Towards Objective Priors in Nonparametric Regression and Classification

Merlise Clyde
(joint with Robert Wolpert and Zhi Ouyang)

Department of Statistical Science
Duke University

10th Annual Winter Workshop
Bayesian Model Selection and Objective Methods
Jan 11-12, 2008

Problem Setting

Consider the nonparametric regression problem with data $\{Y_i, \mathbf{x}_i\} \quad i = 1, \ldots n$

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

Prior Distributions on f:

- Gaussian Process Priors
- Dirichlet Process priors
- Expansions of f (finite and infinite)

Basis Expansions

Need to place a prior distribution on unknown function $f \in \mathcal{F}$
Expansions $f(\mathbf{x}_i) = \sum_j \psi_j(\mathbf{x}_i)\beta_j$

- $\{\psi_j\}$: basis functions for some function space $\mathcal{F}$
- $\{\beta_j\}$ unknown coefficients
- Commonly used basis functions:
  - Polynomials
  - Fourier
  - Splines
  - Wavelets
  - Kernels

Over-complete Dictionaries (OCD)

In recent years OCD have received considerable attention

- Collection $\{\psi_j(\mathbf{x})\}$ "more than a basis"
- Examples:
  - "Large p, small n"
  - Unions of two (or more) bases
  - Translation Invariant Wavelets
  - Free-knot splines
  - Gabor frames
  - Kernels: $\psi_j(\mathbf{x}) = k(\mathbf{x}; \boldsymbol{\omega}_j)$ with kernel specific scale & location parameters
- Expand f in terms of OCD

$$f(\mathbf{x}_i) = \sum_{j \in \mathcal{J}} \psi_j(\mathbf{x}_i)\beta_j, \qquad f \in \mathbb{F} = \overline{\{\psi_j\}}$$

Why Over-complete Dictionaries?

+ More flexible - local adaptivity
+ Potential for sparse representations

− Non-unique coefficients
− Computationally intensive search over (uncountable) dictionary

+/- If we are careful, can use improper priors (!) (at least in theory)

Prior Distributions

Consider finite expansions for some collection of dictionary elements ψ_j

$$\begin{aligned} f(\mathbf{x}) &= \sum_{j \le J} \psi_j(\mathbf{x}) \beta_j \qquad \{\psi_j \in \mathbb{F}\} \\ \mathbf{f} &= \boldsymbol{\Psi} \boldsymbol{\beta} \end{aligned}$$

Choice of prior distribution on β_j

- g-priors and mixtures of g priors (Zellner-Siow Cauchy priors)
- Independent normal or mixtures of normals

g-priors

Zellner-Siow Cauchy Prior:

$$\begin{aligned} \beta \mid \mathbf{\Psi} &\sim \mathrm{N}(0, g\sigma^2(\mathbf{\Psi}'\mathbf{\Psi})^-) \\ g &\sim \mathrm{G}(1/2, n/2) \\ p(\sigma^2) &\propto 1/\sigma^2 \end{aligned}$$

+ Prior on f invariant to choice of basis
- Bayes factors break down if $\mathrm{rank}(\mathbf{\Psi}) = n$ (cannot distinguish model from null model)
- Consistent with $\mathbf{f} \sim \mathbf{N}(\mathbf{0}, \mathbf{g}\sigma^2\mathbf{I_n})$

Independent Priors

Independent normals and scale mixtures of normals used by

- Silverman & Johnstone (wavelets)
- Tipping (relevance vector machines)
- Chakraborty, Ghosh & Mallick (large p, small n nonlinear regression)

$$\beta_j \mid \phi_j \overset{ind}{\sim} \mathrm{N}(0, \varphi_j^{-1})$$
$$\phi_j \overset{iid}{\sim} \mathrm{G}(a, b)$$

Tipping considers modal estimates in the case $a = b = 0$ (improper prior/posterior)

What about the infinite dimensional case $J \to \infty$?

Lévy Adaptive Regression Kernels (Clyde & Wolpert 2007)

Representation

$$f(x) = \sum_{j \le J} k(\mathbf{x}; \omega_j)\beta_j \equiv \int_{\boldsymbol{\Omega}} k(x; \omega)\mathcal{L}(d\omega)$$

Gaussian kernel: $k(\mathbf{x}, \omega_j) = \exp\{-(\mathbf{x} - \chi_j)'\boldsymbol{\Lambda}_j(\mathbf{x} - \chi_j)\}$
$\mathcal{L}$ is a Signed Measure:

$$\mathcal{L}(d\omega) = \sum_{j \le J} \beta_j \delta_{\omega_j}(d\omega)$$

- support points of $\mathcal{L}$: $\{\omega_j\} = \{\chi_j, \boldsymbol{\Lambda}_j\}$
  - "location" of kernel: $\chi_j \in \mathcal{X}$
  - "scale" of kernel: $\boldsymbol{\Lambda}_j \in \mathbb{R}^+$
- jump sizes of measure: β_j
- number of support points J

