BAYESIAN STATISTICS 8, pp. 1–24.
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)

Nonparametric Function Estimation Using Overcomplete Dictionaries

Merlise A. Clyde and Robert L. Wolpert
Duke University, U.S.A.
clyde@stat.duke.edu rlw@stat.duke.edu

Summary

We consider the problem of estimating an unknown function based on noisy data using nonparametric regression. One approach to this estimation problem is to represent the function in a series expansion using a linear combination of basis functions. Overcomplete dictionaries provide a larger, but redundant collection of generating elements than a basis, however, coefficients in the expansion are no longer unique. Despite the non-uniqueness, this has the potential to lead to sparser representations by using fewer non-zero coefficients. Compound Poisson random fields and their generalization to Lévy random fields are ideally suited for construction of priors on functions using these overcomplete representations for the general nonparametric regression problem, and provide a natural limiting generalization of priors for the finite dimensional version of the regression problem. While expressions for posterior modes or posterior distributions of quantities of interest are not available in closed form, the prior construction using Lévy random fields permits tractable posterior simulation via a reversible jump Markov chain Monte Carlo algorithm. Efficient computation is possible because updates based on adding/deleting or updating single dictionary elements bypass the need to invert large matrices. Furthermore, because dictionary elements are only computed as needed, memory requirements scale linearly with the sample size. In comparison with other methods, the Lévy random field priors provide excellent performance in terms of both mean squared error and coverage for out-of-sample predictions.

Keywords and Phrases: Gaussian Random Field; Infinitely Divisible; Kernel Regression; Lévy random field; Nonparametric Regression; Relevance Vector Machine; Reversible Jump Markov chain Monte Carlo; Spatial-Temporal Models; Splines; Support Vector Machine; Wavelets.

Merlise Clyde is Associate Professor of Statistics at Duke University, Durham, North Carolina, USA. Robert Wolpert is Professor of Statistics at Duke University. The authors would like to thank Jen-hwa Chu, Leanna House, and Chong Tu for their contributions. This material is based upon work supported by the National Science Foundation under Grant Number DMS-0342172, DMS-0422400 and DMS-0406115. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

1. INTRODUCTION

The canonical setup for the nonparametric regression problem consists of having n measurements $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ of an unknown real valued function $f(\boldsymbol{x})$ defined on some space $\boldsymbol{\mathcal{X}}$,

$$Y_i = f(\boldsymbol{x}_i) + \epsilon_i \tag{1}$$

observed at points $\boldsymbol{x}_i \in \boldsymbol{\mathcal{X}}$. In the regression formulation the errors, ϵ_i, will typically represent white noise, $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$, but the nonparametric model may be extended to other exponential family models where $g(\mathrm{E}[Y_i]) = f(\boldsymbol{x}_i)$ for some link function g, as in generalized additive models. The function $f(\cdot)$ will often be regarded as an element of some separable Hilbert space $\mathcal{H}$ of real-valued functions on a compact space $\boldsymbol{\mathcal{X}}$. For Bayesian inference regarding the unknown mean function f, we must first place a prior distribution on f. If we are to model f nonparametrically, then we *should* place a prior distribution over the infinite dimensional space $\mathcal{H}$ of possible functions. However, in practice it is common to place a prior on the finite dimensional vector $\boldsymbol{f}_n \equiv (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))^T$ for $\boldsymbol{f}_n \in \mathbb{R}^n$, for example, by expressing $\boldsymbol{f}_n$ at the observed points $\boldsymbol{x}_i$ in terms of a finite dimensional basis and placing prior distributions only on the coefficients or coordinates of $\boldsymbol{f}_n$ with respect to the basis. While a class of priors in the finite dimensional version may lead to reasonable behaviour of posteriors with modest sample sizes, one would hope that the finite dimensional prior remains sensible in the infinite dimensional limit and as the sample size n increases. In this paper, we promote the use of Lévy random field priors for stochastic expansions of f and show how Lévy random fields provide a natural limiting extension of certain finite dimensional prior distributions. We begin in Section 2 by reviewing some of the popular choices of priors in the the finite dimensional version of the problem. In Section 3, we present priors for stochastic expansions of f using Lévy random fields and show how these priors arise as natural limits of certain prior distributions on finite dimensional spaces. The connection between Lévy random fields and Poisson random fields provides the key to tractable computation using (reversible jump) Markov chain Monte Carlo sampling for the stochastic expansions. In Section 4 we describe the resulting hierarchical model and discuss prior specifications. In Section 5 we discuss how the Lévy random field priors lead to penalized likelihoods and contrast these expressions with other model selection criteria. We highlight some of our applications of Lévy random fields in Section 6. For many problems, Lévy random fields provide an attractive alternative to Gaussian random field priors. We conclude by discussing some areas for future research.

2. PRIOR DISTRIBUTIONS ON FUNCTIONS

When it comes to placing prior distributions over nonparametric functions of explanatory variable(s) $\boldsymbol{x}$, Gaussian process (or random field) priors are perhaps the most accessible. If the function f has a Gaussian Process (GP) prior with mean μ and covariance function $\boldsymbol{\Sigma}(\cdot, \cdot; \boldsymbol{\theta})$ (a positive definite function on $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}$),

$$f(\cdot) \sim \mathrm{GP}\left(\mu, \boldsymbol{\Sigma}(\cdot, \cdot; \boldsymbol{\theta})\right) \tag{2}$$

then this implies that any finite dimensional vector $(f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))^T$ has an n dimensional multivariate normal distribution with mean μ and $n \times n$ covariance

matrix, $\mathrm{Cov}[f(\boldsymbol{x}_i), f(\boldsymbol{x}_j)] = \boldsymbol{\Sigma}(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$. The hyperparameter vector $\boldsymbol{\theta}$ control various features of the covariance function, and hence the process. While widely used in spatial-temporal modelling (Cressie, 1993, Chapter 3) and smoothing splines (Wahba, 1990), these priors are well suited for the general nonparametric regression and classification problems (O'Hagan, 1978; Neal, 1999). For Gaussian error models, Gaussian random field priors are particularly appealing as the unknown function f may be integrated out analytically, leaving the marginal likelihood of the (typically) much lower dimensional parameters $\boldsymbol{\theta}$ and σ^2. While sampling from the posterior distribution of the parameters $\boldsymbol{\theta}$ and σ^2 is generally straightforward using MCMC algorithms, implementation typically require repeated inversion of $n \times n$ matrices within the MCMC loop, limiting the applicability to modest n.

One may avoid matrix inversions by working with the Karhunen-Loève (Karhunen, 1946; Lo'eve, 1955) expansion of f,

$$f(\boldsymbol{x}) = \sum_j \psi_j(\boldsymbol{x})\beta_j \tag{3}$$

where ψ_j are the orthonormal eigenfunctions of the integral operator based on the covariance function and β_j are the coordinates of f. For the mean zero Gaussian process, the β_j are independent normal random variables with mean 0 and variance equal to the eigenvalues λ_j. In practice, one may obtain a finite dimensional approximation to f by using the eigenfunctions corresponding to the n largest eigenvalues and setting the remaining $\beta_j, j > n$ to zero. However, starting with a covariance function and determining the corresponding orthonormal eigenfunctions is a challenging task (see Xia and Gelfand, 2005, for discussion and alternative solutions).

The connection between covariance functions in Gaussian processes and kernels in a reproducing Kernel Hilbert space $\mathcal{H}$ has led to a tremendous resurgence of interest in kernel representations of f (see http://kernel-machines.org) in both the statistics and machine learning communities. A Hilbert space associated with a positive definite function $k(\cdot,\cdot)$ on $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{X}}$, such as the covariance function $\boldsymbol{\Sigma}$, may be constructed as the collection of all finite linear combinations of the form $\sum k(\cdot, \boldsymbol{x}_j)$ and their limits (under the norm induced by the inner product $\langle k(\cdot, \boldsymbol{x}_i), k(\cdot, \boldsymbol{x}_j)\rangle = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (Wahba, 2002)). The representer theorem (Kimeldorf and Wahba, 1971) leads to a finite dimensional representation of f:

$$f_\lambda(\cdot) = \sum_{j=1}^{n} k(\cdot, \boldsymbol{x}_j)\beta_j \tag{4}$$

as the solution to the optimization problem of finding $f \in \boldsymbol{\mathcal{H}}$ to minimize the penalized loss functional

$$\sum_i L\left(y_i, f(\boldsymbol{x}_i)\right) + \lambda||f||_k^2 \tag{5}$$

where L is typically (although not necessarily) convex in f. The equivalence between the penalized loss (5) and negative log posterior under a Gaussian prior on β leads to a maximum *a posteriori* (MAP) estimate of f using kernels at the observed data points $\boldsymbol{x}_i$ and provides a Bayesian interpretation of support vector machines (SVM) (Law and Kwok, 2001; Sollich, 2002). These representations, however, use as many basis functions as observations.

Starting with the finite dimensional representation based on the MAP estimate, Tipping (2001) and Chakraborty *et al.* (2004) provide extensions of the SVM model

by replacing the Gaussian prior on the β_j with independent scale mixtures of normals,

$$\beta_j \mid \lambda_j \overset{ind}{\sim} \mathrm{N}(0, 1/\lambda_j) \tag{6}$$

where λ_j is given a Gamma prior distribution. As with the LASSO (Tibshirani, 1996), which is based on a scale mixture of normals corresponding to a double exponential prior, the posterior *modes* for a subset of the β_j may be zero, resulting in a MAP estimate of f based on fewer than the n basis functions used in the original Bayesian SVM solution with Gaussian priors.

