Nonparametric Function Estimation Using Overcomplete Dictionaries

Merlise Clyde[1]

Institute of Statistics and Decision Sciences
Duke University

Graybill Conference on MultiScale Methods
June 12, 2006

[1]Thanks to JenHwa Chu, Leanna House, Feng Liang, Chong Tu & Robert Wolpert

Problem Setting

- Model:
  - Observe data $\{Y_i, \mathbf{x}_i\} \quad i = 1, \ldots n$
  - $\mathrm{E}[Y \mid \mathbf{x}] = f(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}$
  - Goal: inference about unknown $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}$
- Examples:
  - Proteomics: $\mathbf{Y}_i$ is the measured intensity at Time-of-Flight t_i in MALDI-TOF mass spectroscopy
  - Criteria Pollutants: $\mathbf{Y}_i$ is the measured concentration of pollutants at time t_i and spatial location s_i.
- Features:
  - Non-negative intensity
  - Non-stationarity

Nonparametric Regression

Need to place a prior on unknown function $f \in \mathcal{F}$

- Dirichlet and Spatial Dirichlet Process priors
- Gaussian Process Priors

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, Kernels, Wavelets, etc.

Lévy Process (Random Field) Priors

Nonparametric Regression

Need to place a prior on unknown function $f \in \mathcal{F}$

- Dirichlet and Spatial Dirichlet Process priors
- Gaussian Process Priors

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, Kernels, Wavelets, etc.

Lévy Process (Random Field) Priors

Which Basis Expansion?

Representations with (orthonormal) bases may be inefficient ...

- Canonical basis:

$$f = \begin{pmatrix} 42 \\ 42 \\ 42 \end{pmatrix} = 42 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 42 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 42 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

- Enlarging the *dictionary* with element $(1, 1, 1)'$ allows parsimony:

$$f = 42 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 0 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 0 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + 0 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + 0 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

Which Basis Expansion?

Representations with (orthonormal) bases may be inefficient ...

- Canonical basis:

$$f = \begin{pmatrix} 42 \\ 42 \\ 42 \end{pmatrix} = 42 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 42 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 42 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

- Enlarging the *dictionary* with element $(1, 1, 1)'$ allows parsimony:

$$f = 42 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} + 0 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + 0 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + 0 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + 0 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + 0 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}$$

Overcomplete Dictionaries (OCD)

- 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: $\psi_j(\mathbf{x}) = k(\mathbf{x}; \boldsymbol{\omega})$ or other generating functions
- Expand f in terms of OCD

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

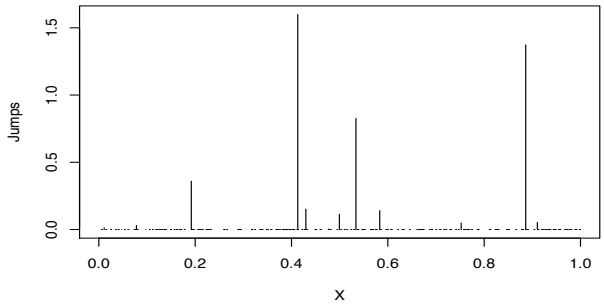
Lévy Adaptive Regression Kernels

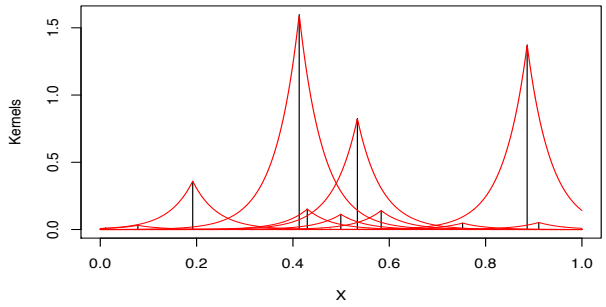
Generating function: $g(\mathbf{x}, \omega) : \mathcal{X} \times \mathbf{\Omega} \to \mathbb{R}$

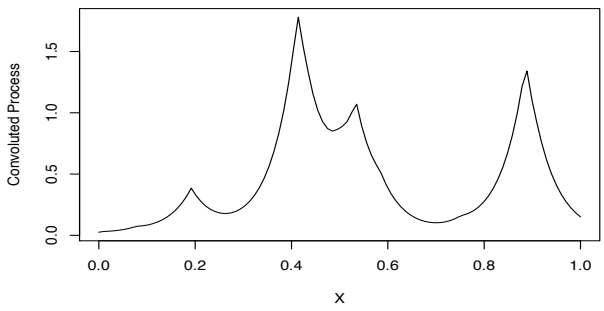
$$
\begin{aligned}
g(x, \omega_j) &= \exp\left(-\lambda_j |x - \chi_j|^{\rho_j}\right) \\
f(x) &= \sum_{j \le J} g(x; \omega_j)\beta_j \equiv \int_{\mathbf{\Omega}} g(x; \omega)\mathcal{L}(d\omega) \\
\mathcal{L}(d\omega) &= \sum_{j \le J} \beta_j \delta_{\omega_j}(d\omega)
\end{aligned}
$$

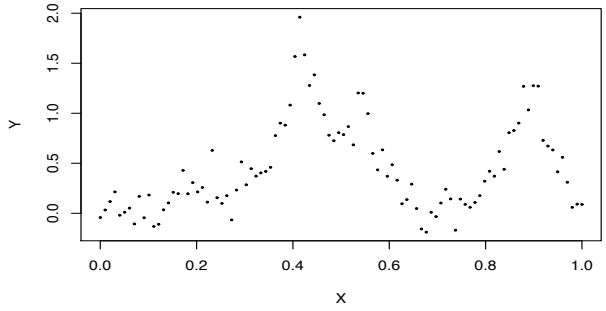
- support points of $\mathcal{L}$: $\{\omega_j\} = \{(\chi_j, \lambda_j, \rho_j)\}$
  - "location" of kernel: $\chi_j \in \mathcal{X}$ (unknown)
  - "scale" of kernel: $\lambda_j \in \mathbb{R}^+$ (unknown)
- jump sizes of measure: β_j (unknown)
- number of support points J (unknown)

Kernel Convolution of a Pure Jump Process









Why Overcomplete Dictionaries?

\+ More flexible - local adaptivity

\+ Potential for sparse representations

\- Non-unique coefficients

\- Computationally intensive search over (uncountable) dictionary

+/- If we are careful, no need to restrict to proper priors (!) (at least in theory)

Model Space with OCDs

"Space is big. Really big. You just won't believe how vastly hugely mind-bogglingly big it is." – D. Adams

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

Sufficient condition:

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

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

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

- Finite number of "big" coefficients $|\beta_j|$
- Possibly infinite number of $\beta \in [-\epsilon, \epsilon]$
- Jumps $|\beta_j|$ are absolutely summable[2]

[2] 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 & Selected ID Random Fields

- Gamma: $\nu(d\beta, d\boldsymbol{\omega}) = \beta^{-1} \exp(-\tau\beta) d\beta \; \gamma(d\boldsymbol{\omega})$

$$\beta_j \overset{iid}{\sim} \mathrm{Ga}(0, \tau)$$

- Symmetric Gamma:
  $\nu(d\beta, d\boldsymbol{\omega}) = |\beta|^{-1} \exp(-\tau|\beta|) d\beta \; \gamma(d\boldsymbol{\omega})$
- Cauchy: $\nu(d\beta, d\boldsymbol{\omega}) = |\beta|^{-2} d\beta \; \gamma(d\boldsymbol{\omega})$

$$\beta_j \mid \lambda_j \overset{ind}{\sim} \mathrm{N}(0, 1/\lambda_j) \qquad \lambda_j \overset{iid}{\sim} \mathrm{Ga}(1/2, 0)$$

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

$$\beta_j \mid \lambda_j \overset{ind}{\sim} \mathrm{N}(0, 1/\lambda_j) \qquad \lambda_j \overset{iid}{\sim} \mathrm{Ga}(\alpha/2, 0) \quad 0 < \alpha < 2$$

Provides a generalization of Generalized Ridge Priors to infinite dimensional case

Approximating Lévy Random Fields

In practice, cannot use infinite 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

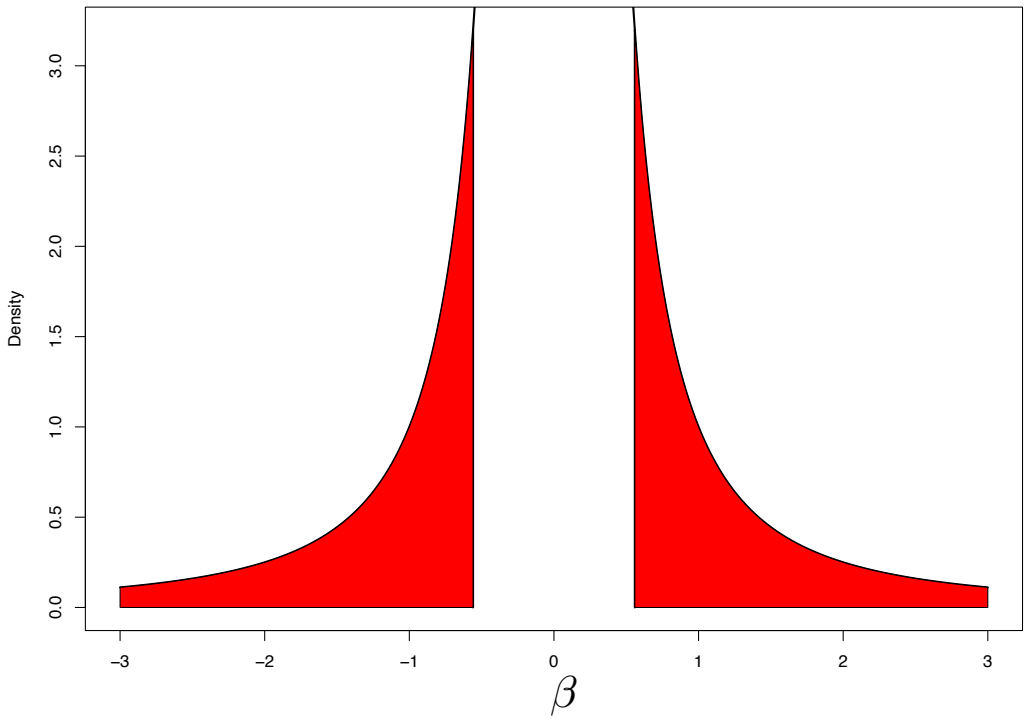
$$\nu_\epsilon(d\beta, d\boldsymbol{\omega}) \equiv \nu(d\beta, d\boldsymbol{\omega})\mathbf{1}(|\beta| > \epsilon)$$

