Bayesian Nonparametric Models using Levy Random Fields and Overcomplete Dictionaries

Merlise Clyde

Department of Statistical Science
Duke University

WISE June 5, 2011

Problem Setting

Consider the nonparametric regression problem where we observe noisy measurements $\{Y_i\} \quad i \in I$ of an unknown function $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$

$$E[Y \mid \mathbf{x}] = f(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}$$

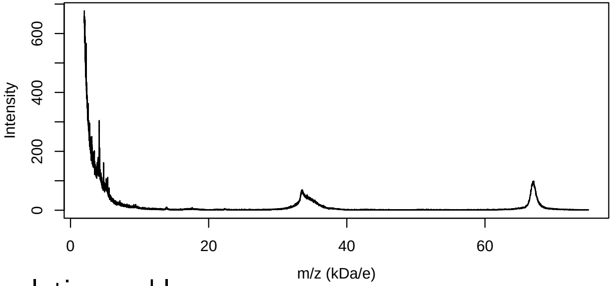
Need flexible prior distributions on functions
Usual Suspects:

- Gaussian Process Priors
- Dirichlet Process priors
- Expansions of f

Focus on Lévy Processes! (related to GP & DP)

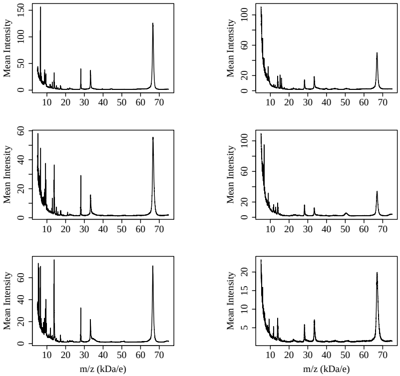
Goals

- "Machine Learning" Classification/Prediction problems
- Learning "features" of f
  - which variables in $\mathbf{x}$ are important for classification
  - location of peaks/proteins in spectra
  - sources/spread of pollutants in space/time models



Deconvolution problem

Multiple Spectra



Learning "features" that are common versus those that separate groups of functions (spectra)

Stochastic Expansions

Expand $f(\mathbf{x}_i) = \sum_{j=0}^{J} \psi_j(\mathbf{x}_i)\beta_j$ in terms of an Overcomplete Dictionary where

- $\{\psi_j\}$: dictionary elements
- $\{\beta_j\}$: unknown coefficients
- J: number of elements in expansion (finite and infinite)

Advantages: may lead to more flexibility in choosing "building" blocks to match features than GP or DP models.

How should we choose prior distributions so that the resulting f is well defined with desired smoothness properties?

Finite Expansions

Consider finite expansions for some collection J of dictionary elements ψ_j

$$f(\mathbf{x}) = \sum_{j \le J} \psi_j(\mathbf{x})\beta_j \qquad \{\psi_j \in \mathcal{F}\}$$

Independent scale mixtures of normals (Generalized Ridge Priors)

$$\begin{aligned} \beta_j \mid \varphi_j &\overset{ind}{\sim} \mathsf{N}(0, \varphi_j^{-1}) \\ \varphi_j &\overset{iid}{\sim} \mathsf{G}(a, b) \end{aligned}$$

Tipping considers modal estimates in the case $a = b = 0$

Improper prior and posterior!

Model Selection Priors

For selection of dictionary elements from a Bayesian perspective, add a point mass at zero in the distribution for β_j

$$\begin{aligned}
\beta_j \mid \varphi_j &\overset{ind}{\sim} \mathrm{N}(0, \gamma_j \varphi_j^{-1}) \\
\varphi_j &\overset{iid}{\sim} \mathrm{G}(a, b) \\
\gamma_j &\overset{iid}{\sim} \mathrm{Ber}(\pi)
\end{aligned}$$

- Choice of global scale as J increases?
- Choice of π as J increases?
- Number of non-zero coefficients is Binomial with mean $J\pi$. Limit as $J \to \infty$ such that $J\pi \to \nu_+$ is Poisson with mean ν_+

Consider model to unify finite and infinite dimensional models!