As suggested by the above expansions, an alternative to using a Gaussian process prior is to expand f directly in terms of a countable collection of basis functions $\phi_j(\cdot), j \in \mathcal{J}$

$$f(\boldsymbol{x}_i) = \sum_{j \in \mathcal{J}} \psi_j(\boldsymbol{x}_i)\beta_j \tag{7}$$

where the $\{\beta_j, j \in \mathcal{J}\}$ are the unique (but unknown) coefficients of f with respect to the basis $\{\psi_j(\cdot), j \in \mathcal{J}\}$ for $\mathcal{H}$. This includes basis functions determined by kernels $g(\cdot, \boldsymbol{x}_j; \boldsymbol{\theta})$, piecewise polynomials, Fourier series, splines, wavelets, etc. One advantage of this approach is that there is no need to restrict the generator to be a positive (Mercer) kernel (Tipping, 2001) allowing for a more flexible representations. A problem with many classical bases, such as a Fourier basis or spline basis, is that they are *non-local*, meaning that many basis functions may contribute to values of the decomposition at "spatially" distant points. Typically, this will lead to decompositions with many non-zero coefficients and is inefficient in the sense that there may be significant "cancellation" of coefficients in order to determine the value of the function at a given point. Wavelet bases and bases generated from kernels with compact support, on the other hand, are "local", leading to more adaptive and parsimonious representations. Even so, orthonormal wavelet bases have a disadvantage in that the positions and the scales of the basis functions are subject to dyadic constraints and may require potentially more basis functions than with a non-orthonormal wavelet basis.

Recent developments using overcomplete representations through frames and dictionaries where the number of functions $|\mathcal{J}|$ in the expansion of $\boldsymbol{f}$ is potentially greater than n (in the finite dimensional case) show great promise (Wolfe *et al.*, 2004; Johnstone and Silverman, 2005). While inherently redundant, overcomplete representations have been advocated due to their increased flexibility and adaptation over orthonormal bases, particularly for finding sparse representations of f. Examples of overcomplete dictionaries include unions of bases, Gabor frames, nondecimated or translational invariant wavelets and wavelet packets. Because of the redundancy of overcomplete representations, coefficients β_j in the expansion (7) using all dictionary elements are not unique. This lack of uniqueness *a priori* is advantageous, permitting more parsimonious representations to be extracted from the dictionary than those obtained using any single basis as shown by Wolfe *et al.* (2004) .

Working with Gabor frames, Wolfe *et al.* (2004) used the variable selection priors popularized by the Stochastic Search Variable Selection (SSVS) model of George and McCulloch (1997). Dictionary elements are identified by introducing an indicator variable γ_j, such that

$$\beta_j \mid \lambda_j, \gamma_j \overset{ind}{\sim} \mathrm{N}(0, \sigma^2\gamma_j/\lambda_j). \tag{8}$$

When γ_j is zero, the distribution of β_j is degenerate at 0, so that function ψ_j is not included in the expansion. A hierarchical prior for the indicators variables $\gamma_j, j \in \mathcal{J}$ completes the specification; by default the γ_j are taken as independent Bernoulli random variables with $\Pr(\gamma_j = 1) = \pi \in (0, 1)$. Bayesian variable selection has been used successfully to identify sparse solutions in the finite dimensional regression formulation and provides the canonical model for soft and hard thresholding of wavelet coefficients (Abramovich *et al.*, 1998; Clyde *et al.*, 1998; Clyde and George, 2000; Johnstone and Silverman, 2005; Vidakovic, 1999) and models for automatic curve fitting using splines or piecewise polynomials (Smith and Kohn, 1996; Denison *et al.*, 1998a,b) (see Clyde and George (2004) for an overview and additional references).

2.1. *Limiting Version of SSVS*

As the sample size n increases, it is common to consider a richer collection of potential dictionary elements, and an important challenge is to characterize the limiting behaviour of the SSVS prior. This is particularly relevant in overcomplete frames in $\mathbb{R}^n$ for fixed sample size n, as the number of dictionary elements increases. If we let $J \equiv \sum_{j \in \mathcal{J}} \gamma_j$ denote the number of functions in the expansion with nonzero coefficients, then the independent Bernoulli priors on the γ_j lead to a $\text{Bi}(|\mathcal{J}|, \pi)$ distribution for J. As $|\mathcal{J}| \to \infty$ and $\pi \to 0$, with $\pi\,|\mathcal{J}|$ converging to a constant ν_+, the number of dictionary elements in the expansion will have a limiting Poisson distribution with mean ν_+, and the number of terms in the expansion will be finite almost surely. Assuming independent prior distributions for β_j (as in Wolfe *et al.*, 2004), the resulting limiting prior distribution on f is a *compound Poisson random field*, which is a special case of a Lévy random field. The examples in Wolfe *et al.* (2004) used a finite (but large) dictionary. Note, however, that the collection of dictionary elements may not even be countable, for example, in the case of free knot splines (DiMatteo *et al.*, 2001), continuous wavelets (Vidakovic, 1999, Chapter 3) or kernels evaluated at points other than the observed data points. Lévy random fields as limits of sequences of compound Poisson fields provide a natural choice for priors on expansions of functions using overcomplete dictionaries. While proper prior distributions on β_j are required in trans-dimensional problems in order for marginal likelihoods to be well determined (Clyde and George, 2004), we will see that the Lévy random field priors lead to improper measures for the coefficients and infinite ν_+ in the limiting version. This provides a new avenue for objective Bayesian analysis in nonparametric modeling.

3. LÉVY RANDOM FIELDS AND STOCHASTIC EXPANSIONS

For illustration, we will consider overcomplete dictionaries created by a generating function g,

$$\phi_{\boldsymbol{\omega}}(\boldsymbol{x}) \equiv g(\boldsymbol{x}, \boldsymbol{\omega}) \qquad \boldsymbol{\omega} \in \boldsymbol{\Omega} \tag{9}$$

a Borel measurable function $g : \mathcal{X} \times \boldsymbol{\Omega} \to \mathbb{R}$ where $\boldsymbol{\Omega}$ is a complete separable metric space and for fixed $\boldsymbol{\omega}$, $g(\cdot, \boldsymbol{\omega})$ is an element in $\mathcal{H}$. Examples of generating functions include density functions (either normalized or un-normalized) from location-scale families of distributions. In the one dimensional setting, the choice

$$g(x, \boldsymbol{\omega}) = \exp\{-\lambda |x - \chi|^p\} \tag{10}$$

includes the un-normalized Gaussian ($p = 2$) and Laplace densities ($p = 1$), where $\boldsymbol{\omega}^T \equiv (\chi, \lambda)^T$ and $\chi \in \mathbb{R}$ and $\lambda \in \mathbb{R}^+$ (p may also be included in $\boldsymbol{\omega}$ as well). Of course,

there is no need to restrict attention to location families or symmetric generators. For example, the un-normalized exponential density or other asymmetric densities may be more appropriate for modeling pollution dissipation over time.

Shifts and scales of a wavelet function ψ,

$$g(x,\boldsymbol{\omega}) = \lambda^{1/2}\psi\Big(\lambda(x-\chi)\Big) \tag{11}$$

provide other generating functions with desirable features for functions exhibiting non-stationarity, where the choice of wavelet may be based on a particular smoothness class of the function. The original Haar wavelet with $\boldsymbol{\mathcal{X}} = [0,1]$

$$g(x,\boldsymbol{\omega}) \equiv \mathbf{1}_{\{0<\lambda(x-\chi)\le 1\}} \tag{12}$$

is the simplest. Many popular wavelets with compact support do not have explicit closed forms, but fast numerical calculation of the continuous wavelet at any point may be obtained using filters and the Daubechies-Lagarias (Vidakovic, 1999, p. 89).

3.1. *Lévy Random Fields*

With the goal of extracting potentially sparse representations for f from an overcomplete dictionary, we consider expansions of the form

$$f(x) \equiv \sum_{0\le j<J} g(\boldsymbol{x},\boldsymbol{\omega}_j)\,\beta_j \tag{13}$$

for a random number $J \le \infty$ of randomly drawn pairs $\beta_j \in \mathbb{R}$, $\boldsymbol{\omega}_j \in \boldsymbol{\Omega}$. Note that the sum is equivalent to an integral with respect to a random *signed* Borel measure $\mathcal{L}(d\boldsymbol{\omega}) \equiv \sum \beta_j \delta_{\boldsymbol{\omega}_j}(d\boldsymbol{\omega})$ on $\boldsymbol{\Omega}$ with random support points $\boldsymbol{\omega}_j$ and "jumps" at $\boldsymbol{\omega}_j$ of size β_j. This leads to the equivalent stochastic integral representation:

$$f(x) = \int_{\boldsymbol{\Omega}} g(x,\boldsymbol{\omega})\mathcal{L}(d\boldsymbol{\omega}) \tag{14}$$

for the random prior measure $\mathcal{L}$. Our goal is to deduce the random measure $\mathcal{L}$ based on information in the data, a form of inverse problem (Wolpert *et al.*, 2003).