$\Rightarrow$ $J \sim \text{Po}(\nu_\epsilon^+)$; $\nu_\epsilon^+ = \nu([-\epsilon, \epsilon]^c, \boldsymbol{\Omega})$

- $\beta_j, \boldsymbol{\omega}_j \overset{iid}{\sim} \pi(d\beta, d\boldsymbol{\omega}) \equiv \nu_\epsilon(d\beta, d\boldsymbol{\omega})/\nu_\epsilon^+$
- use RJ-MCMC to update $J, \{\beta_j, \boldsymbol{\omega}_j\}$

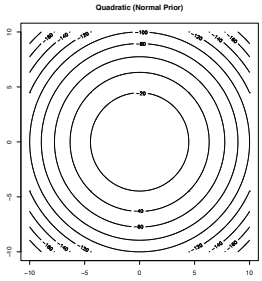
Truncated Cauchy

Restriction $|\beta| > \epsilon$

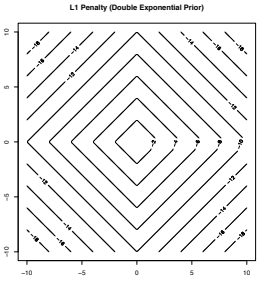


Contours of Log Prior (in $\mathbb{R}^2$) – Penalties

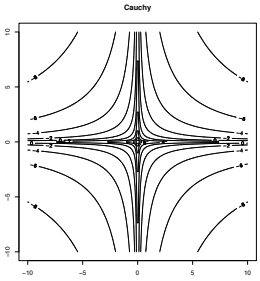
Normal



DE



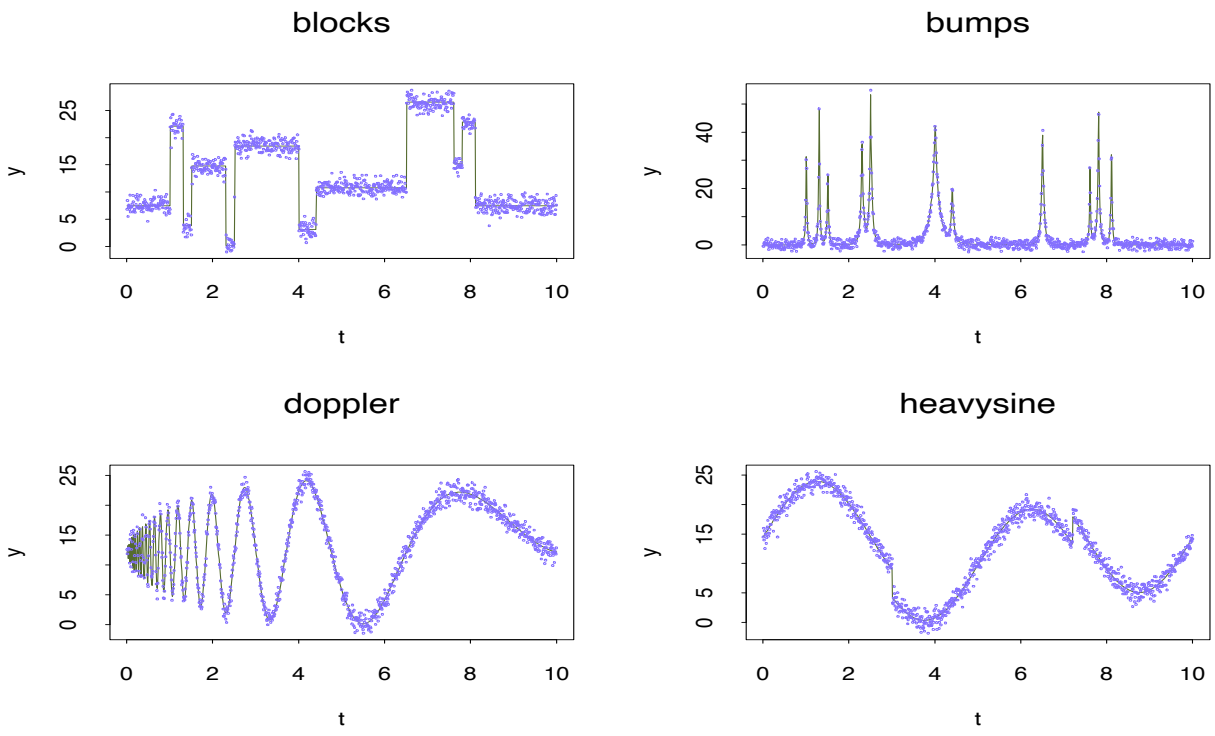
Cauchy



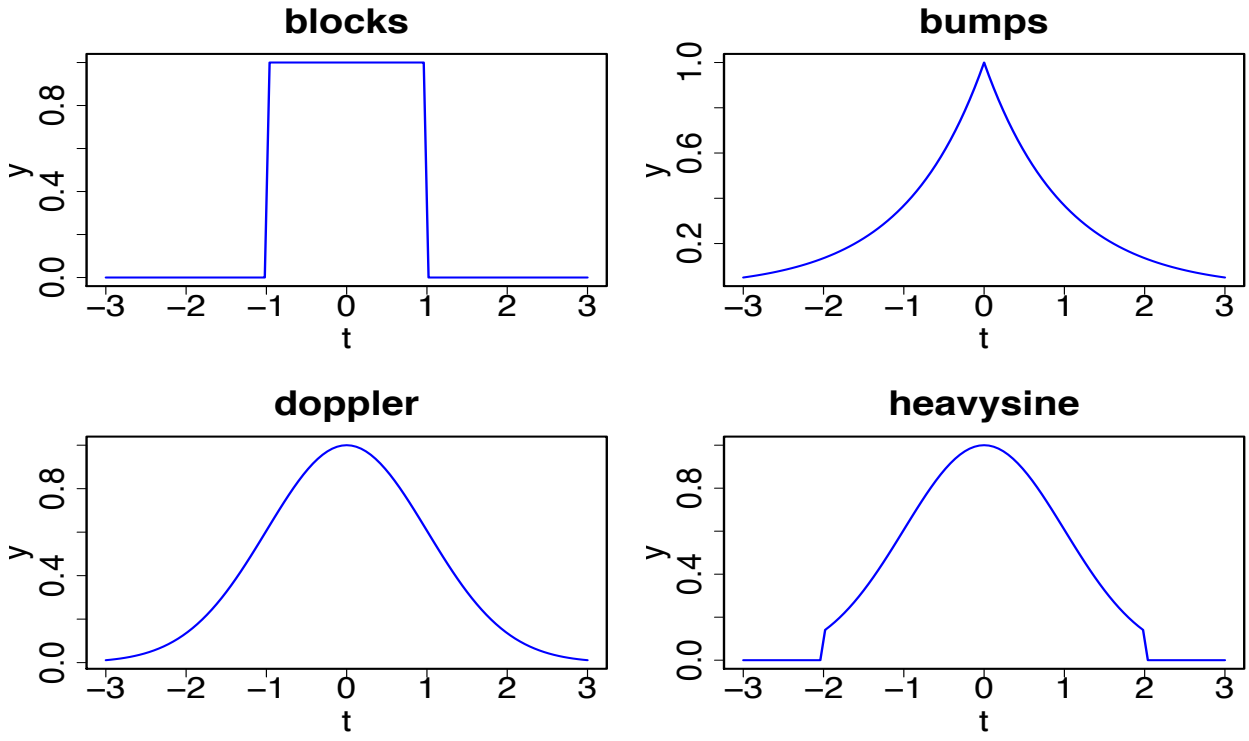
Penalized Likelihood:

$$-\frac{1}{2\sigma^2}\sum_i \left(Y_i - f(\mathbf{x}_i)\right)^2 - \log(J!) - (\alpha+1)\sum_j \log(|\beta_j|) - \nu_\epsilon^+ \ldots$$

Wavelet Test Functions (SNR = 7)