Lévy Priors

Hierarchical Student-t prior

$$\begin{aligned} \beta_j \mid \varphi_j &\overset{ind}{\sim} \mathsf{N}(0, \varphi_j^{-1}) \\ \varphi_j &\overset{ind}{\sim} \mathsf{G}\left(\frac{\alpha}{2}, \frac{\alpha\epsilon^2}{2}\right) \\ J &\sim \mathsf{P}(\nu_\epsilon^+) \end{aligned}$$

where $\nu_\epsilon^+ = \nu_\epsilon(\mathbb{R}) = \frac{\alpha^{1-\alpha/2}\Gamma(\alpha)\Gamma(\alpha/2)}{\epsilon^\alpha \pi^{1/2}\Gamma(\frac{\alpha+1}{2})} \sin(\frac{\pi\alpha}{2})$

Limit as $\epsilon \to 0$ leads to Lévy *alpha*-Stable process!

Lévy Adaptive Regression Kernels

Stochastic Integral Representation

$$f(x) = \sum_{j \leq J} \psi(\mathbf{x}; \boldsymbol{\omega}_j)\beta_j \equiv \int_{\boldsymbol{\Omega}} \psi(x; \boldsymbol{\omega})\mathcal{L}(d\boldsymbol{\omega})$$

$$\psi(\mathbf{x}, \boldsymbol{\omega}_j) \equiv g(\boldsymbol{\Lambda}_j(\mathbf{x} - \boldsymbol{\chi}_j)) \quad \text{"generator"}$$

$\mathcal{L}$ is a Signed Measure:

$$\mathcal{L}(d\boldsymbol{\omega}) = \sum_{j \leq J} \beta_j \delta_{\boldsymbol{\omega}_j}(d\boldsymbol{\omega})$$

- support points of $\mathcal{L}$: $\{\boldsymbol{\omega}_j\} = \{\boldsymbol{\chi}_j, \boldsymbol{\lambda}_j\}$
  - "location" parameters: $\boldsymbol{\chi}_j \in \mathcal{X}$
  - "scaling" parameters: $\boldsymbol{\lambda}_j \in \mathbb{R}^+$
- jump sizes of measure: β_j
- number of support points J

Generators

Dictionary elements generated by $\psi(\mathbf{x}, \omega_j) \equiv g(\mathbf{\Lambda}_j(\mathbf{x} - \chi_j))$
translation and scaling as in wavelets

- kernels (as in kernel regression or SVM)
- densities of location-scale families Gaussian or Cauchy kernel (for mass spect)
- exponential densities (pollutant concentrations)
- wavelet families

No need for symmetric kernels as in SVM
Generates continuous dictionaries

Lévy Random Fields

- $\mathcal{L}(d\omega)$ is a random (signed) measure on $\mathbf{\Omega}$
- Convenient to think of a random measure as stochastic process where $\mathcal{L}$ assigns random variables to sets $A \in \mathbf{\Omega}$
- Take

$$\mathcal{L} \sim \mathrm{Lv}(\nu) \text{ with Lévy measure } \nu(d\beta, d\boldsymbol{\omega})$$

where ν satisfies integrability condition:

$$\int_{\mathbb{R}\times\Omega} \min(1, \beta^2)\, \nu(d\beta, d\boldsymbol{\omega}) < \infty \tag{1}$$

Poisson Representation of Lévy Random Fields is the key to Bayesian Inference!

Poisson Representation

Goal: $f(x) = \sum_{j<J} \psi(\mathbf{x}, \boldsymbol{\omega}_j)\beta_j = \sum_{j<J} g(\boldsymbol{\Lambda}_j(\mathbf{x} - \chi_j))$

Sufficient condition for bounded g:

$$\int_{\mathbb{R}\times\Omega} \min(1, |\beta|)\nu(d\beta, d\boldsymbol{\omega}) < \infty \tag{2}$$

$$\Rightarrow \; J \sim \mathsf{P}(\nu_+), \qquad \nu_+ \equiv \nu(\mathbb{R} \times \boldsymbol{\Omega})$$

$$\Rightarrow \; \beta_j, \boldsymbol{\omega}_j \mid J \overset{iid}{\sim} \pi(d\beta, d\boldsymbol{\omega}) \propto \nu(d\beta, d\boldsymbol{\omega}).$$