As suggested by the limiting version of the SSVS priors and other priors in transdimensional problems, an intuitive construction of such random measures begins by choosing any positive number $\nu_+ > 0$ and assigning J a Poisson distribution, $J \sim \mathrm{Po}(\nu_+)$. Then, conditionally on J, accord the $(\beta_j, \boldsymbol{\omega}_j) \in \mathbb{R}\times\boldsymbol{\Omega}$ independent identical distributions, $(\beta_j,\boldsymbol{\omega}_j) \sim \pi(d\beta, d\boldsymbol{\omega})$, where π is a probability distribution on $\mathbb{R}\times\boldsymbol{\Omega}$. In that case, the random measure $\mathcal{L}$ assigns independent infinitely-divisible (ID) random variables,

$$\mathcal{L}(A_i) \equiv \sum_{0\le j<J} \mathbf{1}_{A_i}(\boldsymbol{\omega}_j)\,\beta_j \tag{15}$$

to disjoint Borel sets $A_i \subset \boldsymbol{\Omega}$. The random variables $\mathcal{L}(A_i)$ for a collection of Borel sets A_i is an example of a Lévy random field with "Lévy measure" $\nu(d\beta, d\boldsymbol{\omega}) = \nu_+\pi(d\beta, d\boldsymbol{\omega})$, the product of the Poisson rate ν_+ for J and the distribution $\pi(d\beta, d\boldsymbol{\omega})$ for $\{(\beta_j,\boldsymbol{\omega}_j)\}$. Here π is proper distribution and $\nu(\mathbb{R}\times\boldsymbol{\Omega})$ is finite (by construction).

More generally, Khinchine and Lévy (1936) showed that any ID random variable and indeed any ID random measure (Rajput and Rosiński, 1989, Prop. 2.1) has characteristic function

$$\mathrm{E}\left[e^{it\mathcal{L}(A)}\right] = \exp\left\{it\delta(A) - \tfrac{1}{2}t^2\Sigma(A) + \int\int_{\mathbb{R}\times A}\left(e^{it\beta} - 1 - ith(\beta)\right)\nu(d\beta, d\boldsymbol{\omega})\right\} \tag{16}$$

determined uniquely by the characteristic triplet of sigma-finite measures (δ, Σ, ν) consisting of a signed measure $\delta(d\boldsymbol{\omega})$ and a positive measure $\Sigma(d\boldsymbol{\omega})$ on $\boldsymbol{\Omega}$, and a positive measure $\nu(d\beta, d\boldsymbol{\omega})$ on $\mathbb{R}\times\boldsymbol{\Omega}$ that for each compact $K \subset \boldsymbol{\Omega}$ satisfies

$$\int\int_{\mathbb{R}\times K}(1\wedge\beta^2)\nu(d\beta, d\boldsymbol{\omega}) < \infty \tag{17}$$

and $\nu(\{0\}, \boldsymbol{\Omega}) = 0$. The function $h(\beta) \equiv \beta\mathbf{1}_{[-1,1]}(\beta)$ is known as the "compensator" and is required to make the last integrand in (16) bounded and $O(\beta^2)$ for β in a neighborhood of zero. When the Lévy measure satisfies the more restrictive condition

$$\int\int_{\mathbb{R}\times K}\Big(1\wedge|\beta|\Big)\nu(d\beta, d\boldsymbol{\omega}) < \infty \tag{18}$$

for each compact $K \subset \boldsymbol{\Omega}$, we may take the compensator $h(\beta)$ to be zero. From the characteristic function, $\mathcal{L}(A)$ may be recognized as the sum of two independent parts: a Gaussian component (with mean and covariance determined by the first two terms in (16)) and a "pure jump" component with Lévy measure ν. We will restrict attention to random measures $\mathcal{L}$ with $\delta \equiv 0$ and $\Sigma \equiv 0$ (no Gaussian component).

A random measure $\mathcal{L}$ satisfying (16) (without Gaussian component) induces a *random field*, a linear mapping from functions $\phi : \boldsymbol{\Omega} \to \mathbb{R}$ to random variables

$$\mathcal{L}[\phi] \equiv \int_{\boldsymbol{\Omega}}\phi(\boldsymbol{\omega})\mathcal{L}(d\boldsymbol{\omega}) \tag{19}$$

with characteristic function

$$\mathrm{E}\left[e^{it\mathcal{L}[\phi]}\right] = \exp\left\{\int\int_{\mathbb{R}\times\boldsymbol{\Omega}}\left(e^{it\phi(\boldsymbol{\omega})\beta} - 1 - it\phi(\boldsymbol{\omega})\,h(\beta)\right)\nu(d\beta, d\boldsymbol{\omega})\right\}. \tag{20}$$

When the bound (18) holds, the class of functions ϕ and spaces $\boldsymbol{\Omega}$ for which (20), and hence the random field $\mathcal{L}[\phi]$, is well defined includes all bounded measurable compactly supported functions. More generally, Rajput and Rosiński (1989)show that the space Φ of functions that are integrable with respect to an ID random measure $\mathcal{L}(d\boldsymbol{\omega})$ with no Gaussian component are certain *Musielak-Orlicz* modular spaces.

For expansions of f, let $\mathcal{G}$ denote the linear space of measurable functions $g : \boldsymbol{\mathcal{X}}\times\boldsymbol{\Omega} \to \mathbb{R}$ for which $g(\boldsymbol{x}, \cdot) \in \Phi$ for each $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$. For any generator function $g \in \mathcal{G}$, we can construct a random function $f : \boldsymbol{\mathcal{X}} \to \mathbb{R}$ by $f(\boldsymbol{x}) \equiv \mathcal{L}[g(\boldsymbol{x}, \cdot)]$; in particular this holds for the generators in Sec 3 (see Tu *et al.*, 2006 for more details on conditions and examples) . Moments of $f(x) = \mathcal{L}[g(x, \cdot)]$, when they exist, are easy to compute from the characteristic function (20) for any $\mathcal{L}[\phi]$:

$$\mathrm{E}\Big[f(x)\Big] = \int\int_{\mathbb{R}\times\Omega} g(x, \boldsymbol{\omega})\,\Big[\beta - h(\beta)\Big]\nu(d\beta, d\boldsymbol{\omega}) \tag{21}$$

$$\mathrm{Cov}\Big[f(x_1), f(x_2)\Big] = \int\int_{\mathbb{R}\times\Omega} g(x_1, \boldsymbol{\omega})\,g(x_2, \boldsymbol{\omega})\,\beta^2\,\nu(d\beta, d\boldsymbol{\omega}). \tag{22}$$

3.2. *Poisson Representation*

Tu *et al.* (2006) give an equivalent construction of the Levy random fields from a sequence of compound Poisson random fields, which is key to tractable posterior inference. Let $N(d\beta, d\boldsymbol{\omega}) \sim \text{Po}(\nu)$ denote a Poisson random measure on $(\mathbb{R} \times \Omega)$ that assigns independent random variables with $\text{Po}(\nu(C_i))$ distributions to disjoint Borel sets $C_i \subset (\mathbb{R} \times \Omega)$ and denote the *compensated* or centered Poisson measure $\tilde{N}(d\beta, d\boldsymbol{\omega}) \equiv N(d\beta, d\boldsymbol{\omega}) - \nu(d\beta, d\boldsymbol{\omega})$ (Sato, 1999, page 38). Then equivalently

$$\mathcal{L}[\phi] \stackrel{d}{=} \int\int_{\mathbb{R}\times\Omega} [\beta - h(\beta)]\, \phi(\boldsymbol{\omega}) N(d\beta, d\boldsymbol{\omega}) + \int\int_{\mathbb{R}\times\Omega} h(\beta)\, \phi(\boldsymbol{\omega}) \tilde{N}(d\beta, d\boldsymbol{\omega}) \tag{23}$$

which simplifies to

$$\mathcal{L}[\phi] \stackrel{d}{=} \int\int_{\mathbb{R}\times\Omega} \beta\phi(\boldsymbol{\omega}) N(d\beta, d\boldsymbol{\omega}) \tag{24}$$

in the case ν satisfies (18).

In the case of infinite Lévy measure, the number of points $N(\mathbb{R}, \boldsymbol{\Omega})$ will be infinite almost surely. While the integrals $\mathcal{L}[g]$ will be well behaved, stochastic expansions with an infinite number of terms are not practical for posterior simulation of the distribution of f. One solution is to truncate jumps less than a certain threshold in absolute value, since we are primarily interested in sparse solutions based on the jumps with largest absolute magnitudes. When the integrability condition (18) holds, we can always approximate $\mathcal{L}$ and $\mathcal{L}[\phi]$ by choosing some small $\epsilon > 0$, and replacing $\mathbb{R}$ in (24) by $[-\epsilon, \epsilon]^c$. Thus,

$$\mathcal{L}_\epsilon[\phi] \equiv \int\int_{[-\epsilon,\epsilon]^c\times\boldsymbol{\Omega}} \phi(\boldsymbol{\omega})\beta\, N(d\beta, d\boldsymbol{\omega}) = \int\int_{\mathbb{R}\times\boldsymbol{\Omega}} \phi(\boldsymbol{\omega})\beta\, N_\epsilon(d\beta, d\boldsymbol{\omega}) \tag{25}$$

where N_ϵ is a Poisson measure on $\mathbb{R} \times \boldsymbol{\Omega}$ with intensity measure

$$\nu_\epsilon(d\beta, d\boldsymbol{\omega}) \equiv \nu(d\beta, d\boldsymbol{\omega}) \mathbf{1}_{\{|\beta|>\epsilon\}}. \tag{26}$$

Consequently, the Lévy random field $\mathcal{L}[\phi]$ may be approximated by a sequence of compound Poisson random fields, $\mathcal{L}_\epsilon[\phi]$ where $\mathcal{L}_\epsilon[\phi]$ converges in distribution to $\mathcal{L}[\phi]$ as $\epsilon \to 0$. Similar approximations are possible in the case the Lévy measure satisfies the more general bound (17), however, one may need to include an ϵ dependent deterministic adjustment due to the compensator (Tu *et al.*, 2006).