Kernel Functions



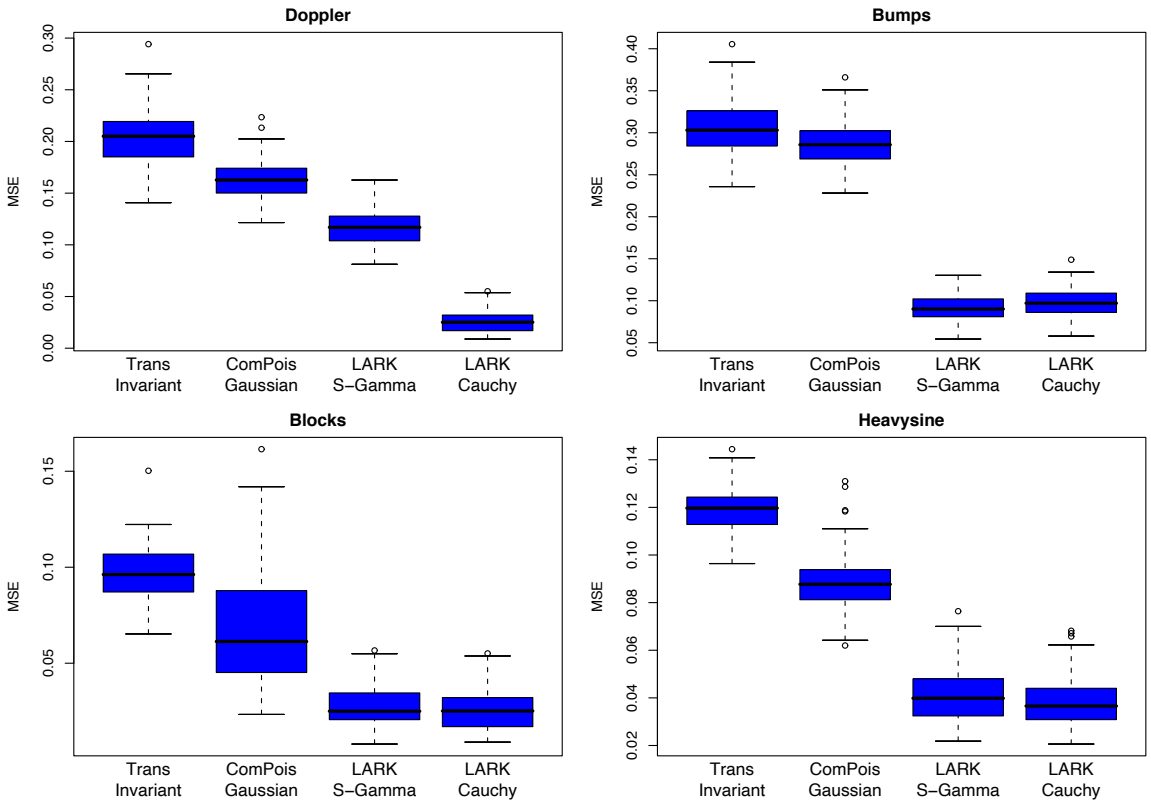
Comparisons of OCD Methods

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

Range of Overcomplete Dictionaries and Priors

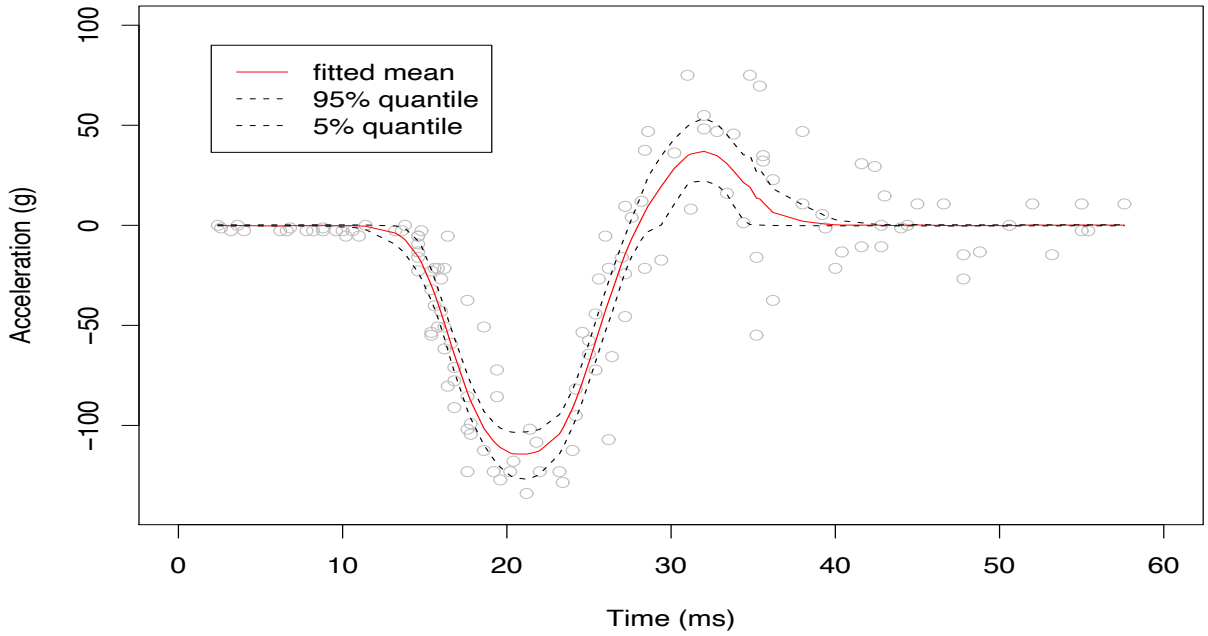
Comparison of Mean Square Error w/ OCDs

100 realizations of each function



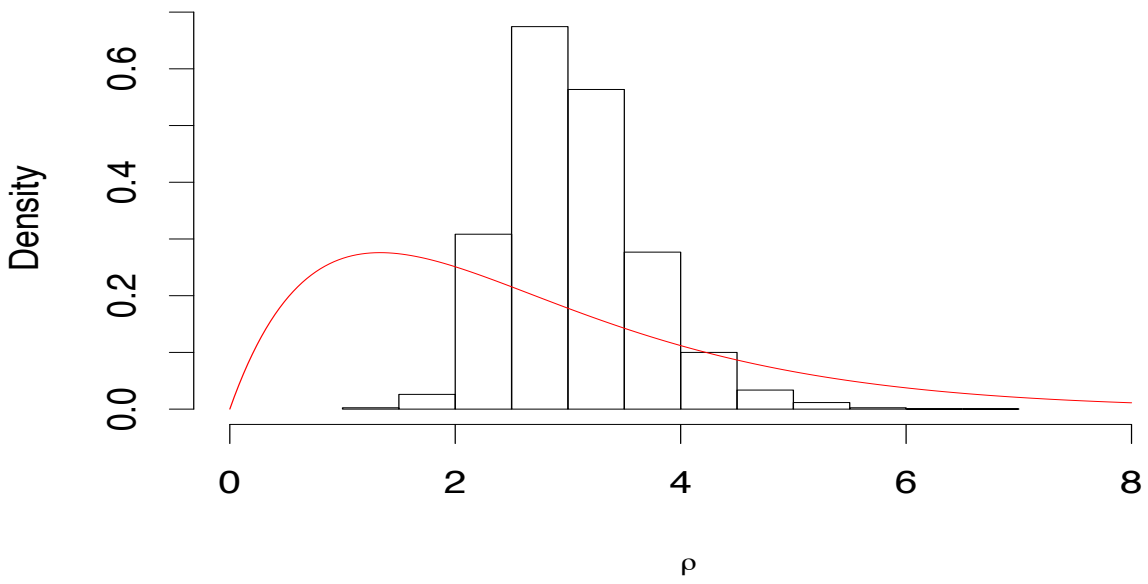
Motorcycle Crash Data

On average, only $E[J \mid Y] \approx 4$ jumps are needed for fit:

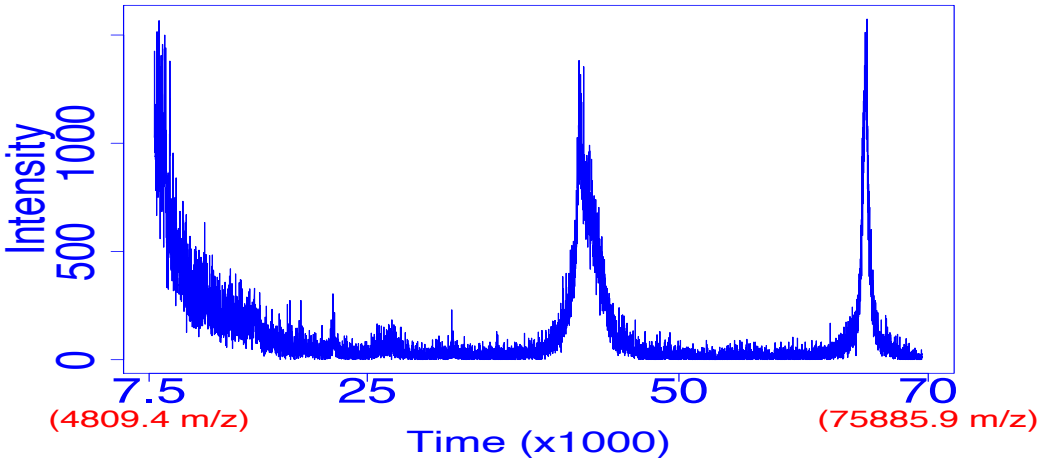


Posterior on Kernel Power

$$g(x_i;\omega) = e^{-\lambda_j |x_i - \chi_j|^\rho}$$



MALDI-TOF Mass Spectroscopy



Proteins correspond to peaks in the spectrogram

Single Spectrum Likelihood

Model for Measured Intensity at TOF t

- Expected intensity: $\mathrm{E}[Y(t)] = b + f(t)$
- b = Baseline Shift (constant, unimportant nuisance)
- EDA suggests variance proportional to mean:

$$Y(t) - b \overset{ind}{\sim} \mathrm{Ga}\left(f(t) \cdot \phi,\ \phi\right)$$

Propose: non-parametric model with Gamma random field prior $\mathcal{L} \equiv \Gamma$

$$f(t) = \sum_{j}^{J} k(t, \boldsymbol{\theta}_j)\, \beta_j = \int_{\Theta} k(t, \boldsymbol{\theta}) \Gamma(d\boldsymbol{\theta})$$

Natural Basis Functions

Peaks for a single spectrum often exhibit Gaussian form in the time domain where the "spread" is primarily induced from initial velocity distribution.

Use kernels with parameters $\theta = (\tau, \lambda) \in \Theta \subset \mathbb{R}^2_+$:

$$k(t, \boldsymbol{\theta}) = \frac{\lambda^{1/2}}{\sqrt{2\pi}} \exp\{-\frac{\lambda}{2}(t-\tau)^2\} \qquad (1)$$

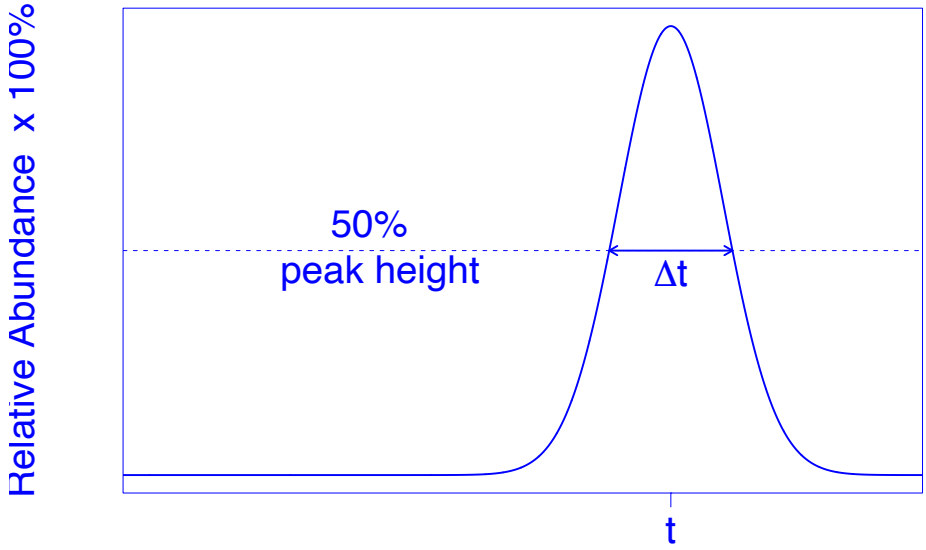
- $\{\tau\}$ expected TOF of protein (unknown)
- $\{\lambda\}$ controls peak spread (unknown)

Both τ and λ may vary from peak to peak

Induced Prior on f

- $f(t) = \sum_j^J k(t, \boldsymbol{\theta}_j)\beta_j$
- J is the number of kernels (unknown but finite) corresponding to the unknown number of proteins (Poisson *a priori*)
- $\{\beta_j\}$ concentration (ϵ-truncated Gamma process)
- $\boldsymbol{\theta} = (\tau, \lambda)$
  - $\{\tau_j\}$ Expected TOF of protein (unknown) (Uniform)
  - $\{\lambda_j\}$ peak width – determined by prior on resolution

Mass Resolution



Resolution $\varrho \equiv \Delta t/\tau$

The longer the TOF (larger mass), the wider the peak.

