes it clear why this set of basis functions is inappropriate for describing surfaces whose behavior is highly localized in space. Intuitively, in order to get zeros out in the oceans (where there is rarely any news) using a harmonic basis, one needs a large sum of harmonics to cancel to exactly zero. This happens in the limit of infinite resolution, but not in practice where estimates are subject to truncation error.

No such nonlocalities exist in the fitted intensity surface corresponding to the shrinkage estimator. In essence, the needlet coefficients loading in these areas are treated as noise, and are shrunk almost completely to zero. This example conveys the intuition of

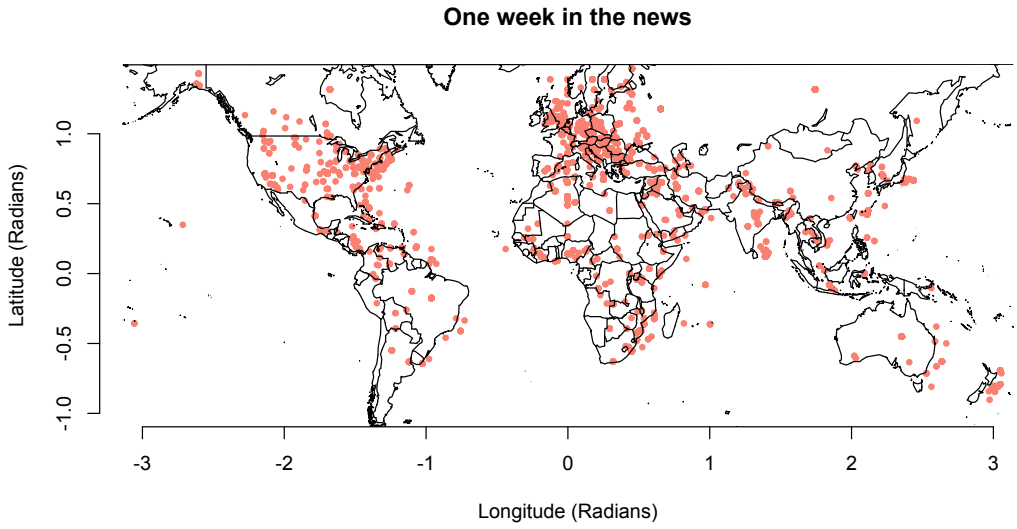

Figure 4: The disambiguated physical coordinates associated with place names in the TR-CoNLL reference corpus, comprising Reuters news filings in late August, 1996.

why needlet-based shrinkage rules may perform better for many spherical data sets, and why much attention needs to be paid to the character and performance of the shrinkage rule itself.

One caveat is that the needlet estimator, while clearly superior to the harmonic estimator, still resulted in a noticeable degree of overshrinkage, with many smaller features (e.g. those in South America and the Carribean) shrunk by a considerable factor. This is likely due to the highly localized nature of the intensity surface—many essentially zero areas, and many areas of highly concentrated newsworthy events—coupled with the fact that needlet basis functions are not orthogonal. One alternative approach, not explored here, would involve orthogonalizing the design matrix using a singular value decomposition, and applying the shrinkage estimator in the orthogonal space instead.

References

Bae, K. and Mallick, B. (2004). "Gene selection using a two-level hierarchical Bayesian model." *Bioinformatics*, 20(18): 3423–30. 315

Baldi, P., Kerkyacharian, G., Marinucci, D., and Picard, D. (2009a). "Adaptive density estimation for directional data using needlets." *The Annals of Statistics*, 37(6A): 3362–95. 308