Truncating the support is not the only way to construct suitable approximating sequences of finite Lévy measures $\nu_\epsilon(d\beta, d\boldsymbol{\omega})$ for which the integrals $\mathcal{L}_\epsilon[\phi]$ converge in distribution to $\mathcal{L}[\phi]$. The Lévy measure $\nu(d\beta, d\boldsymbol{\omega}) = \alpha(d\boldsymbol{\omega})\beta^{-1}e^{-\beta\tau}\mathbf{1}_{\mathbb{R}_+}(\beta)d\beta$ for the Gamma $\text{Ga}(\alpha(d\boldsymbol{\omega}), \tau)$ random field, for example, may be approximated by $\nu_\epsilon(d\beta, d\boldsymbol{\omega}) \equiv \gamma(d\boldsymbol{\omega})\beta^{\epsilon-1}e^{-\beta/\tau}\mathbf{1}_{\mathbb{R}_+}(\beta)d\beta$, a finite measure (if $\gamma(\boldsymbol{\Omega}) < \infty$ and $\epsilon > 0$) with full support. Truncation has an advantage, in that it maintains the exact conditional distribution for all included mass points $(\beta_j, \boldsymbol{\omega}_j)$, and merely sets to zero the smallest coefficients. This focus on the large magnitude coefficients is desirable, as we are interested in finding sparse expansions.

3.3. *Examples of Lévy measures*

Examples of Lévy random fields with infinite Lévy measures include the symmetric α-stable (SαS) family for $0 < \alpha < 2$ (including the Cauchy process with $\alpha = 1$), with Lévy measure

$$\nu(d\beta, d\boldsymbol{\omega}) = c_\alpha \gamma(d\boldsymbol{\omega})|\beta|^{-1-\alpha}\, d\beta \tag{27}$$

for some constant $c_\alpha > 0$, giving $\mathcal{L}[A] \sim \mathrm{St}(\alpha, 0, \gamma(A), 0)$. For $1 \leq \alpha < 2$, the construction of $\mathcal{L}[\phi]$ requires compensation as in (23). Interestingly, the Lévy measure for the symmetric stable may be represented as a mixture of normals

$$\nu(d\beta, d\boldsymbol{\omega}) \propto \gamma(d\boldsymbol{\omega}) \int_0^\infty \xi^{1/2} \exp\left(-\frac{1}{2}\beta^2\xi\right) \xi^{\alpha/2-1}\, d\xi \tag{28}$$

where the mixing distribution on ξ is a limiting version of a $\mathrm{Ga}(\alpha/2, b)$ distribution with $b = 0$. This suggests a way of extending the independent normal priors in SVMs such that the limiting infinite distribution on f is an integral with respect to a Stable random field. The prior used by [**?**], in fact corresponds to taking $\alpha = b = 0$, which is not equivalent to using a Lévy measure for an α-stable random field, and does not lead to a proper posterior.

The Gamma random field is another important example used in constructing non-negative functions, with Lévy measure

$$\nu(d\beta, d\boldsymbol{\omega}) = \gamma(d\boldsymbol{\omega})\beta^{-1}e^{-\beta/\tau}\mathbf{1}_{\{\beta>0\}}\, d\beta \tag{29}$$

for some σ-finite measure $\gamma(d\boldsymbol{\omega})$ on $\boldsymbol{\Omega}$, giving $\mathcal{L}[A] \sim \mathrm{Ga}(\gamma(A), 1/\tau)$ for $A \subset \boldsymbol{\Omega}$ of finite γ measure. A symmetric analogue of the Gamma random field (29) has Lévy measure

$$\nu(d\beta, d\boldsymbol{\omega}) = |\beta|^{-1}e^{-|\beta|/\tau}\, d\beta\, \gamma(d\boldsymbol{\omega}) \tag{30}$$

on all of $\mathbb{R} \times \Omega$, leading to random variables $\mathcal{L}(A)$ distributed as the difference of two independent $\mathrm{Ga}(\gamma(A), 1/\tau)$ variables, with characteristic functions

$$\mathrm{E}[e^{it\mathcal{L}(A)}] = (1 + t^2\tau^2)^{-\gamma(A)}.$$

Both the standard positive and this symmetric Gamma random measures satisfy the bound (18), thus no compensation is required and we may use (25) to approximate (24).

The means $\mathrm{E}[f(x)]$ are available directly from (21); they vanish for the symmetric Gamma random field, $\mathrm{E}[f(x)] = 0$, or for any other Lévy random field with a symmetric (in $\pm\beta$) Lévy measure that satisfies (18). Covariances (when they exist) are available from (22). Nearly all of the commonly used isotropic geostatistical covariance functions (see Chilès and Minières, 1999, Ch. 2.5) may be achieved by the choice of a suitable generating kernel $g(x, \cdot)$ and Lévy measure $\nu(d\beta, d\boldsymbol{\omega})$. The Gaussian generating kernel $|\,(x)$ of (10)with $p = 2$ and the symmetric Gamma Lévy measure ν from (30) with $\gamma(d\boldsymbol{\omega}) = \gamma\pi(d\lambda)\, d\chi$ for some $\alpha > 0$ and $\pi(d\lambda)$ a point mass at $\lambda_0 > 0$ lead to

$$\mathrm{Cov}\Big[f(x_1), f(x_2)\Big] = 2\alpha\tau^2\sqrt{\pi/\lambda_0}\, e^{-\lambda_0(x_1-x_2)^2/4},$$

the isotropic Gaussian covariance function, while the choice $\pi(d\lambda) = \mathrm{Ga}(a, b)$ for some $a > \frac{1}{2}$ and $b > 0$ leads to the Generalized Cauchy model (Yaglom, 1987, p. 365):

$$\mathrm{Cov}\Big[f(x_1), f(x_2)\Big] = 2\alpha\tau^2\sqrt{\pi b}\, \frac{\Gamma(a - 1/2)}{\Gamma(a)} \left[1 + \frac{(x_1 - x_2)^2}{4b}\right]^{1/2-a}.$$

For the Laplace generating kernel (10) with $p = 1$ with the same Lévy measure ν, the covariance is

$$\begin{aligned}\mathrm{Cov}\Big[f(x_1), f(x_2)\Big] &= 2\alpha\tau^2 \int_{\mathbb{R}^+} \frac{1}{\lambda} e^{-\lambda|x_1-x_2|}\Big(1+\lambda|x_1-x_2|\Big)\pi(\lambda)d\lambda \\ &= \frac{2\alpha\tau^2}{a-1}\,\frac{b+a|x_1-x_2|}{[1+|x_1-x_2|/b]^a}\,.\end{aligned}$$

Other examples of kernels and resulting covariance functions may be found in Tu *et al.* (2006).

4. HIERARCHICAL MODEL

Capitalizing on the (approximate) compound Poisson representation of $\mathcal{L}$ based on the truncation of small jumps, we may state the likelihood and prior of f in hierarchical fashion

$$Y_i \mid f(\boldsymbol{x}_i) \overset{ind}{\sim} \mathrm{N}(f(\boldsymbol{x}_i), \sigma^2) \qquad f(\boldsymbol{x}_i) = \sum_{0\le j<J} g(\boldsymbol{x}_i, \boldsymbol{\omega}_j)\beta_j \tag{31}$$

$$J \sim \mathrm{Po}(\nu_+) \quad \text{where } \nu_+ \equiv \nu_\epsilon(\mathbb{R}, \boldsymbol{\Omega}) \tag{32}$$

$$(\beta_j, \boldsymbol{\omega}_j) \mid J \overset{iid}{\sim} \pi(d\beta_j, d\boldsymbol{\omega}_j) \equiv \frac{\nu_\epsilon(d\beta_j, d\boldsymbol{\omega}_j)}{\nu_\epsilon(\mathbb{R}, \boldsymbol{\omega})} \qquad \text{for } j = 1, \ldots, J \tag{33}$$

where J is the random number of terms in the stochastic expansion, $(\beta_1, \ldots, \beta_J)$ represents the unknown coefficients and $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_J)$ represents the collection of generator specific parameters. A prior distribution on σ^2 completes the prior specification for the first stage of the hierarchy. While we have stated the model in terms of the Gaussian error model (1), other distributions for Y, of course, can replace the normal assumption without any loss of generality.