- Finite number of "big" coefficients $|\beta_j|$
- Possibly infinite number of $\beta \in [-\epsilon, \epsilon]$
- Coefficients $|\beta_j|$ are absolutely summable[1]

[1] need to add a term to "compensate" the infinite number of tiny jumps that are not absolutely summable under the more general integrability condition Equation (1)

Existence Theorem

Theorem

Let ν be a Lévy measure on $\mathbb{R} \times \Omega$ satisfying (1). Then $f(\mathbf{x})$ is well-defined if $\psi(\omega) \equiv g(\mathbf{\Lambda}(\cdot - \chi))$ satisfies

$$\iint_{[-1,1]^c \times \Omega} (1 \wedge |\beta \psi(\omega)|)\, \nu(d\beta\ d\omega) < \infty \tag{3a}$$

$$\iint_{[-1,1] \times \Omega} (|\beta \psi(\omega)| \wedge |\beta \psi(\omega)|^2)\, \nu(d\beta\ d\omega) < \infty. \tag{3b}$$

For ν satisfying Equation (2) the condition simplifies

$$\iint_{\mathbb{R} \times \Omega} (1 \wedge |\beta \phi(\omega)|)\, \nu(d\beta\ d\omega) < \infty \tag{3c}$$

g is in a "Musielak-Orlicz space"

Function Spaces

The Besov space $\mathbb{B}^s_{pq}$ consists of those $f \in L_p(\mathbb{R}^d)$ whose Besov semi-norms are finite

$$\|f\|^s_{pq} = \|f\|_p + |f|^s_{pq} < \infty \tag{4}$$

For $p, q \geq 0$ and $s > d(1/p - 1)_+$ and for any integer $m > s$ set

$$|f|^s_{pq} = \left(\int_{|h| \leq 1} |h|^{-sq} \, \|\Delta^m_h f\|^q_p \; dh/|h|^d \right)^{1/q}$$

where Δ^m_h denotes the mth forward finite difference,

$$\begin{aligned} \Delta^0_h f(x) &= f(x) \\ \Delta^m_h f(x) &= \left[\Delta^{m-1}_h f(x+h) - \Delta^{m-1}_h f(x)\right] \\ &= \sum_{k=0}^{m} \binom{m}{k} (-1)^{m-k} f(x+kh). \end{aligned} \tag{5}$$

LARK Models and Besov Spaces

Let $f(\mathbf{x}) = \int_\Omega \psi(\mathbf{x}, \omega)\mathcal{L}(d\omega)$ where $\psi(\mathbf{x}, \boldsymbol{\omega}) = g(\boldsymbol{\Lambda}(\mathbf{x} - \chi))$ and $\mathcal{L} \sim \mathrm{Lv}(\nu)$

Theorem

Fix $g \in \mathbb{B}^s_{pq}(\mathbb{R}^d)$ for some $p, q \geq 1$ and $s > 0$ and a Lévy measure ν on $\mathbb{R} \times \Omega$ with $\Omega = (\mathcal{S}^d_+ \times \mathbb{R}^d)$ of translation-invariant product form $\nu(d\beta\ d\omega) = \tilde{\nu}(d\beta\, d\boldsymbol{\Lambda})d\chi$ for a σ-finite measure $\tilde{\nu}(d\beta\, d\boldsymbol{\Lambda})$ on $\mathbb{R} \times \mathcal{S}^d_+$ that satisfies the integrability condition Equation (2) and the Existence Theorem. Then $f \in \mathbb{B}^s_{pq}$ almost surely if $\tilde{\nu}$ satisfies:

$$L_p : \quad \iint_{\mathbb{R}\times\mathcal{S}^d_+} \left(1 \wedge |\beta||\boldsymbol{\Lambda}|^{-1/p}\right) \tilde{\nu}(d\beta\, d\boldsymbol{\Lambda}) < \infty \tag{6a}$$

$$\mathbb{B}^s_{pq} : \quad \iint_{\mathbb{R}\times\mathcal{S}^d_+} \left(1 \wedge |\beta||\boldsymbol{\Lambda}|^{s-1/p}\right) \tilde{\nu}(d\beta\, d\boldsymbol{\Lambda}) < \infty. \tag{6b}$$

Lévy Measures

α-Stable measure: $\nu(d\beta, d\boldsymbol{\omega}) = c_\alpha |\beta|^{-(\alpha+1)} \, \gamma(d\boldsymbol{\omega})$
For α- Stable $\nu^+(\mathbb{R}, \boldsymbol{\Omega}) = \infty$
Fine in theory, but not in practice for MCMC!