Lévy Random Fields

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

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

where ν satisfies integrability condition:

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

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

Poisson Representation

Goal: $f(x) = \sum_{j<J} k(\mathbf{x}, \boldsymbol{\omega}_j)\beta_j$

Sufficient condition:

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

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

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

- Finite number of "big" coefficients $|\beta_j|$
- Possibly infinite number of $\beta \in [-\epsilon, \epsilon]$
- Jumps $|\beta_j|$ 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

Lévy Measures

α-Stable measure: $\nu(d\beta, d\omega) = c_\alpha |\beta|^{-(\alpha+1)}\ \gamma(d\omega)$

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

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 a generalization of Generalized Ridge Priors to infinite dimensional
- Infinite dimensional analog of Cauchy priors

Approximating Lévy Random Fields

For α- Stable $\nu^{+}(\mathbb{R}, \boldsymbol{\Omega}) = \infty$

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

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

$\Rightarrow$ $J \sim \mathsf{P}(\nu_\epsilon^{+})$ 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:

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

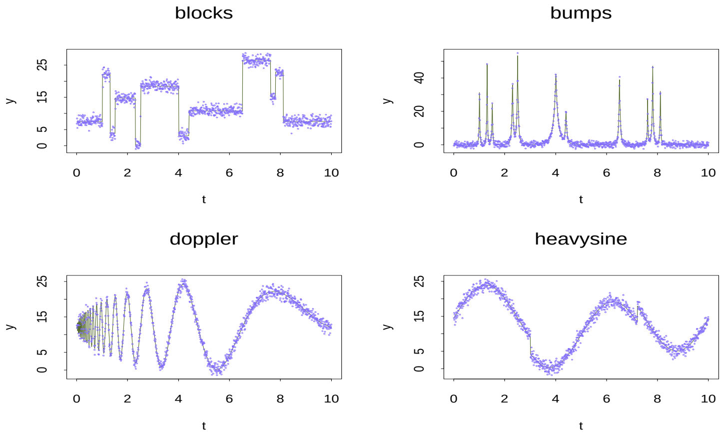
Based on the following hierarchical prior

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

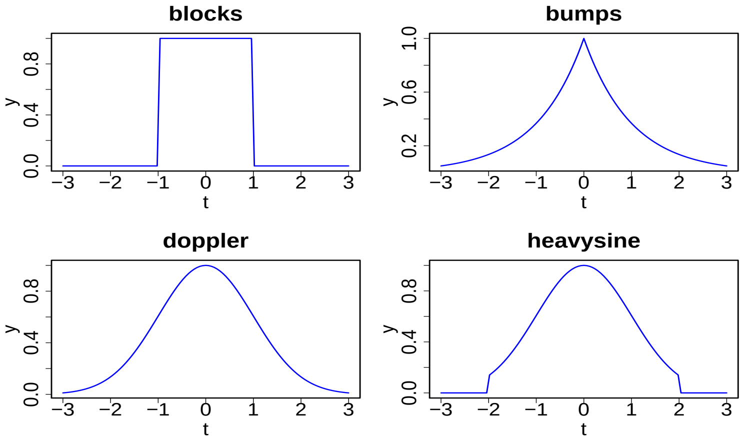
where $\nu+_\epsilon = \nu_\epsilon(\mathbb{R}, \boldsymbol{\Omega}) = \frac{\alpha^{1-\alpha/2}\Gamma(\alpha)\Gamma(\alpha/2)}{\epsilon^\alpha \pi^{1/2}\Gamma(\frac{\alpha+1}{2})} \sin(\frac{\pi\alpha}{2})\gamma(\boldsymbol{\Omega})$

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

Wavelet Test Functions (SNR = 7)