We may also place a prior distribution on parameters in the Lévy measure. With either the Gamma (29) or Stable (27) random fields and default stationary measure $\gamma(d\boldsymbol{\omega}) \equiv \gamma\pi(d\lambda)\,d\chi$ for some distribution $\pi(d\lambda)$, we may place a Gamma prior on γ, which leads to J having a Negative Binomial distribution. This provides robustness to a fixed choice of γ by providing over-dispersion. Hyperparameters in the Negative-Binomial may be elicited by specifying various quantiles, for example fixing a probability that there are no components (just a constant mean) and specifying the 95^{th} quantile for J. This has given reasonable behaviour in a wide variety of problems. The prior distribution $\pi(d\lambda)$ needs more careful specification to keep the kernel generating functions from having support that is too narrow. We have used a $\mathrm{Ga}(a_\lambda, b_\lambda)$ prior for λ, with selection of the hyperparameters based on subjective information about the problem.

While expressions for posterior modes or posterior distributions of quantities of interest do not exist in closed form, the prior construction using Lévy random fields permits tractable posterior simulation via a reversible jump Markov chain Monte Carlo algorithm Green (1995). Efficient computation is possible because updates to f based on adding/deleting or updating single dictionary elements bypass the need to invert large matrices. Furthermore, because dictionary elements $g(\boldsymbol{x}, \boldsymbol{\omega})$ are only computed as needed, memory requirements scale linearly with the sample size. For generating functions with compact support, further improvements in computational speed are possible.

5. MODEL COMPLEXITY

Given observations $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_n)^T$, the log of the posterior distribution

$$\log\left[\pi\left(\{\beta_j, \boldsymbol{\omega}_j\}_{j=1}^J,\ J,\ \boldsymbol{\theta} \mid \boldsymbol{Y}\right)\right]$$

where $\boldsymbol{\theta}$ represents the fixed dimensional parameters (σ^2, hyperparameters in the Lévy measure, etc), takes the form of a penalized or regularized likelihood

$$\text{constant} \ -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - f(\boldsymbol{x}_i))^2 - \log(J!) + \sum_{j=1}^{J}\log(\nu_\epsilon(d\beta_j, d\boldsymbol{\omega}_j)). \tag{34}$$

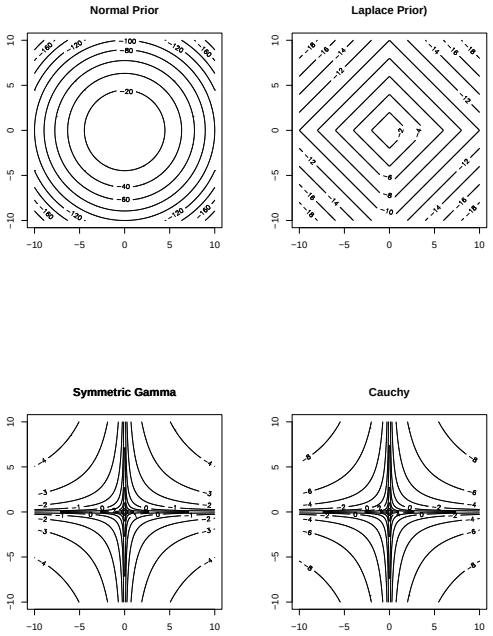

Figure 1: *Log contours for the joint distribution of two β coefficients in the expansion of f under independent normal, independent double exponential, the Gamma random eld prior which is proportional to (30)), and the Cauchy random field prior, which is proportional to (27).*

Modelcomplexity is penalized directly through the $\log(J!)$ term, as in ℓ_0 penalties which penalizes the number of coefficients in the expansion. As in other Bayesian model selection examples, model complexity is also indirectly penalized through the choice of prior on the regression coefficients, in this case through the Lévy measure ν. Figure 1 contrasts the contours of the log prior for two coefficients $(\beta_j, \beta_{j'})$ under the independent normal priors (spherical contours), the independent double exponential prior (diamond) and the priors based on the Lévy measures for the

symmetric Gamma and Cauchy random fields. The approximate Lévy measure in the symmetric Gamma and Cauchy random field models may be seen as inducing a sparsity penalty for the addition of generator functions to the function $f(x)$, similar (or stronger in fact) to the L_1 penalties of the LASSO (Tibshirani, 1996). The shape of the joint prior makes it much harder to keep redundant components in the model than with the LASSO.

The Gamma and Cauchy processes are both examples of infinite Lévy measures, and in order to restrict the expansion to a finite number of terms, we must restrict $|\beta| > \epsilon$. This may be related to the idea of practical significance in the non-conjugate version of the SSVS algorithm George and McCulloch (1993); Chipman *et al.* (1997) where the prior distribution on β is a mixture of two normal distributions; one fairly dispersed and the other concentrated around zero. The variance in the concentrated distribution is selected to reflect values of the coefficient that for practical purposes suggest that the variable could be dropped from the model. The choice of ϵ in the Lévy random field framework may be guided by this idea of practical significance for estimating f in the presence of noise, in that dictionary elements with coefficients larger than ϵ in absolute value will be retained. In the limit as $\epsilon \to 0$, the distribution for any β taken on its own is actually improper, with an infinite spike at zero, however, the prior distribution on f with infinite J and infinite measure ν is well defined based on (20).

6. APPLICATIONS

We illustrate the Lévy random field priors in several examples, using both simulated and real applications to highlight the versatility of the priors.

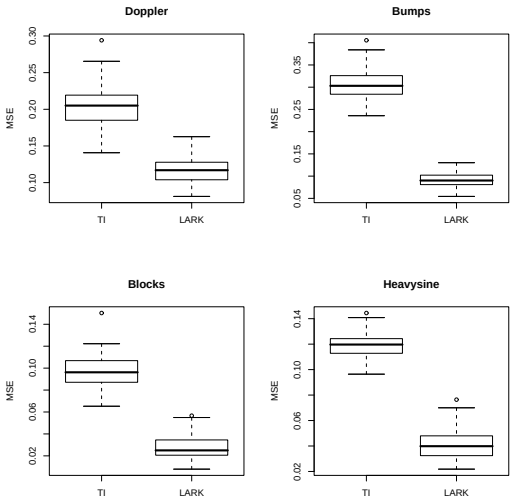

Figure 2: *Average MSE over 100 simulations using the Empirical Bayes estimator with translation-invariant wavelet transform (TI) (Johnstone and Silverman, 2005) and Lévy Adaptive Regression Kernels (LARK) using the ϵ-truncated symmetric Gamma Lévy random field prior (Tu et al, 2006).*

Example 1 (Wavelets). In Tu *et al.* (2006) we compare the performance of the Lévy random field priors to estimators based on translational invariant wavelets (Johnstone and Silverman, 2005), another overcomplete representation, using simulated data and several of the now standard wavelet test functions: Blocks, Bumps, Doppler and Heavysine (Donoho and Johnstone, 1994). Figure 2 illustrates the significant improvements in mean squared error for estimating the true function using the symmetric Gamma Lévy random field priors.

Because parameters in the generators g are allowed to vary with location, the Lévy random field priors lead to adaptive estimation. Like translation-invariant wavelets, the prior representation leads to an overcomplete representation, however, the posterior has a much sparser representation than with the non-decimated wavelets. We have illustrated the methods using generating functions based on kernels, thus the reduction in MSE may be due to choice of prior, generator, or both. The methodology is easily extended to continuous wavelets using the Daubechies-Lagarias algorithm for evaluating wavelets at arbitrary locations/scales. While preliminary results of Chu *et al.* (2006) using continuous wavelets with the compound Poisson priors with finite J and a normal prior on β as suggested in Abramovich *et al.* (2000) provide improvements over the translational invariant wavelets (using the same wavelet generators), we expect additional reductions may be possible using the heavier-tailed symmetric Gamma or Cauchy random field prior distributions with the overcomplete wavelet dictionary.

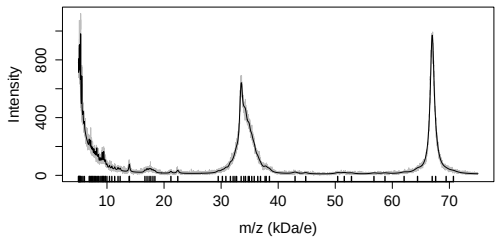

Figure 3: *A raw spectrum and estimates of mean intensity from the highest posterior draw; the rug plot at the bottom indicates locations of the latent proteins (peaks).*

Example 2 (Mass Spectroscopy). In Clyde *et al.* (2006) we use Gamma random field priors to construct models for the latent relative abundance of proteins as a function of their mass/charge (or equivalently time of flight) using data from Matrix Assisted Laser Desorption/Ionization Time of Flight mass spectroscopy. Normalized Gaussian kernels with time varying scale parameters provide a natural choice of generating functions to capture the variation in time of flight of proteins of a given mass/charge. Combined with the positive Gamma measure ν in (29), f will be non-negative. Unlike wavelets or spline models, the parameters in the adaptive kernel model have interesting biological interpretations: J is the number of unknown proteins in the sample, β_j is the unknown concentration for a protein with expected time of flight τ_j, and the resolution parameter ρ_j governs the peak widths λ_j (here we take $\boldsymbol{\omega}_j \equiv (\tau_j, \rho_j)$). This interpretability is a key feature of the Lévy random field models, as it allows us to incorporate subjective prior information regarding resolution and time of flight (a transformation of the mass/charge). In this example,

we drop the normality assumption and use a Gamma distribution for Y_i without any additional computational complexity. Figure 3 illustrates a draw from the RJ-MCMC output, where the rug plot on the x-axis indicates locations of peaks by their time of flight.