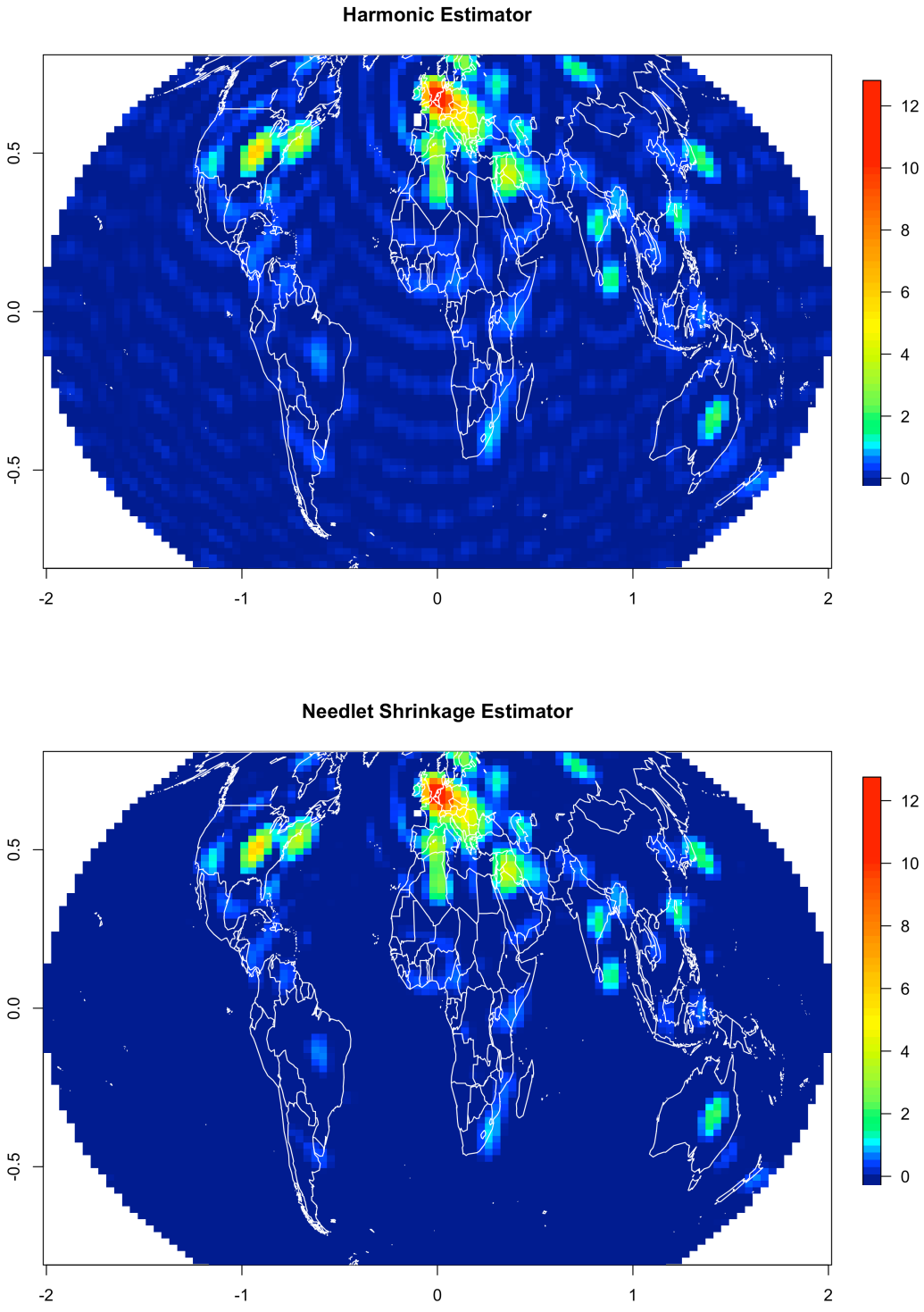

Figure 5: Estimated news intensity surfaces for the TR-CoNLL data set.