Hierarchical Prior on Resolution

- Individual Peak Resolution

$$\rho_j \overset{iid}{\sim} \text{LN}(\varrho, \sigma_\rho)$$

- Change of variables between

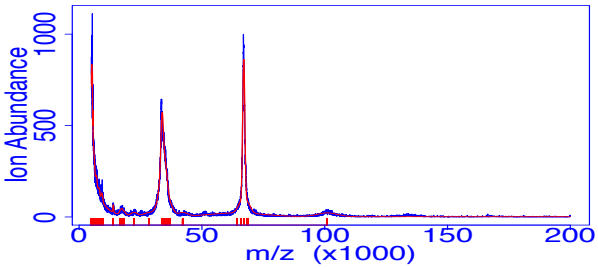
$$\lambda_j = g(\rho_j, \tau_j)$$

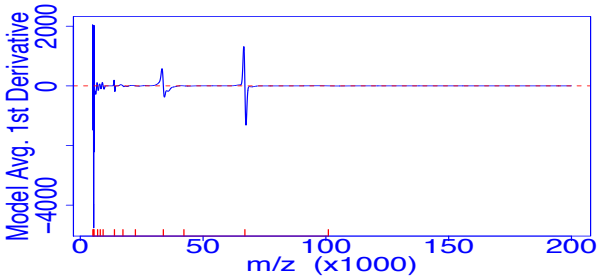
- Prior on overall machine resolution: selected so that 95% interval covers (20, 100) based on prior info

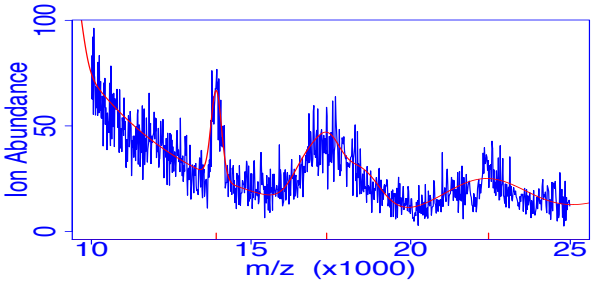
Posterior Inference

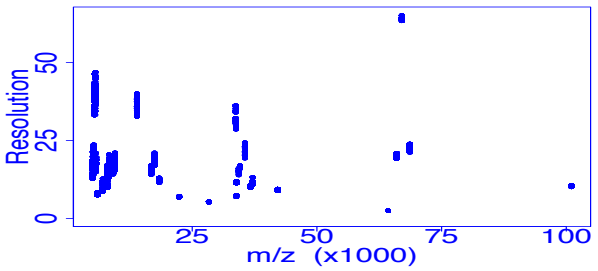
- No closed form for posterior distribution
- Use stochastic simulation via Reversible Jump Markov chain Monte Carlo (RJ-MCMC) to make inference about
  - unknown number of proteins (M)
  - TOF $\{\tau_j\} \rightarrow$ mass/charge
  - concentration $\{\beta_j\}$
  - peak resolution $\{\rho_j\}$
- Find configuration with highest posterior mode, or
- "Model Averaging"— posterior mean $\mathrm{E}[f(\cdot)]$

Estimated Spectrum









Features

Nonparametric model addresses several issues in MALDI-TOF data:

- Expected intensity is non-negative
- Peak identification ($\{\tau_j\}$)
- Measure of relative protein quantity (area) ($\{\beta_j\}$)
- Scaling
- Baseline shift and exponential decay (initial peaks)
- Non-Gaussian Errors
- Use of expert information on resolution in prior on λ_j
- Hierarchical Model: multiple spectra

Multiple Spectra

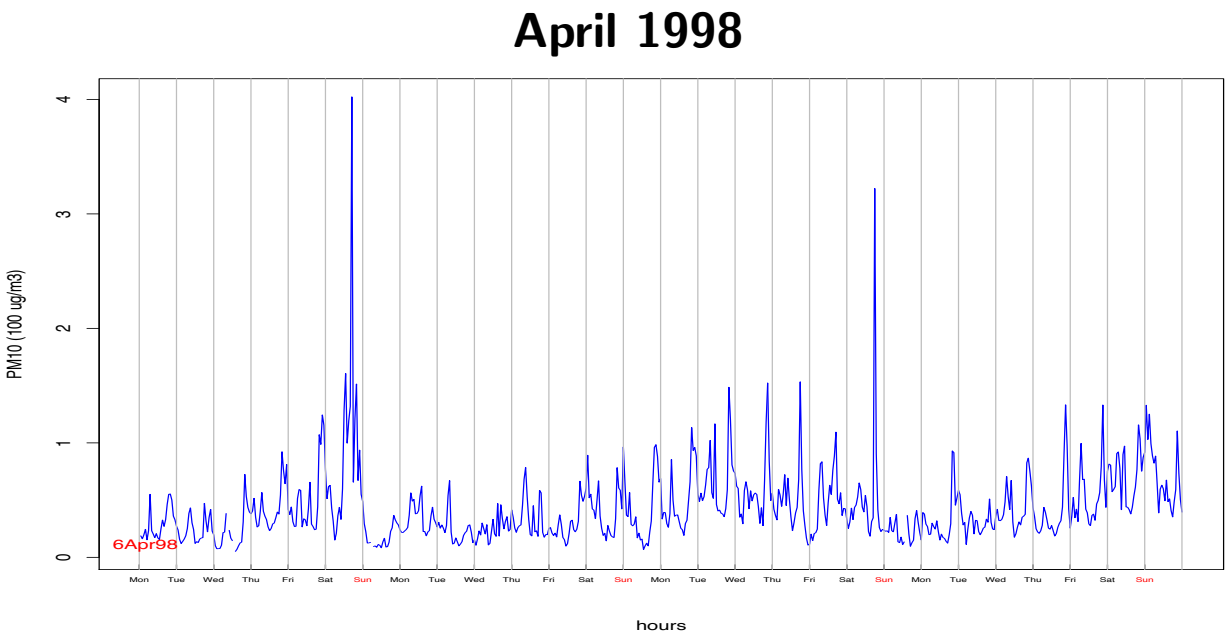
Simultaneous

- Alignment of spectra
- Identification of peaks (proteins)
- Differential Expression of Proteins Across Groups

Add "Mark" to ω: { Control, Shared, Disease }

Goal: Classification of subjects from biologically relevant features (peaks/proteins not valleys)

Hourly PM_{10} concentration in Maricopa County, AZ



The "Spikey" concentration profiles don't fit ARMA well. Semi-periodic with possible daily and meteorlogically driven patterns.

Marked Lévy model

$$
\begin{aligned}
Y_{t_i} &= f(t_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \\
f(t) &= b_0 + \int_{\Omega} k(t;\omega)\, \Gamma(d\omega) \quad \swarrow \text{ Marks} \\
\Omega &= [0, 720] \times \mathbb{R}_+ \times \{0, 1\} \\
k(t;\omega) &= k\big(t; (\tau, \lambda, a)\big) \\
&= \begin{cases} e^{-\lambda |t-\tau|} & a = 0, \quad \textit{Aperiodic} \text{ part} \\ e^{-\lambda |(t-\tau) \pmod{24}|} & a = 1, \quad \textit{Daily} \text{ part} \end{cases}
\end{aligned}
$$

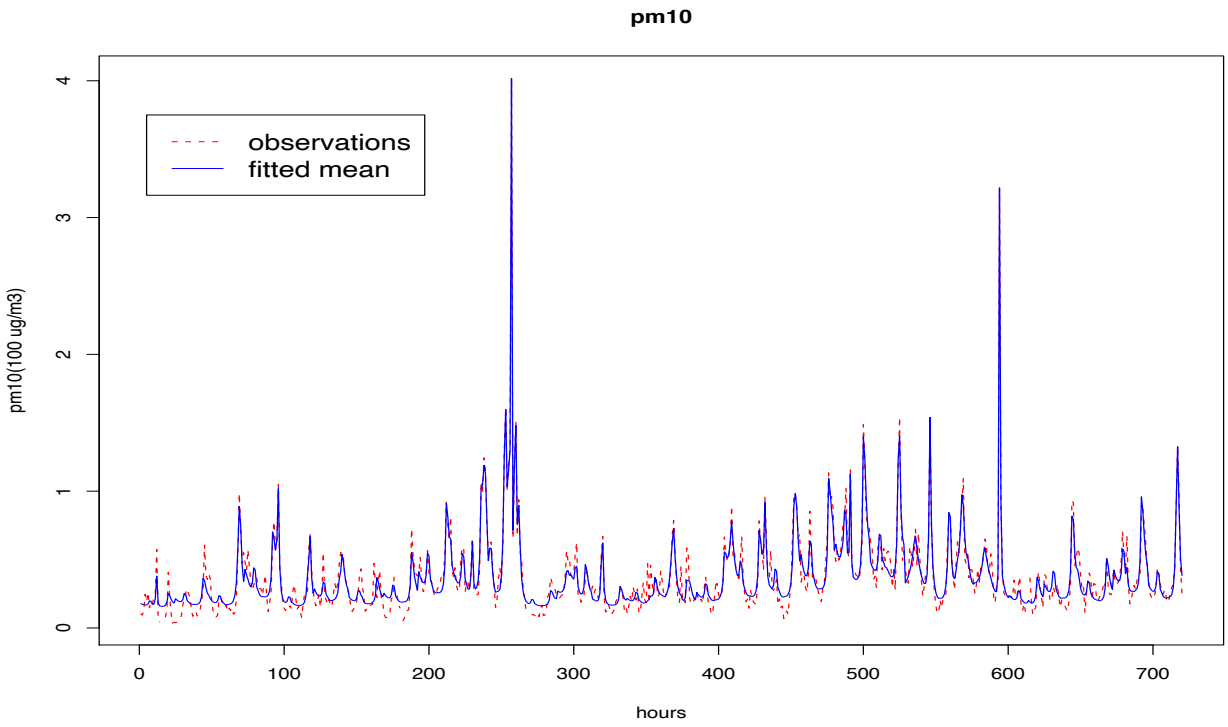
The Prior Distribution:

Aperiodic part			Daily periodic part		
J_0	$\sim$	$\mathrm{Po}\big(\alpha_0\, T\, E_1(\beta_0\epsilon)\big)$	J_1	$\sim$	$\mathrm{Po}\big(\alpha_1\, 24\, E_1(\beta_1\epsilon)\big)$
$\{\tau_j\}$	$\overset{iid}{\sim}$	$\mathrm{Un}[0, T]$	$\{\tau_j\}$	$\overset{iid}{\sim}$	$\mathrm{Un}[0, 24]$
$\{\lambda_j\}$	$\overset{iid}{\sim}$	$\mathrm{Ga}(a_\lambda, b_\lambda)$	$\{\lambda_j\}$	$\overset{iid}{\sim}$	$\mathrm{Ga}(a_\lambda, b_\lambda)$
$\{u_j\}$	$\propto$	$u^{-1}e^{-\beta_0 u}\,\mathbf{1}_{\{u>\epsilon\}}$	$\{u_j\}$	$\propto$	$u^{-1}e^{-\beta_1 u}\,\mathbf{1}_{\{u>\epsilon\}}$

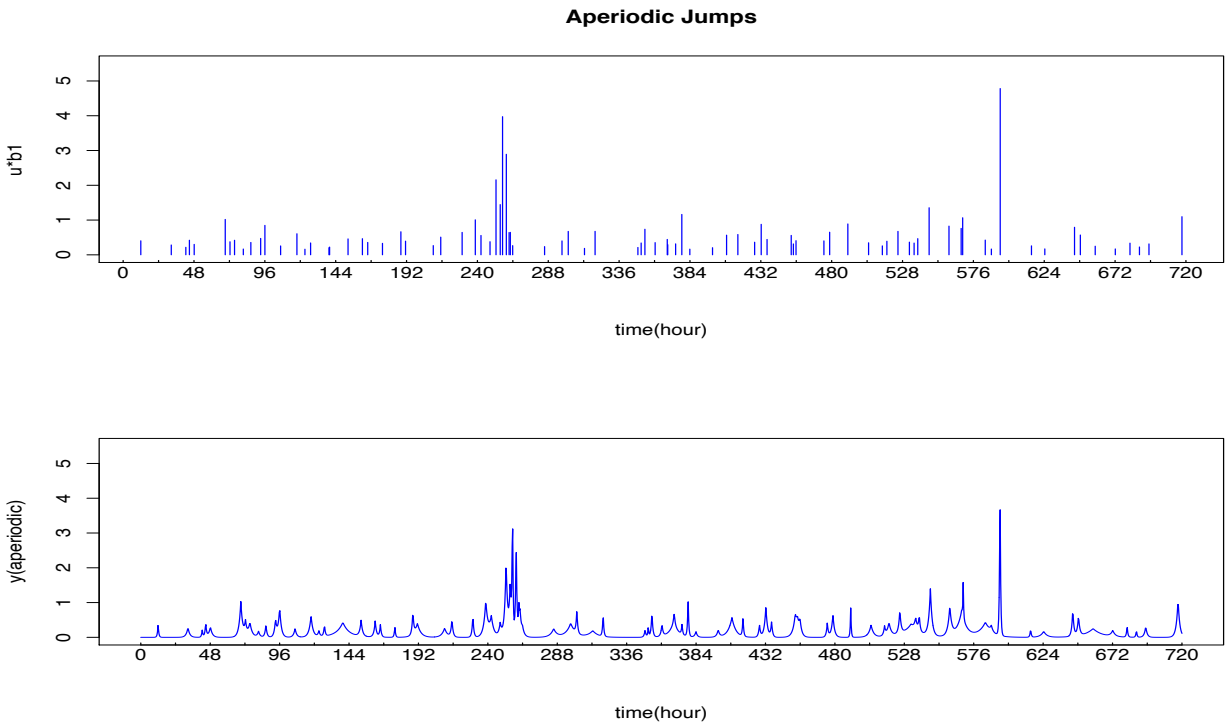
$$f(t) = b_0 + \sum_{j=1}^{J_0} u_j e^{-\lambda_j|t-\tau_j|} + \sum_{j=J_0+1}^{J_0+J_1} u_j e^{-\lambda_j|(t-\tau_j) \pmod{24}|}$$

Note: May include daily meteorological data $\mathbf{x_k}$ (temp, wind, etc.) through coefficient $b_{\lceil t/24\rceil} \sim \mathrm{LN}(\mathbf{x_k}'\gamma_m, \sigma^2_m)$ in daily term.