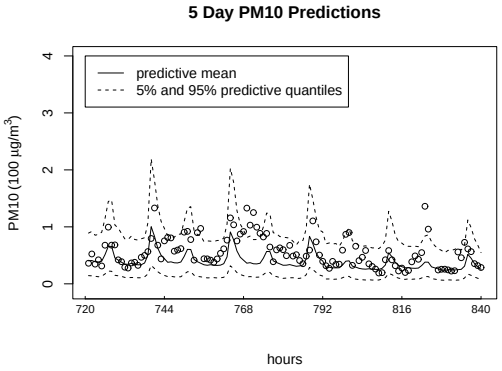

Figure 4: *Five day forecast of PM10 predictions with 90% credible intervals (pointwise).*

Example 3 (Non-Stationary Spatial-Temporal Models). The third area of application concerns development of non-stationary temporal (as well as spatial and spatial-temporal models) for concentrations of one or more criteria pollutants. As expected pollution concentrations are inherently non-negative, the Lévy random field prior based on a Gamma random field ensure that the expected functions are non-negative, and is a natural alternative to the commonly adopted Gaussian random field priors in spatial-temporal models. The models here may be seen as natural extensions of the work of Wolpert and Ickstadt (1998), who developed conjugate Poisson-Gamma spatial temporal models. Unlike Gaussian process convolution models (Higdon *et al.*, 1999; Lee *et al.*, 2002; Xia and Gelfand, 2005), there is no need to evaluate the kernels on a regular lattice, thus the Lévy random field models have the potential to be more parsimonious, as in the case of critically sampled wavelets versus continuous wavelets. The spatial-temporal locations of jumps in the Lévy random field may be interpreted as point sources of pollution, with dispersal over time and space controlled by additional parameters in the kernels. Hierarchical models for parameters in the Lévy measure allow incorporation of meteorological variables which influence both the dispersal parameters and expected concentrations. An extension of the models described here is to utilize marked random fields that allow common jumps (shared impulses) between two or more pollutants or that represent periodic (repetitive) events over time (for example, peaks in concentration due to morning/evening commutes which occur regularly).

Figure 4 illustrates out-of-sample predictions for a time series of particulate matter from Phoenix, AZ. The model incorporates a periodic process that repeats daily and automatically captures features of the morning/evening traffic, plus a non-periodic process. Unlike in-fill predictions, in the out-of-sample forecasts the non-periodic events are driven by the prior process and are more sensitive to hyperparameter specifications. In comparison with standard methods, the Lévy random field priors provide excellent performance in terms of both mean squared error

(RMSE = $0.28\mu g/m^3$) and 91.5% coverage for nominal 90% credible intervals for out-of-sample model predictions. Additional details and more examples of the multi-pollutant and spatial-temporal models may be found in Tu *et al.* (2006).

7. SUMMARY

In this paper, we have tried to illustrate the potential of Bayesian nonparametric modelling using Lévy random field priors for function estimation. The model is based on a stochastic expansion of functions in an overcomplete dictionary, which may be formulated as a stochastic integration problem with a (signed) random measure. The unknown function may be approximated as a finite sum of generating functions at arbitrary locations where the number of components is a free parameter. The generator parameters are location-specific and thus are adaptively updated given the data. The adaptability of the generators is especially useful for modeling "spatially" inhomogeneous functions. Unlike many wavelet based methods, there is no requirement that the data are equally spaced. The RJ-MCMC algorithm developed for fitting the models provides an automatic search mechanism for finding sparse representations of a function.

REFERENCES

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. B* **60**, 725–749.

Abramovich, F., Sapatinas, T. and Silverman, B. W. (2000). Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields* **117**, 133–144.

Chakraborty, S., Ghosh, M. and Mallick, B. K. (2004). Bayesian nonlinear regression for large p small n problem. *Tech. Rep.*, 2004-01, University of Florida, USA.

Chilès, J.-P. and Minières, P. D. (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Amer. Statist. Assoc.* **92**, 1413–1421.

Chu, J., Clyde, M. A. and Liang, F. (2006). Bayesian function estimation using an overcomplete continuous wavelet dictionary. Discussion Paper 06–11, ISDS, Duke University.

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. Roy. Statist. Soc. B* **62**, 681–698.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statist. Science* **19**, 81–94.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–401.

Clyde, M. A., House, L. L. and Wolpert, R. L. (2006). Nonparametric models for proteomic peak identification and quantification. *Bayesian Inference for Gene Expression and Proteomics* (K. A. Do, P. Müller and M. Vannucci, eds.). Cambridge: University Press.

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

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998a). Automatic Bayesian curve fitting. *J. Roy. Statist. Soc. B* **60**, 333–350.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998b). Bayesian MARS. *Statist. Computing* **8**, 337–346.

DiMatteo, I., Genovese, C. R. and Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88**, 1055–1071.

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

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–374.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Higdon, D., Swall, J. and Kern, J. (1999). Non-stationary spatial modeling. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 761–768.

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

Karhunen, K. (1946). Zur Spektraltheorie Stochasticher Prozesse. *Ann. Acad. Sci. Fennicae* **34**, 1–7.

Khinchine, A. Y. and Lévy, P. (1936). Sur les lois stables. *C. R. Acad. Sci. Paris* **202**, 374–376.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Applications* **33**, 82–95.

Law, M. H. and Kwok, J. T. (2001). Bayesian support vector regression. *Proc. 8th International Workshop on Artificial Intelligence and Statistics (AISTATS)*. Florida: Key West, 239–244.

Lee, H. K. H., Holloman, C. H., Calder, C. A. and Higdon, D. M. (2002). Flexible gaussian processes via convolution. *Tech. Rep.*, 09–09, ISDS, Duke University, USA.

Loève, M. M. (1955). *Probability Theory*. Princeton: University Press.

Neal, R. M. (1999). Regression and classification using Gaussian process priors. *Bayesian Statistics 6* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) Oxford: University Press, 475–501.

O'Hagan, T. (1978). Curve fitting and optimal design for prediction (with discussion). *J. Roy. Statist. Soc. B* **40**, 1–42.

Rajput, B. S. and Rosiński, J. (1989). Spectral representations of infinitely divisible processes. *Probab. Theory Rel.* **82**, 451–487.

Sato, K.-i. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge: University Press.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–343.

Sollich, P. (2002). Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning* **46**, 21–52.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B* **58**, 267–288.

Tipping, M. E. (2001). Sparse Bayesian learning and the Relevence Vector Machine. *J. Mach. Learn. Res.* **1**, 211–244.

Tu, C. (2006). *Nonparametric Modelling using Lévy Process Priors with Applications for Function Estimation, Time Series Modeling and Spatio-Temporal Modeling*. Ph.D. Thesis, ISDS, Duke University, USA..

Tu, C., Clyde, M. A. and Wolpert, R. L. (2006). Lévy adaptive regression kernels. *Tech. Rep.*, 06–08, ISDS, Duke University, USA.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. New York: Wiley.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia, PA: SIAM.

Wahba, G. (2002). Soft and hard classification by reproducing kernel Hilbert space methods. *Proc. National Academy of Sciences* 99, 524–530.

Wolfe, P. J., Godsill, S. J. and Ng, W.-J. (2004). Bayesian variable selection and regularisation for time-frequency surface estimation. *J. Roy. Statist. Soc. B* **66**, 575–589.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 251–267.

Wolpert, R. L., Ickstadt, K. and Hansen, M. B. (2003). A nonparametric Bayesian approach to inverse problems (with discussion). *Bayesian Statistics 7* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford: University Press, 403–418.

Xia, G. and Gelfand, A. (2005). Stationary process approximation for the analysis of large spatial datasets. *Tech. Rep.*, 05-24, ISDS, Duke University, USA.

Yaglom, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions: Basic Results*, vol. I. Berlin: Springer.

DISCUSSION

BRANI VIDAKOVIC (*Georgia Institute of Technology, USA*)

The authors consider the nonparametric regression problem in which the estimation of a function of interest is done in the atomic decomposition domains by Bayesian modeling. It can be shown that many of currently used models for regularizing functions by shrinkage of coefficients in their atomic decompositions are either special cases or approximations of the Clyde-Wolpert model. Examples include SSVS in orthogonal regression, wavelet thresholding, pursuit methods. The authors construct the prior on the parameters of atoms using Lévy Random Field which leads to tractable posterior simulations. The reversible-jump MCMC method used for efficient computation is well suited because updates based on adding/deleting or updating a single dictionary element at the time. The dictionary elements are only computed when needed which leads to calculational and memory storage efficiency.

Introduction and Overview

The authors are to congratulate for an excellent contribution that is unifying for several state-of-art Bayesian statistical models in the "atomic domains." It can be shown that many of current models for regularizing functions by shrinkage in the atomic domains are either special cases or approximations of the Clyde-Wolpert model. Examples include SSVS in orthogonal regression, Bayesian wavelet shrinkage of various flavors, Bayesian pursuit methods.