— (2009b). "Asymptotics for Spherical Needlets." *The Annals of Statistics*, 37(3): 1150–71. 308, 309

— (2009c). "Subsampling Needlet Coefficients on the Sphere." *Bernoulli*, 15(2): 438–63. ArXiv:0706.4169v1 [math.ST]. 308

Barndorff-Nielsen, O., Kent, J., and Sorensen, M. (1982). "Normal variance-mean mixtures and z distributions." *International Statistical Review*, 50: 145–59. 316

Butler, R. and Wood, A. (2002). "Laplace Approximations for Hypergeometric Functions With Matrix Argument." *The Annals of Statistics*, 30: 1155–77. 318

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling Sparsity via the Horseshoe." *Journal of Machine Learning Research: Workshops and Case Proceedings*, 5: 73–80. 314

— (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–80. 308, 313, 314, 315, 316

Clyde, M. and George, E. I. (2000). "Flexible Empirical Bayes Estimation for Wavelets." *Journal of the Royal Statistical Society, Series B (Methodology)*, 62(4): 681–98. 308

Damien, P., Wakefield, J. C., and Walker, S. G. (1999). "Bayesian nonconjugate and hierarchical models by using auxiliary variables." *J. R. Stat. Soc. Ser. B, Stat. Methodol.*, 61: 331–44. 319

Faucheur, X., Vidakovic, B., and Tannenbaum, A. (2007). "Bayesian spherical wavelet shrinkage: applications to shape analysis." In Truchetet, F. and Laligant, O. (eds.), *Wavelet Applications in Industrial Processing V*, volume 6763. International Society for Optics and Photonics (SPIE). 308

Figueiredo, M. (2003). "Adaptive sparseness for supervised learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9): 1150–9. 315

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models." *Bayesian Anal.*, 1(3): 515–33. 318

Gordy, M. B. (1998). "A generalization of generalized beta distributions." Finance and Economics Discussion Series 1998-18, Board of Governors of the Federal Reserve System (U.S.). 317

Gradshteyn, I. and Ryzhik, I. (1965). *Table of Integrals, Series, and Products*. Academic Press. 317

Johnstone, I. and Silverman, B. W. (2004). "Needles and Straw in Haystacks: Empirical-Bayes estimates of possibly sparse sequences." *The Annals of Statistics*, 32(4): 1594–1649. 313, 320

Johnstone, I. M. and Silverman, B. W. (2005). "Empirical Bayes Selection Of Wavelet Thresholds." *Ann. Statist*, 33: 1700–1752. 320

Lan, X. and Marinucci, D. (2009). "On the dependence structure of wavelet coefficients for spherical random fields." *Stochastic Processes and their Applications*, 119(10): 3749–66. 308

Leidner, J. L. (2008). *Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Universal Press. 323

Marinucci, D., Pietrobon, D., Balbi, A., Baldi, P., Cabella, P., Kerkyacharian, G., Natoli, P., Picard, D., and Vittorio, N. (2008). "Spherical needlets for cosmic microwave background data analysis." *Monthly Notices of the Royal Astronomical Society*, 8(2): 539–45. 308, 310, 312

Masreliez, C. (1975). "Approximate non-Gaussian filtering with linear state and observation relations." *IEEE. Trans. Autom. Control*, 20(1): 107–10. 316

Mayeli, A. (2010). "Asymptotic uncorrelation for Mexican needlets." *Journal of Mathematical Analysis and Applications*, 363(1): 336–44. 308

Narcowich, F., Petrushev, P., and Ward, D. (2006). "Localized tight frames on spheres." *SIAM J. Math. Anal.*, 38: 574–94. 307, 309

Pericchi, L. R. and Smith, A. (1992). "Exact and Approximate Posterior Moments for a Normal Location Parameter." *Journal of the Royal Statistical Society (Series B)*, 54(3): 793–804. 316

Polson, N. G. (1991). "A representation of the posterior mean for a location model." *Biometrika*, 78: 426–30. 316

Scott, J. G. and Berger, J. O. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136(7): 2144–2162. 315

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *J. Royal. Statist. Soc B.*, 58(1): 267–88. 314

Tipping, M. (2001). "Sparse Bayesian learning and the relevance vector machine." *Journal of Machine Learning Research*, 1: 211–44. 314, 315

Vidakovic, B. and Müller, P. (1999). *Bayesian Inference in Wavelet Based Models*, chapter An introduction to wavelets. Springer–Verlag. 311

Acknowledgments

The author wishes to thank two anonymous referees; the associate editor; the editor; and Kary Myers, the copy editor, for their many helpful suggestions for improving the article.