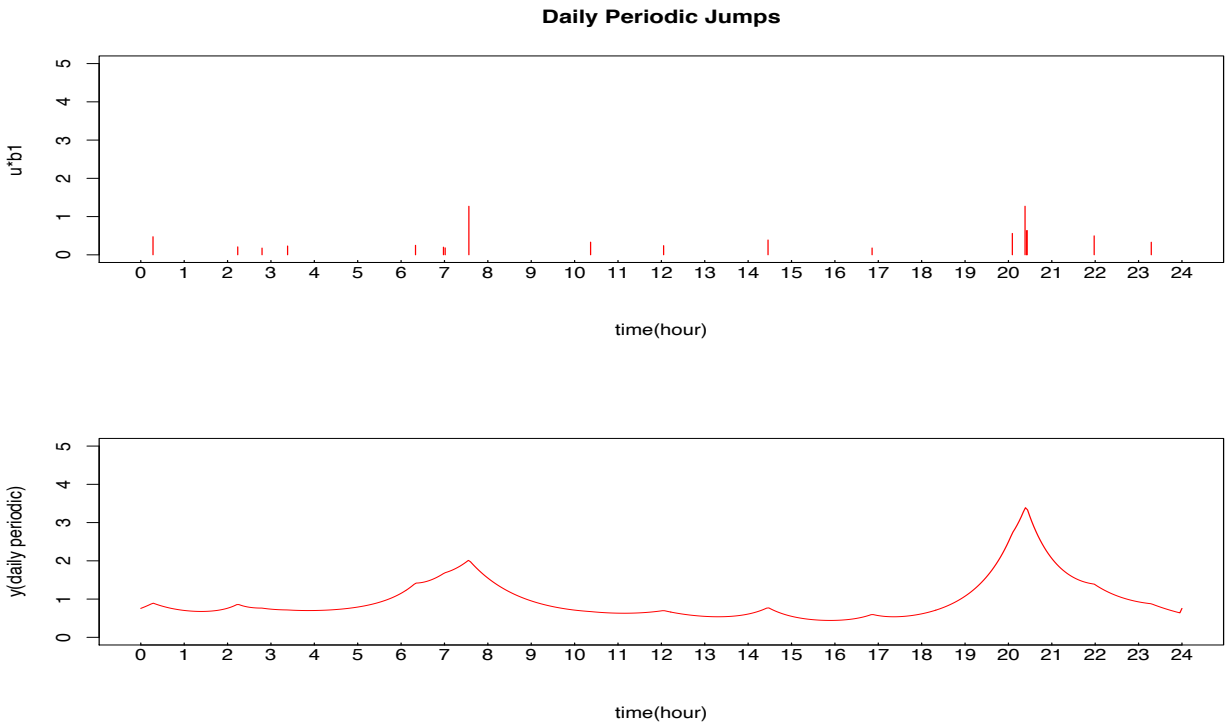
The Fitted Model



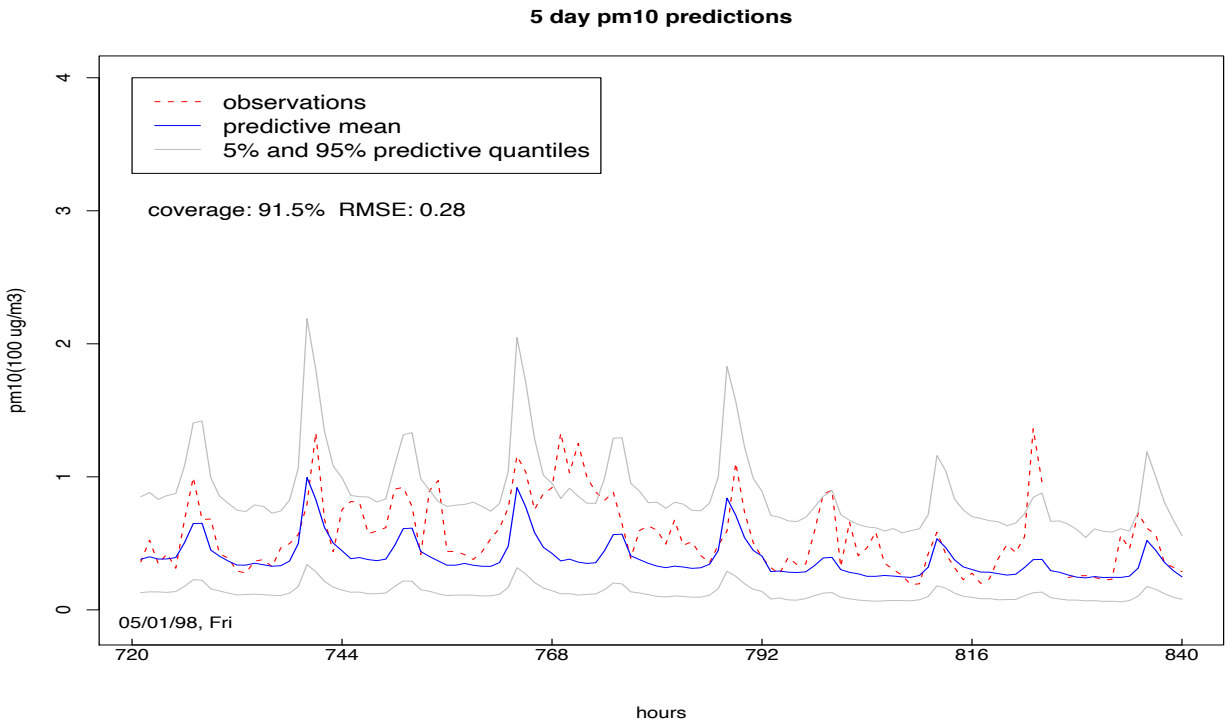
The Fitted Model: Aperiodic Part



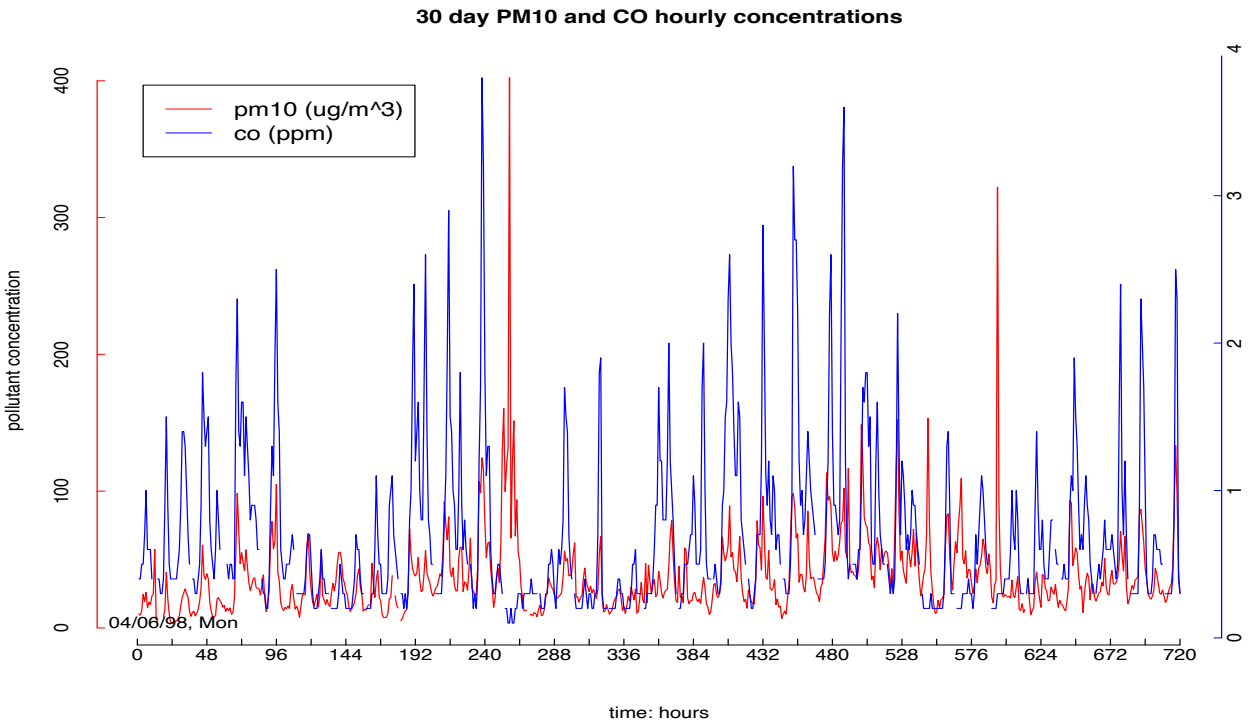
The Fitted Model: Daily Part



The Fitted Model: Predictions



Two Pollutants: PM_{10} and CO



Marks

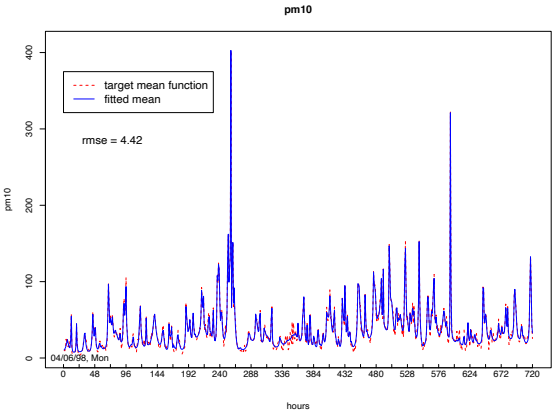
six marks

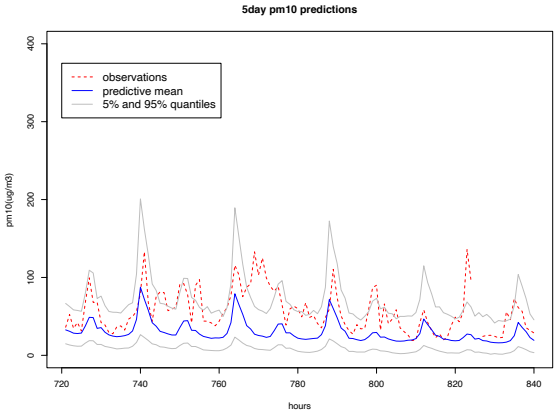
$$\boldsymbol{\Omega} = [0, 720] \times \mathbb{R}_+ \times \mathcal{A}, \qquad \mathcal{A} \equiv \{0, 1, 2, 3, 4, 5\}$$

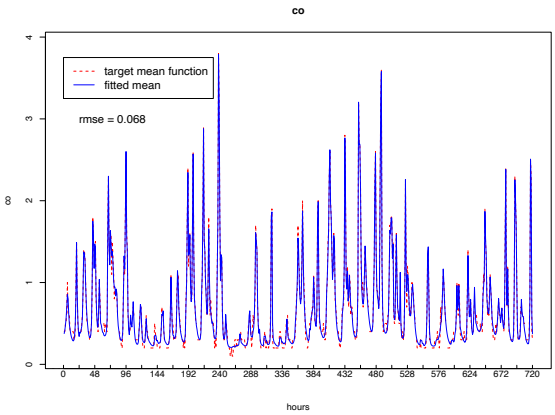
where

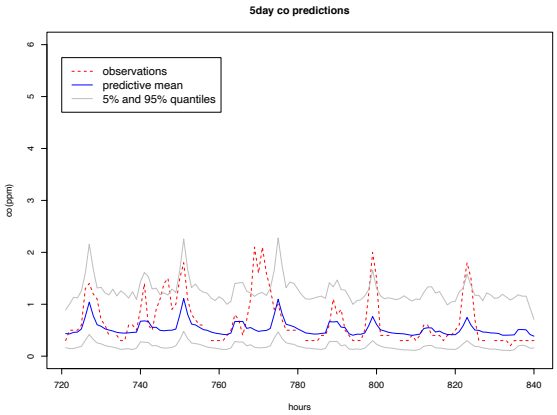
$$a = \begin{cases} 0 & \text{Aperiodic, } \mathrm{PM}_{10} \text{ (only)} \\ 1 & \text{Daily, } \mathrm{PM}_{10} \text{ (only)} \\ 2 & \text{Aperiodic, CO (only)} \\ 3 & \text{Daily, CO (only)} \\ 4 & \text{Aperiodic, } \mathrm{PM}_{10} \textit{ and } \mathrm{CO} \\ 5 & \text{Daily, } \mathrm{PM}_{10} \textit{ and } \mathrm{CO} \end{cases}$$

Fits and Predictions: PM_{10} and CO

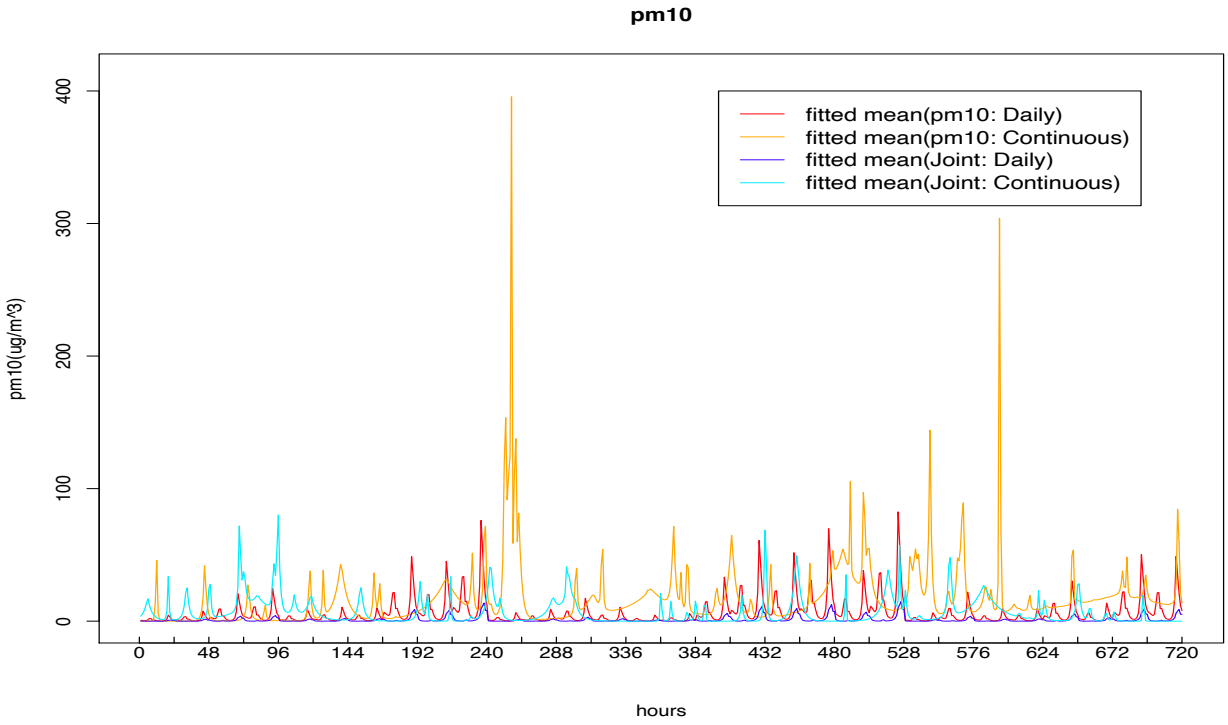




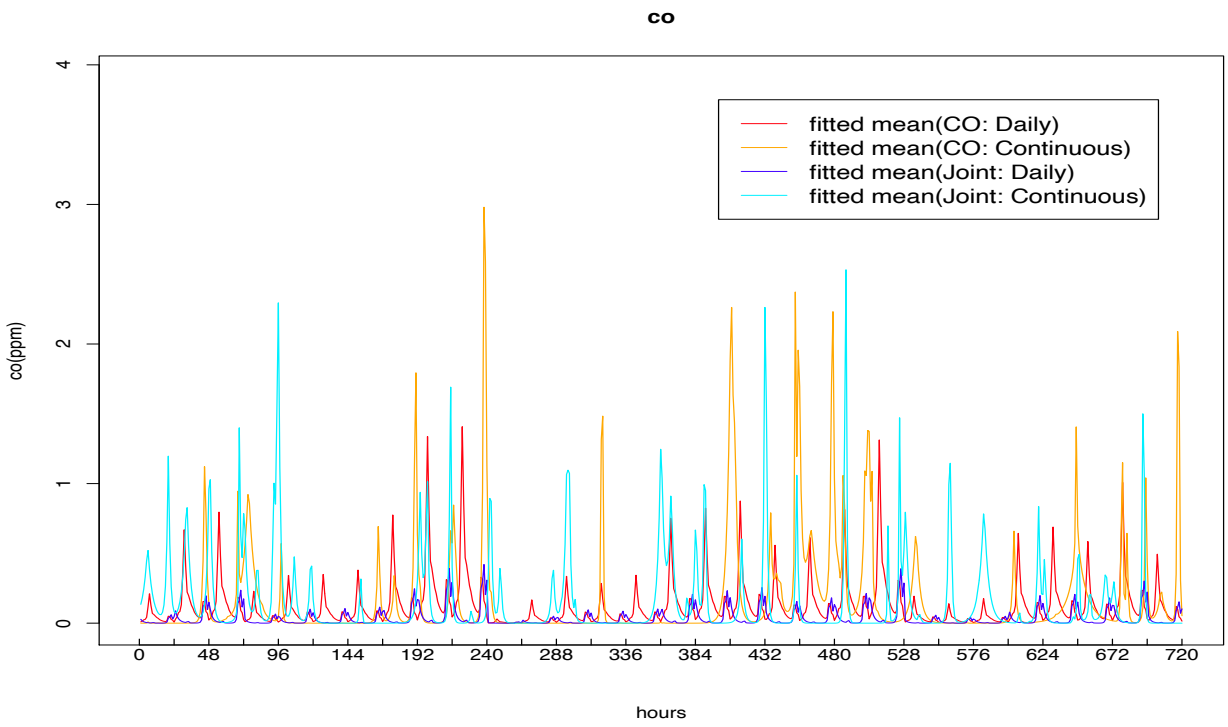




Decomposition, PM_{10}



Decomposition, CO



Benefits of the Method

- Non-negative data (here, $[\mathrm{PM}_{10}]$ and $[\mathrm{CO}]$) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- No problems with unequally spaced data
- No need to invert large matrices (as in Gaussian methods)
- Non-linear dependence structure okay
- Easy interpretability, good out-of-sample predictions, easy dove-tail with other models (e.g. trajectory analysis)
- Our Mov Avg method permits only positive correlations