Truncate measure to obtain a finite expansion:

- The random number of support points $\boldsymbol{\omega}$ with β in $[-\epsilon, \epsilon]^c$ is finite
- Fix ϵ (practical significance)
- Use approximate Lévy measure

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

$\Rightarrow$ $J \sim \mathsf{P}(\nu_\epsilon^+)$ where $\nu_\epsilon^+ = \nu([-\epsilon, \epsilon]^c, \boldsymbol{\Omega})$

$\Rightarrow$ $\beta_j, \boldsymbol{\omega}_j \overset{iid}{\sim} \pi(d\beta, d\boldsymbol{\omega}) \equiv \nu_\epsilon(d\beta, d\boldsymbol{\omega}) / \nu_\epsilon^+$

Approximate Lévy Prior

Continuous Approximation:

$$\nu_\epsilon(d\beta, d\omega) = c_\alpha(\beta^2 + \alpha\epsilon^2)^{-(\alpha+1)/2} d\beta \; \gamma(d\omega)$$

Based on the following hierarchical prior

$$\begin{aligned} \beta_j \mid \phi_j &\stackrel{ind}{\sim} \mathrm{N}(0, \varphi_j^{-1}) \\ \phi_j &\stackrel{ind}{\sim} \mathrm{G}\left(\frac{\alpha}{2}, \frac{\alpha\epsilon^2}{2}\right) \\ J &\sim \mathrm{P}(\nu_\epsilon^+) \end{aligned}$$

where $\nu_\epsilon^+ = \nu_\epsilon(\mathbb{R}, \boldsymbol{\Omega}) = \frac{\alpha^{1-\alpha/2}\Gamma(\alpha)\Gamma(\alpha/2)}{\epsilon^\alpha \pi^{1/2}\Gamma(\frac{\alpha+1}{2})} \sin(\frac{\pi\alpha}{2})\gamma(\boldsymbol{\Omega})$

Advantage: Conjugate prior so β can be integrated out for MCMC

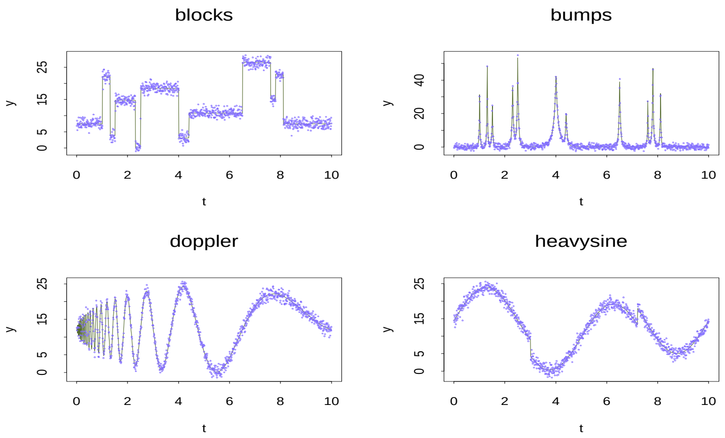
Limiting Case

$$\beta_j \mid \varphi_j \overset{ind}{\sim} \mathrm{N}(0, 1/\varphi_j)$$
$$\varphi_j \overset{iid}{\sim} \mathrm{G}(\alpha/2, "0")$$

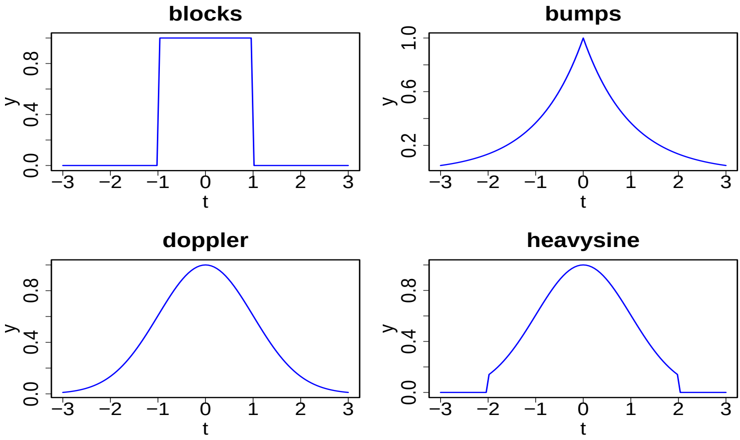
Notes:

- Require $0 < \alpha < 2$ for characteristic function for $\mathcal{L}$ and functionals to exist.
- Cauchy corresponds to $\alpha = 1$
- Tipping's choice corresponds to $\alpha = 0$
- Provides an extension of Generalized Ridge Priors to infinite dimensional
- Infinite dimensional analog of Cauchy priors

Wavelet Test Functions (SNR = 7)



Kernel Functions



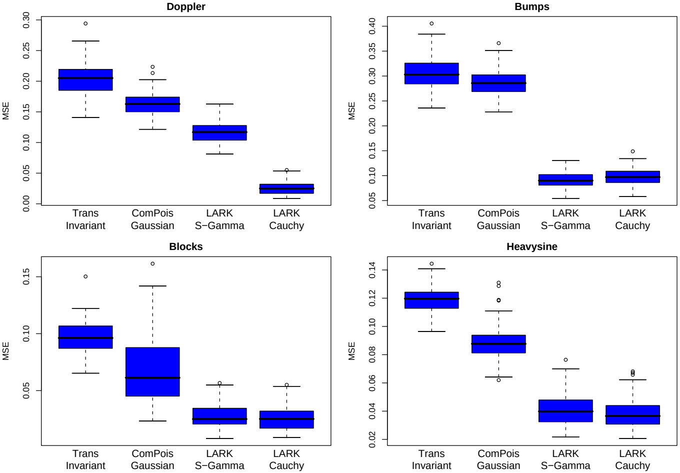
Comparisons of OCD Methods

- Translational Invariant Wavelets – Laplace Priors (Johnstone & Silverman 2005)
- Continuous Wavelet Dictionary – Compound Poisson with Gaussian Priors (Chu, Clyde, Liang 2007)
- LARK Symmetric Gamma
- LARK Cauchy

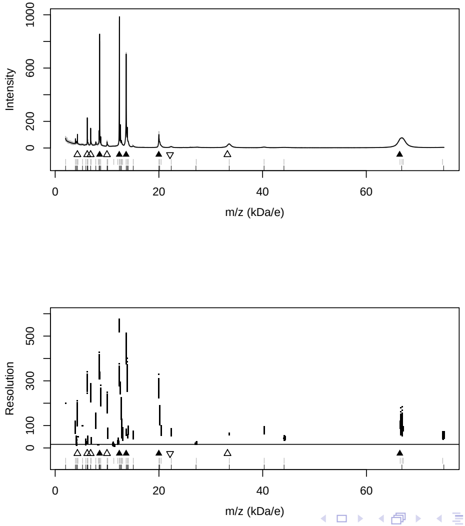
Range of Over-complete Dictionaries and Priors

Comparison of Mean Square Error w/ OCDs

100 realizations of each function



MALDI-TOF Mass Spectroscopy



Higher Dimensional $\mathcal{X}$

MCMC is too slow to allow

- location χ to be arbitrary; restrict to observed $\{\mathbf{x}_i\}$
- scale parameter to vary with location; use common Λ
- arbitrary Λ; restrict to diagonal Λ

$$
\begin{aligned}
k(\mathbf{x}, \omega_j) &= \prod_d \exp\{-\lambda_d (x_d - x_{jd})^2\} \\
f(\mathbf{x}) &= \sum_j k(\mathbf{x}, \omega_j)\beta_j
\end{aligned}
$$

- Product structure allows interactions between variables
- Many input variables may be irrelevant
- Feature selection; if $\lambda_d = 0$ variable x_d is removed