Kernel Functions



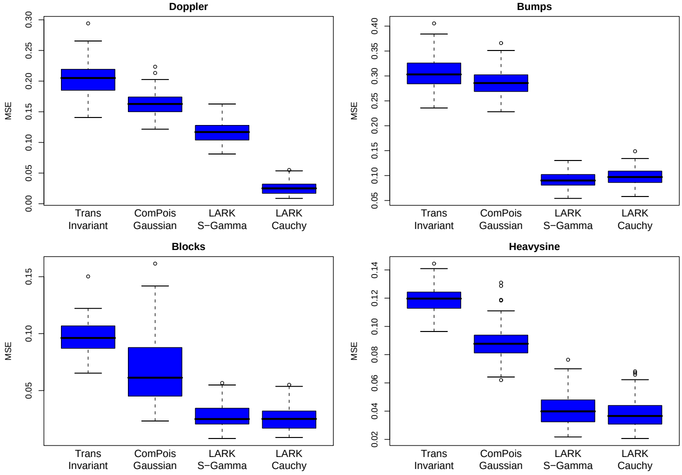
Comparisons of OCD Methods

- 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



Higher Dimensional $\mathcal{X}$

MCMC is too slow to allow

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

$$k(\mathbf{x}, \omega_j) = \prod_d \exp\{-\lambda_d (x_d - x_{jd})^2\}$$

$$f(\mathbf{x}) = \sum_j k(\mathbf{x}, \omega_j)\beta_j$$

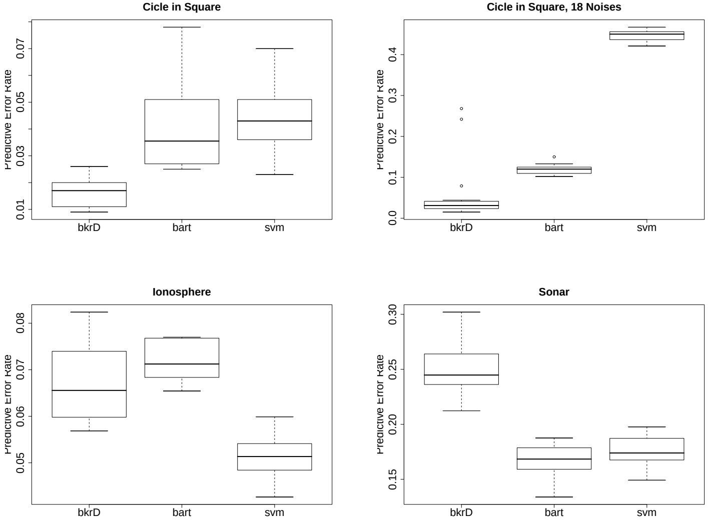
- Product structure allows interactions between variables
- Many input variables may be irrelevant
- Feature selection; as $\lambda_d \to 0$ variable x_d is removed

Classification Examples

Name	d	data type	n (train/test)
Circle	2	simulation	200/1000
Circle (3 null)	5	simulation	200/1000
Circle (8 null)	10	simulation	200/1000
Circle (18 null)	20	simulation	200/1000
Ionosphere	33	real data	351 (5 *cv*)
Sonar	60	real data	208 (5 *cv*)

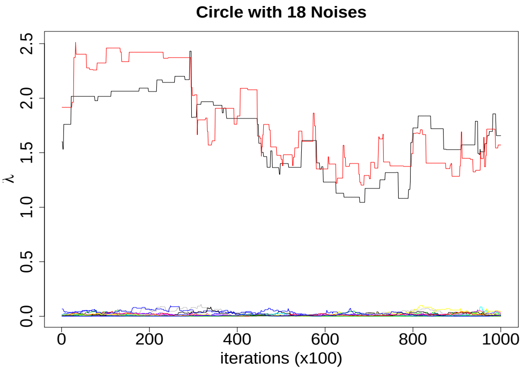
- Add latent Gaussian Z_i for probit regression (as in Albert & Chib)
- Same model as before conditional on $\mathbf{Z}$
- Advantage: Draw β in a block from full conditional

Error Rate for Classification



Feature Selection

Trace plots of λ_d for Circle in Square with 18 null predictors



Regression Examples

Name	d	n (train/test)	Comparison MSE
Friedman 1	10	200/1000	BART<LARK<SVM
Friedman 2	4	200/1000	LARK<BART<SVM
Friedman 3	4	200/1000	BART<LARK<SVM
BostonHousing	13	506 (5 *cv*)	BART<LARK<SVM
Bodyfat	14	252 (5 *cv*)	BART<LARK<SVM
Basketball	4	96 (5 *cv*)	LARK<BART<SVM
Spouse	21	11136/11136	too slow to run

Summary

Lévy Random Field Priors & LARK models:

- Provide limit of finite dimensional priors (GRP & SVSS) to infinite dimensional setting
- Proper posterior distribution
- Allow flexible generating functions (non-parametric)
- Provide sparse representations compared to SVM & RVM

On going work:

- Port to C
- Improve MCMC to allow adaptive λ_{dj} in higher dimensional problems (local interactions & feature selection)