Benefits of the Method

- Non-negative data (here, $[\mathrm{PM}_{10}]$ and $[\mathrm{CO}]$) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- No problems with unequally spaced data
- No need to invert large matrices (as in Gaussian methods)
- Non-linear dependence structure okay
- Easy interpretability, good out-of-sample predictions, easy dove-tail with other models (e.g. trajectory analysis)
- Our Mov Avg method permits only positive correlations

Benefits of the Method

- Non-negative data (here, $[\mathrm{PM}_{10}]$ and $[\mathrm{CO}]$) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- No problems with unequally spaced data
- No need to invert large matrices (as in Gaussian methods)
- Non-linear dependence structure okay
- Easy interpretability, good out-of-sample predictions, easy dove-tail with other models (e.g. trajectory analysis)
- Our Mov Avg method permits only positive correlations

Benefits of the Method

- Non-negative data (here, $[\mathrm{PM}_{10}]$ and $[\mathrm{CO}]$) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- No problems with unequally spaced data
- No need to invert large matrices (as in Gaussian methods)
- Non-linear dependence structure okay
- Easy interpretability, good out-of-sample predictions, easy dove-tail with other models (e.g. trajectory analysis)
- Our Mov Avg method permits only positive correlations

Spatial-Temporal Models

Now observe $Z_i = Z(s_i, t_i)$ at space $\{s_i\} \in \mathcal{S}$ and time $\{t_i\} \in \mathcal{T}$, with separable-kernel model. For each i,

$$
\begin{aligned}
Z_i &= f(s_i, t_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{No}(0, \sigma^2) \\
f(s,t) &= b_0 + \int_{\Omega} k(s,t;\omega)\, \Gamma(d\omega) \\
\Omega &= \mathcal{S} \times \mathcal{T} \times \mathbb{R}_+ \times \mathcal{M}_+ \times \mathcal{A} \\
k(s,t;\omega) &= k\big(s,t;(\sigma,\tau,\lambda,\Lambda,a)\big) \\
&= \begin{cases} e^{-\lambda|t-\tau| - (s-\sigma)'\Lambda(s-\sigma)/2} & a = 0 \text{ (Aper)} \\ e^{-\lambda|(t-\tau) \pmod{24}| - (s-\sigma)'\Lambda(s-\sigma)/2} & a = 1 \text{ (Daily)} \end{cases} \\
(\sigma,\tau) &\in \mathcal{S} \times \mathcal{T} \text{ is space-time center (random, adaptive)} \\
\lambda &\in \mathbb{R}_+ \text{ is Temporal Decay rate (random, adaptive)} \\
\Lambda &\in \mathcal{M}_+ \text{ is Spat dispers'n matrix (random, adaptive)}
\end{aligned}
$$

Spatial-Temporal Models

Now observe $Z_i = Z(s_i, t_i)$ at space $\{s_i\} \in \mathcal{S}$ and time $\{t_i\} \in \mathcal{T}$, with separable-kernel model. For each i,

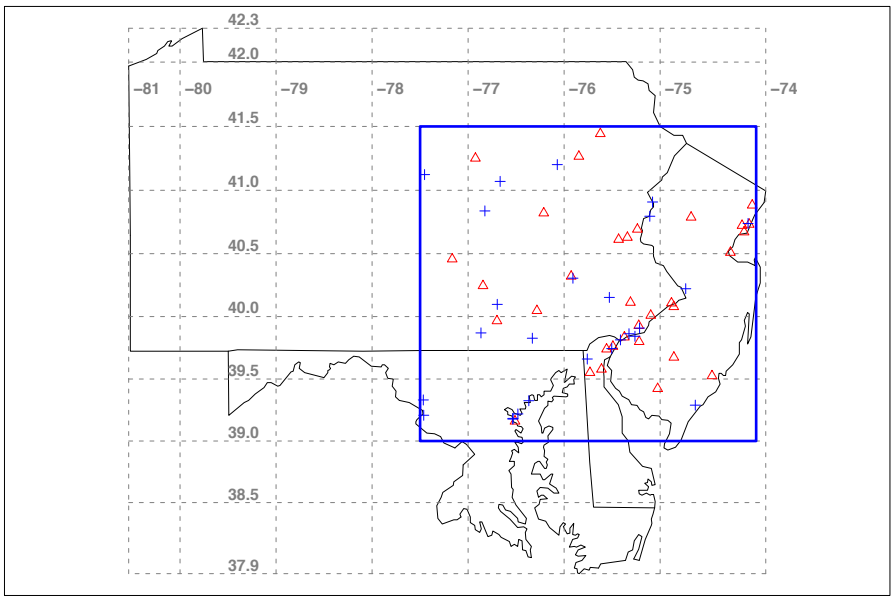
$$
\begin{aligned}
Z_i &= f(s_i, t_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{No}(0, \sigma^2) \\
f(s,t) &= b_0 + \int_{\boldsymbol{\Omega}} k(s,t;\omega)\, \Gamma(d\omega) \\
\boldsymbol{\Omega} &= \mathcal{S} \times \mathcal{T} \times \mathbb{R}_+ \times \mathcal{M}_+ \times \mathcal{A} \\
k(s,t;\omega) &= k\big(s,t;(\sigma,\tau,\lambda,\Lambda,a)\big) \\
&= \begin{cases} e^{-\lambda|t-\tau| - (s-\sigma)'\Lambda(s-\sigma)/2} & a = 0 \text{ (Aper)} \\ e^{-\lambda|(t-\tau) \pmod{24}| - (s-\sigma)'\Lambda(s-\sigma)/2} & a = 1 \text{ (Daily)} \end{cases} \\
(\sigma,\tau) &\in \mathcal{S} \times \mathcal{T} \text{ is space-time center (random, adaptive)} \\
\lambda &\in \mathbb{R}_+ \text{ is Temporal Decay rate (random, adaptive)} \\
\Lambda &\in \mathcal{M}_+ \text{ is Spat dispers'n matrix (random, adaptive)}
\end{aligned}
$$

Spatial-Temporal Models

Now observe $Z_i = Z(s_i, t_i)$ at space $\{s_i\} \in \mathcal{S}$ and time $\{t_i\} \in \mathcal{T}$, with separable-kernel model. For each i,

$$
\begin{aligned}
Z_i &= f(s_i, t_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \text{No}(0, \sigma^2) \\
f(s,t) &= b_0 + \int_{\boldsymbol{\Omega}} k(s,t;\omega)\, \Gamma(d\omega) \\
\boldsymbol{\Omega} &= \mathcal{S} \times \mathcal{T} \times \mathbb{R}_+ \times \mathcal{M}_+ \times \mathcal{A} \\
k(s,t;\omega) &= k\big(s,t;(\sigma,\tau,\lambda,\Lambda,a)\big) \\
&= \begin{cases} e^{-\lambda|t-\tau| - (s-\sigma)'\Lambda(s-\sigma)/2} & a = 0 \text{ (Aper)} \\ e^{-\lambda|(t-\tau) \pmod{24}| - (s-\sigma)'\Lambda(s-\sigma)/2} & a = 1 \text{ (Daily)} \end{cases} \\
(\sigma,\tau) &\in \mathcal{S} \times \mathcal{T} \text{ is space-time center (random, adaptive)} \\
\lambda &\in \mathbb{R}_+ \text{ is Temporal Decay rate (random, adaptive)} \\
\Lambda &\in \mathcal{M}_+ \text{ is Spat dispers'n matrix (random, adaptive)}
\end{aligned}
$$

Sulfur Dioxide Concentrations in PA, MD, NJ:



EPA SO_2 Monitoring stations (real ones!) are shown as Δ's, Point Sources (power plants, *etc.*) as $+$'s