To put the ideas of Clyde and Wolpert in the proper setting, we consider a paradigmatic nonparametric regression model. The observations $Y_i, i = 1, \ldots, n$ have two components: a sampled unknown function f as the systematic part and the zero mean random errors as the stochastic part. The errors are assumed to be iid normal with a constant variance, although more general distributions can be considered.

$$Y_i = f(x_i) + \epsilon_i,\ i = 1, \ldots, n;\quad \epsilon_i \sim N(0, \sigma^2).$$

Assume that the function of interest has atomic decomposition

$$f(x_i) = \sum_{0 \le j \le J} g(x_i, \omega_j)\beta_j = \int g(x_i, \omega)\mathcal{L}(d\omega),$$

where $g(x_i, \omega)$ are ω-indexed atoms evaluated at observations x_i and where $\mathcal{L}(d\omega) = \sum \beta_j \delta_{\omega_j}(d\omega)$ is a random signed Borel measure.

The atoms are generated by a function g that could be modulated trigonometric function, wavelet, or a custom made function. For example, if the generator is a wavelet ψ, then the atoms are given by

$$g(x_i, \omega_j) = \sqrt{\lambda_j}\psi(\lambda_j(x_i - \chi_j)),\quad \omega_j = (\chi_j, \lambda_j).$$

where λ_j is a scale and χ_j is a location parameter. The model in (35) is completed by adopting priors on J, β and ω.

How the measure $\mathcal{L}$ was constructed? A rudimentary idea can be traced back to Abramovich *et al.* (2000) who proposed $J \sim Poi(\nu_+)$ and normal prior on β_j. However, their model is not constructive, that is, does not lead to an effective Bayesian simulation. Effective models that use Gaussian random fields are proposed Godsill and his Cambrige team; see Wolfe *et. al.* (2004).

Clyde and Wolpert propose Compound Poisson Representation of $\mathcal{L}$ which is an approximation to Lèvy Measure $\nu(d\beta, d\omega) = \nu_+\pi(d\beta, d\omega)$. This approximation proved to be a constructive. Examples of specific random fields are given in equations (28)-(30).

Additional strength of the paper is wide applicability of the model. The authors provide several examples: (i) wavelet-based function estimation which is described in more detail in Chu *et al.* (2006); (ii) proteomic peak identification that uses the library of normalized Gaussians in peak modeling (research described in Clyde *et al.* (2006); and (iii) non-stationary time series exemplified on measurements of particulate matter from Phoenix, AZ. Here the model is based on Lévy process generated by Gamma random field, as described in Tu (2006).

In the next section we discuss possible construction of calculationally efficient overcomplete dictionaries that potentially can be modeled by compound Poisson representations of $\mathcal{L}$.

Joy and Sorrow of Redundancy

Albert Einstein once said that models should be simple enough but not simpler. The right measure of complexity versus parsimony in modeling is fundamental philosophical issue that directs the development of modeling methodologies. This trade off is in influenced by the "vocabulary of models." An object has parsimonious representation in a rich vocabulary and vice versa, a complex representation in a minimal dictionary.

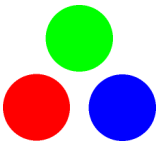

Figure 5: *Three fundamental colors: Red, Green and Blue.*

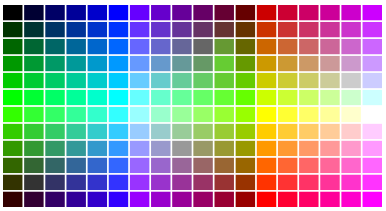

Figure 6: *Pallete with shades*

It is instructive to consider an analogy from chromatic theory. A color painting is to be made and this can be done by one of the two proposed strategies. The first strategy is to paint from a palette with only three basic colors: red, green and blue (RGB, Figure 5), while the second one is paint from a rich palette consisting of many pre-mixed colors of various shades, Figure 6.

The painting can be made using any of the palettes but the redundant palette makes modeling task easier and faster.

There are two camps among atomic modelers (i) *criticalists*: the researchers who strive to produce representations from the (critically) minimal dictionaries, and (ii) *redundantists*, who enjoy benefits of redundancy for more informative and simpler modeling leading to most compact representations. The orthonormal bases are an example of minimal dictionaries: excluding any element from the distionary will make it incomplete, that is, for some objects the representations are not possible.

Here we list several modeling properties in the light of size of dictionaries. Atomic representations with overcomplete dictionaries (OD) (i) enable deeper sparsity, (ii) can contain more meaningful (in the sense of the phenomena they describe) atoms, and (iii) atoms could have nice properties that are impossible in critical dictionaries (symmetry of compactly supported atoms in the wavelet context, shift invariance, more directional information).

On the other hand, critically minimal dictionaries (i) produce unique representations and are mathematically elegant, (ii) are often computationally (and statistically) more convenient, and (iii) have some modeling properties not shared by the overcomplete models; For example, energy preservation is a key property needed in modeling the scaling, long range dependence and (multi)fractality, and controlling the energy distribution in overcomplete models is difficult if possible at all.

We narrow our discussion to construction of dictionaries that are redundant but share some desirable properties of the minimal bases. The dictionaries would be parameterizable in a simple way and thus amenable to Bayesian prior modeling.

It is desirable that the function generating the dictionary is well localized in the time and frequency domains. For most generators this is an impossible requirement. For example, for wavelet bases, the Fourier dual of Haar wavelet is a sinc function which has poor time localization (decays as $O(1/x)$.)

A theoretical result showing the necessity of overcompleteness if locality is important is celebrated Balian-Low theorem (Balian, 1981; Low, 1985). It states that for most minimal atomic expansions, e.g., orthonormal bases or tight frames, simultaneous time/frequency locality of atoms is impossible.

Theorem 1 *Balian-Low Theorem Let g be an L_2 function (window), a and b positive constants, and $(m, n) \in Z \times Z$.*

If the Gabor system of functions (Gabor, 1946)

$\{g_{m,n}(x) = e^{2\pi aimx} g(x - bn)\}$ is a basis, then $ab = 1$, and

$$\left(\int_R |tg(t)|^2 dt\right) \cdot \left(\int_R |\omega \hat{g}(\omega)|^2 d\omega\right) = \infty.$$

For example, it is impossible to form a basis by modulating Gaussian window, $g(x) = e^{-x^2}$.

The solution to Balian-Low obstacle is the redundancy of dictionaries. The examples of standardly used OD are numerous: continuous wavelets, Gabor frames,

non-decimated wavelets, wavelet packets, SLEX bases, complex wavelets, to list a few. The drawback for all OD, as we pointed out, is computational complexity of representing and manipulating the models.

Can we have versatile OD libraries of atoms that retain calculational simplicity of o.n. bases? The answer is positive, and the wavelet packet tables are a nice example. We provide several proposals in the wavelet context and suggest a possibility of Bayesian approaches. The underlying idea is to mix parameterized orthonormal bases.

It is well known that multiresolution analysis by wavelets is fully described by a single filter $\boldsymbol{h}$, called wavelet filter. In fact, the scaling function (informally, the father wavelet) ϕ and its Fourier transform are connected with *transfer function* m_0, and in the sequel, the transfer function uniquely generates the filter $\boldsymbol{h}$. Opposite direction is also possible: given the wavelet filter $\boldsymbol{h}$ one can reconstruct the corresponding scaling function. Schematically,

$$\phi(x) \leftrightarrow \Phi(\omega) \leftrightarrow m(\omega) \leftrightarrow \boldsymbol{h} = (h_0, h_1, \ldots).$$

For example, for Haar scaling function $\phi(x) = \mathbf{1}(0 \le x \le 1)$ the Fourier transform is $\Phi(\omega) = e^{-i\omega/2}\mathrm{sinc}(\omega/2)$, the transfer function is $m_0(\omega) = \frac{1+e^{i\omega}}{2}$, producing $\boldsymbol{h} = \{1/\sqrt{2}\ 1/\sqrt{2}\}$. For details see Vidakovic (1999), page 60.

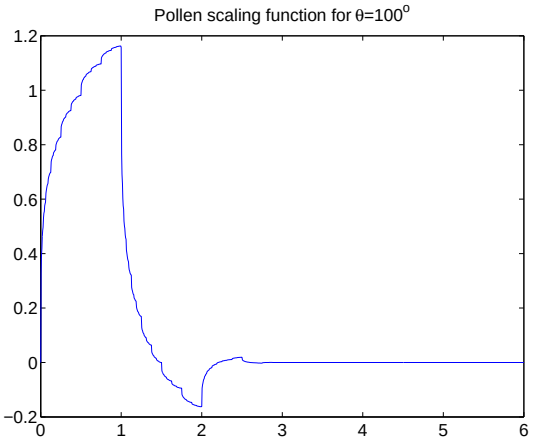

Figure 7: *Scaling function from Pollen family for* $\varphi = 5\pi/9$.

There are many examples of wavelet filters that can be parameterized. Pollen's family (Pollen, 1989) is often used because of its simplicity. An example of four tap Pollen dictionary indexed by a single parameter is provided below. For any value φ from $[0, 2\pi]$ the resulting filter

$$\begin{array}{|l|}\hline h_0 = (1 + \cos(\varphi) - \sin(\varphi))/2^{3/2} \\ h_1 = (1 + \cos(\varphi) + \sin(\varphi))/2^{3/2} \\ h_2 = (1 - \cos(\varphi) + \sin(\varphi))/2^{3/2} \\ h_3 = (1 - \cos(\varphi) - \sin(\varphi))/2^{3/2} \\ \hline \end{array}$$

generates an orthogonal multiresolution analysis, i.e., corresponds to an orthogonal compactly supported wavelet basis. For example, $\varphi = 0$ corresponds to Haar wavelet while $\varphi = \pi/6$ corresponds to Daubechies 4 tap wavelet. Thus the atom $g(x, \omega, \varphi) = \sqrt{\lambda}\phi_\varphi(\lambda x - \chi)$ is parameterized not only by scale λ and location χ, but also by the *shape* parameter φ. Figure 7 shows Pollen scaling function from Shi *et al.* (2005) where it was used the context of wavelet-based classification; the best basis from the Pollen library was selected by entropy minimization had $\varphi = 5\pi/9$. Minent bases (as opposed to maxent priors) are closest to Kahrunen-Loève bases in the sense of "energy packing" and thus produce parsimonious representations.