Regression Out of Sample Prediction

Average Relative MSE to best procedure

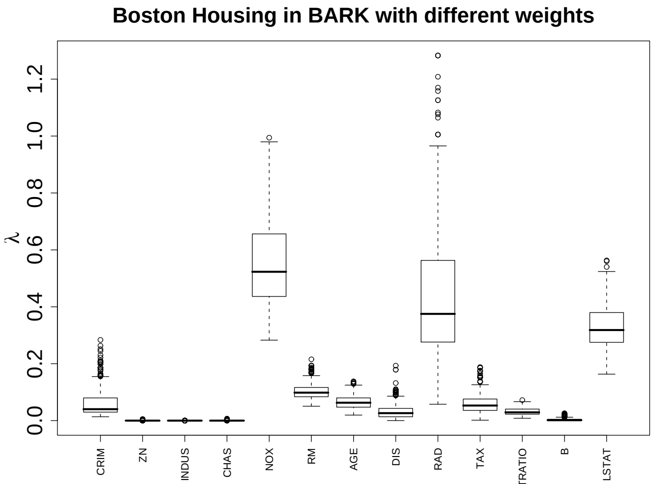
Data Sets	BARK			SVM	BART
	D	S + E	S + D		
Friedman1	1.22	2.26	1.93	5.36	1.97
Friedman2	1.07	1.09	1.04	4.36	3.64
Friedman3	1.46	2.30	1.44	2.70	1.00
Boston Housing	1.09	1.23	1.20	1.56	1.01
Body Fat	1.81	1.01	2.19	4.04	1.68
Basketball	1.01	1.01	1.02	1.16	1.10

D: dimension specific scale λ_d

E: equal scales $\lambda_d = \lambda \forall d$

S: selection $\lambda_d = 0$ with probability ρ

Posterior Distribution of λ_d



Classification Examples

Name	d	data type	n (train/test)
Circle	2	simulation	200/1000
Circle (3 null)	5	simulation	200/1000
Circle (18 null)	20	simulation	200/1000
Swiss Bank Notes	6	real data	200 (5 *cv*)
Breast Cancer	30	real data	569 (5 *cv*)
Ionosphere	33	real data	351 (5 *cv*)

- Add latent Gaussian Z_i for probit regression (as in Albert & Chib)
- Same model as before conditional on $\mathbf{Z}$
- Advantage: Draw β in a block from full conditional

Predictive Error Rate for Classification

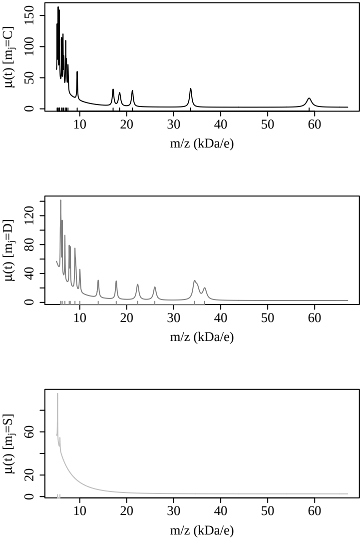
Data Sets	BARK			SVM	BART
	D	S + E	S + D		
Circle 2	4.91%	1.88%	1.93%	5.03%	3.97%
Circle 5	4.70%	1.47%	1.65%	10.99%	6.51%
Circle 20	4.84%	2.09%	3.69%	44.10%	15.10%
Bank	1.25%	0.55%	0.88%	1.12%	0.50%
BC	4.02%	2.49%	6.09%	2.70%	3.36%
Ionosphere	8.59%	5.78%	10.87%	5.17%	7.34%

D: dimension specific scale λ_d

E: equal scales $\lambda_d = \lambda \forall d$

S: selection $\lambda_d = 0$ with probability ρ

Multiple Spectra



Decompose into Control,

Diseased + Shared functions, extending ω to include “marks”

Summary

Lévy Random Field Priors & LARK models:

- Provide limit of finite dimensional priors (GRP & SVSS) to infinite dimensional setting
- Adaptive bandwidth for kernel regression
- Allow flexible generating functions (non-parametric)
- Provide sparse representations compared to SVM & RVM, with coherent Bayesian interpretation
- Incorporation of prior knowledge if available
- relax assumptions of equally spaced data
- Hierarchical Extensions
- Formulation allows one to define stochastic processes on arbitrary spaces (spheres, manifolds) periodicities

Open problems – rates of convergence!