It would be interesting to let the data (signal) to select φ in Bayesian fashion.

Another interesting wavelet that can produce a versatile library of atoms is the GT wavelet. The following result holds

Theorem 2 *For any* $|b| \geq 1$, $h_0 = -\frac{\sqrt{2}}{2b}$, $h_1 = \frac{\sqrt{2}}{2}$, $h_{2k} = \frac{(b^2-1)\sqrt{2}}{2b^{k+1}}$, $h_{2k+1} = 0$, $k = \pm 1, \pm 2, \ldots$ *is a wavelet filter.*

The GT wavelet has exponential decay and various properties depending on b. For example, if $b = -3$, the scaling function has finite second moment, (Figure 9). while for $b = \frac{1+\sqrt{5}}{2}$ the scaling function is orthogonal not only to its integer shifts, but also to its $1/2$ shifts, $\langle \phi(x), \phi(x + 1/2) \rangle = 0$, Figure 8. If the parameter b is taken into account when modeling, the posterior of b may lead to efficient data justified atoms.

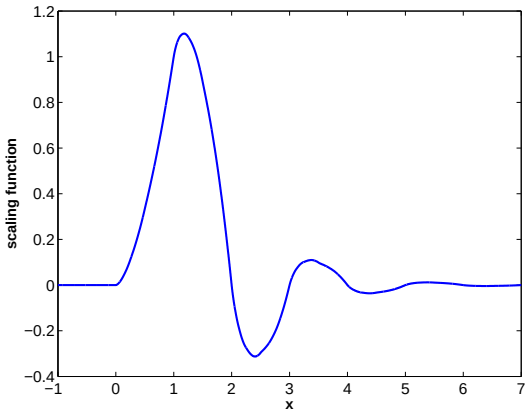

Figure 8: $b = -3$ *has finite* $\phi''(x)$

There is also possibility to form an overcomplete dictionary by manipulating wavelet-filter equations. Consider the system of wavelet-filter equations generating Daubechies minimum phase 4 tap filter. If the standard zero-moment condition $0h_0 - 1h_1 + 2h_2 - 3h_3 = 0$ in the system

$h_0 + h_1 + h_2 + h_3 = \sqrt{2}$
$h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1$
$h_0 h_2 + h_1 h_3 = 0$
$h_0 - h_1 + h_2 - h_3 = 0; \quad 0h_0 - 1h_1 + 2h_2 - 3h_3 = 0$

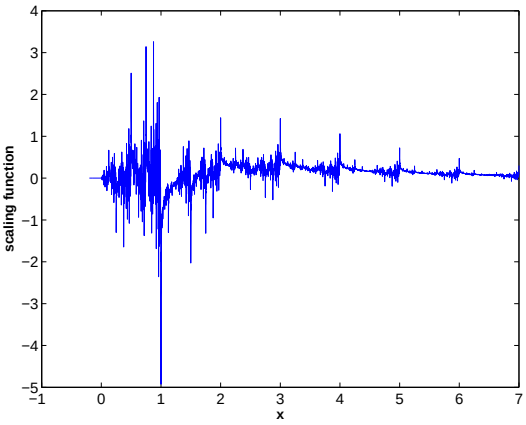

Figure 9: $b = \frac{1+\sqrt{5}}{2}$ *gives* $\langle \phi(x), \phi(x+1/2) \rangle = 0$

is replaced by $h_3 - f(1,c)h_2 + f(2,c)h_1 - f(3,c)h_0 = 0, \quad f(x,c) = e^{-x^2/c}, \ c > 0,$ then each c would correspond to a valid wavelet basis. In the modeling approach a prior could be be placed on c.

We conclude our discussion with the sketch of construction of a custom-made library, via a parameterized wavelet basis with the scaling function matching the features of the data. The steps of the algorithm for such construction are

(i) Select an arbitrary function g appropriate for the modeling problem. In non-parametric regressions the rationale is to select the scaling function that "looks like the data."

(ii) Find Fourier transform of g, $G(\omega) = \hat{g}(\omega)$.

(iii) Normalize G to construct the Fourier transform of a scaling function $\Phi(\omega)$ as

$$\frac{G(\omega)}{\sqrt{\sum_{\ell=-\infty}^{\infty} |G(\omega + 2\ell\pi)|^2}} \, .$$

(iv) Fourier-invert the $\Phi(\omega)$ to obtain scaling function $\phi(t)$. From $\Phi(\omega)$ find the transfer function m_0 and the wavelet filter $\boldsymbol{h}$.

(v) Project the filter $\boldsymbol{h}$ on the space of finite wavelet filters if compact support of atoms is desired and parameterize the wavelet.

Conclusions

The paper by Clyde and Wolpert is a milestone contribution in Bayesian modeling in atomic domains. It unifies several well known atomic estimation strategies and is practicable. In our discussion we focus on the possibility to improve performance by customizing and mixing the dictionaries.

REPLY TO THE DISCUSSION

We would like to begin by thanking Prof. Vidakovic for his enlightening comments on atomic decompositions and extensions of the models that we have presented here. Prof. Vidakovic traces features of our construction back to Abramovich *et al.* (2000). Indeed, their stochastic expansion may be viewed as a special case of our Lévy random field approach, with a Lévy measure constructed as the product of three terms: a proper Gaussian measure on the coefficient parameter β (conditional on the scale parameter λ); a uniform measure on location $\chi \in [0, 1]$; and a potentially improper marginal measure on λ. Even through the coefficients in their expansion have proper conditional distributions, the joint measure may be infinite. As with the infinite Lévy measures presented here, this will lead to a decomposition with an infinite number of jumps or support points in the measure $\mathcal{L}$. Using a finite approximation to the Lévy measure (for example, by truncating the scale parameter's support), one can pursue posterior simulation and inference as described in more detail in Tu *et al.* (2006).

Starting from the stochastic approximations of Abramovich *et al.* (2000), Chu *et al.* (2006) developed posterior inference using continuous wavelet dictionaries. Using the same test functions that were considered in our Example 1, they also found that the overcomplete expansions provided improved MSE performance and more sparse representations when compared to the nondecimated wavelets of Johnstone and Silverman (2005). Comparisons between the results of Chu *et al.* (2006) and Tu *et al.* (2006) suggest that the LARK method leads to improved MSE— however, it remains uncertain whether this is due to the different choices of prior distributions or of generating functions. In applications we have used separable measures for the scaling parameters and coefficients, leading to independent prior distributions; more general non-separable measures (hence non-stationary priors) are worth further investigation.

Combining the heavy tailed priors based on the Lévy random measures (symmetric gamma or stable families) with wavelet generating functions, particularly the parametrized filters and wavelets that Prof. Vidakovic suggests, is an extension worth pursuing. With parametric families of generating functions (such as that of Pollen, 1989, cited by Prof. Vidakovic), one could place a prior distribution on the indexing parameter and provide posterior inference for functions by either selecting an "optimal" family or averaging over families, a form of "model averaging". One could also incorporate the index parameter into the support of the Lévy measure, allowing additional mixing of the dictionaries; in this case each atom used in the expansion may come from any family. Larger dictionaries could of course be created by mixing over wavelets and kernels. While certainly leading to more flexible representations, the increased computational complexity may become overwhelming. Parametrized generating functions adapted to features of the data offer a more promising avenue for exploration.

While the computational challenges that arise with the Lévy random field priors and wavelet generating functions are straightforward to overcome, there remains the interesting theoretical question of which function space contains the limiting infinite stochastic expansions for infinite Lévy measures. Abramovich *et al.* (2000) prove that the stochastic expansion will remain in the Besov space of the generating wavelet, in their Gaussian context, but for more general Lévy measures and wavelets this remains an open question.

ADDITIONAL REFERENCES IN THE DISCUSSION

Abramovich, F., Sapatinas, T. and Silverman, B. W. (2000). Stochastic expansions in an overcomplete wavelet dictionary. *Probability Theory and Related Fields* **117**, 133–144.

Balian, R. (1981). A strong uncertainty principle in signal theory or in quantum mechanics. *C. R. Acad. Sci.* **292**, 1357–1362.

Gabor, D. (1946). Theory of communication, *J IEE London* **93**, 429–457.

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

Low, F. (1985). Passion for Physics. *Essays in Honor of Geoffrey Chew.* (C. DeTar *et al.*, eds.) Singapore: World Scientific, 17.

Pollen, D. (1989). Parametrization of compactly supported wavelets. *Tech. Rep.*, Aware Inc, USA.

Shi, B., Moloney, K. P., Pan, Y., Leonard, V. K., Vidakovic, B., Jacko, J. and Sainfort, F. (2006). Classification of high frequency pupillary responses using Schur monotone descriptors in multiscale domains. *J. Statist. Computation and Simulation* **76**, 431–446.