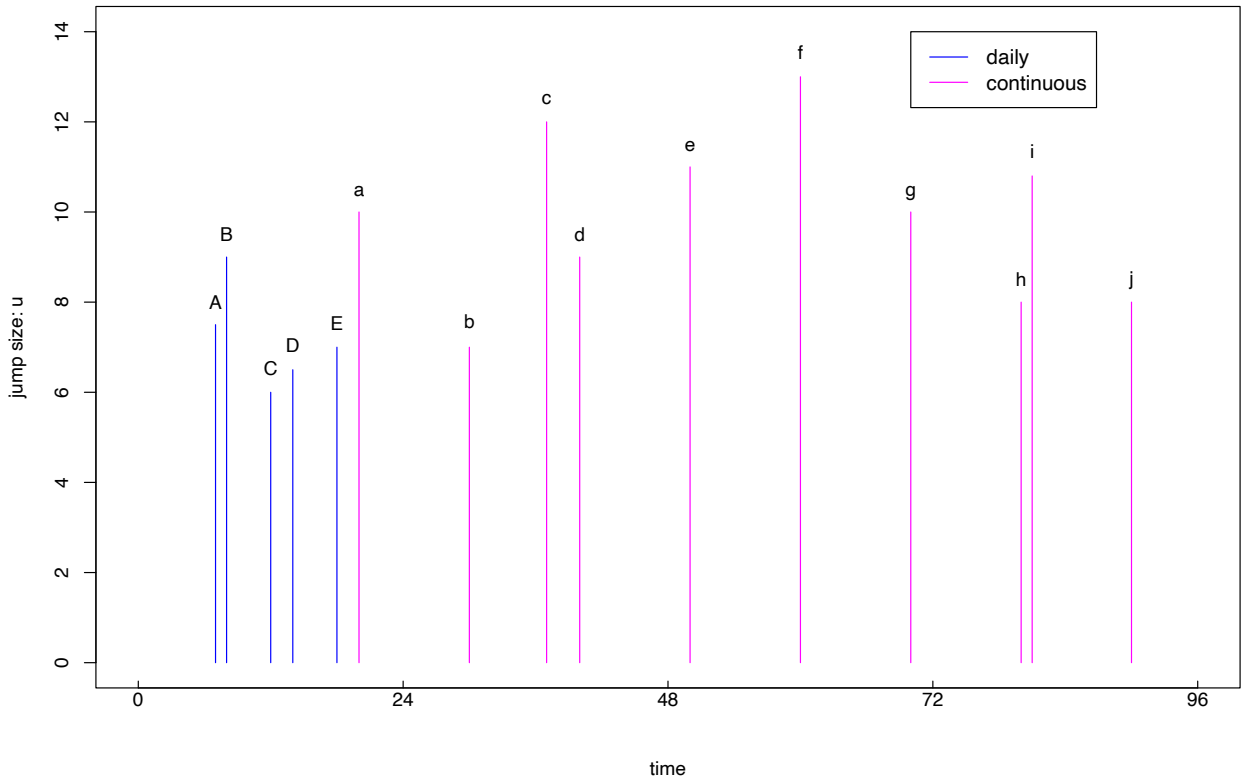
Simulated Data

Simulated data in a Table

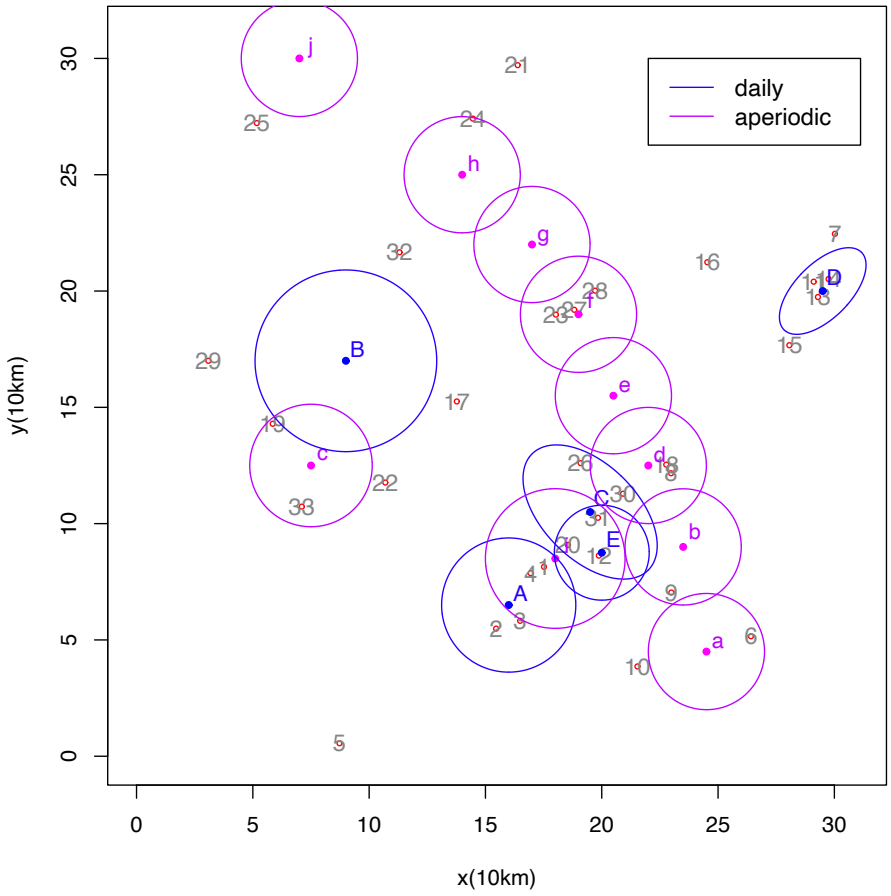
$J_0 = 10,\ J_1 = 5,\ \{(u_j, \omega_j)\}_{j \le 15} =$

	u	$(\sigma_x,$	$\sigma_y)$	τ	λ_t	$(\lambda_x,$	$\lambda_y)$	ω	a
a	10.0	(24.5,	4.50)	20	0.50	(4.5,	4.5)	1.571	0
b	7.0	(23.5,	9.00)	30	0.50	(4.5,	4.5)	1.571	0
c	12.0	(7.5,	12.50)	37	0.55	(5.0,	5.0)	1.571	0
d	9.0	(22.0,	12.50)	40	0.50	(4.5,	4.5)	1.571	0
e	11.0	(20.5,	15.50)	50	0.50	(4.5,	4.5)	1.571	0
f	13.0	(19.0,	19.00)	60	0.50	(4.5,	4.5)	1.571	0
g	10.0	(17.0,	22.00)	70	0.50	(4.5,	4.5)	1.571	0
h	8.0	(14.0,	25.00)	80	0.50	(4.5,	4.5)	1.571	0
i	10.8	(18.0,	8.50)	81	0.60	(6.5,	6.5)	1.571	0
j	8.0	(7.0,	30.00)	90	0.50	(4.5,	4.5)	1.571	0
A	7.5	(16.0,	6.50)	7*	0.45	(6.0,	6.0)	1.571	1
B	9.0	(9.0,	17.00)	8*	0.35	(11.0,	11.0)	1.571	1
C	6.0	(19.5,	10.50)	12*	0.60	(9.0,	3.0)	0.785	1
D	6.5	(29.5,	20.00)	14*	0.40	(4.0,	1.0)	-0.785	1
E	7.0	(20.0,	8.75)	18*	0.75	(3.0,	3.0)	1.571	1

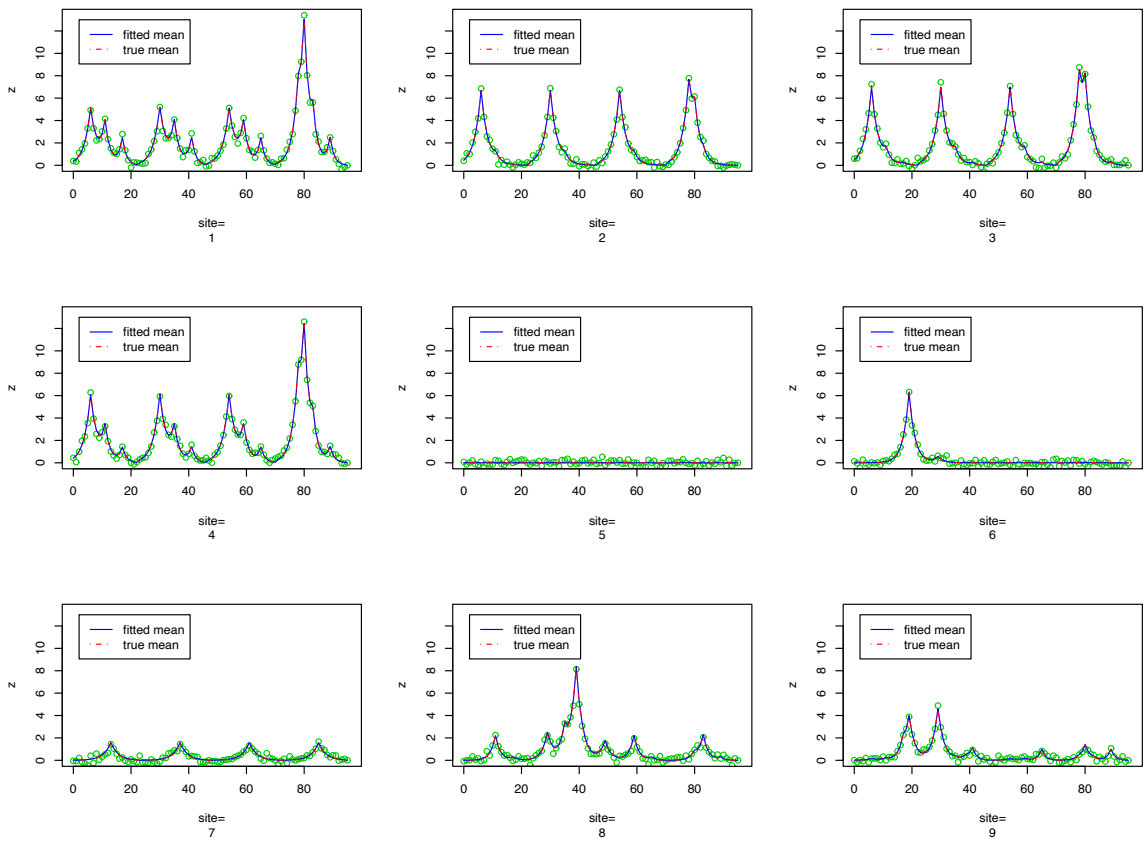
Simulated data in a Plot:



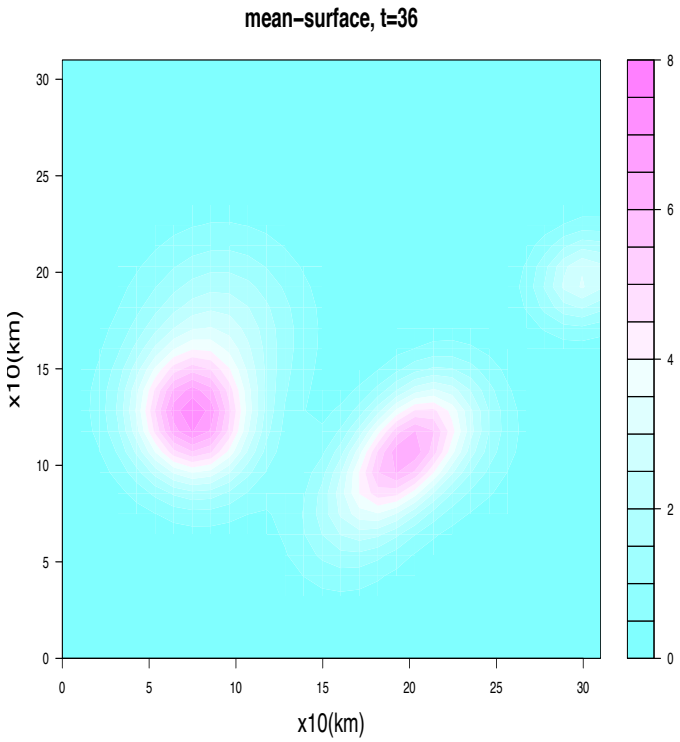
Simulated data in a Picture:

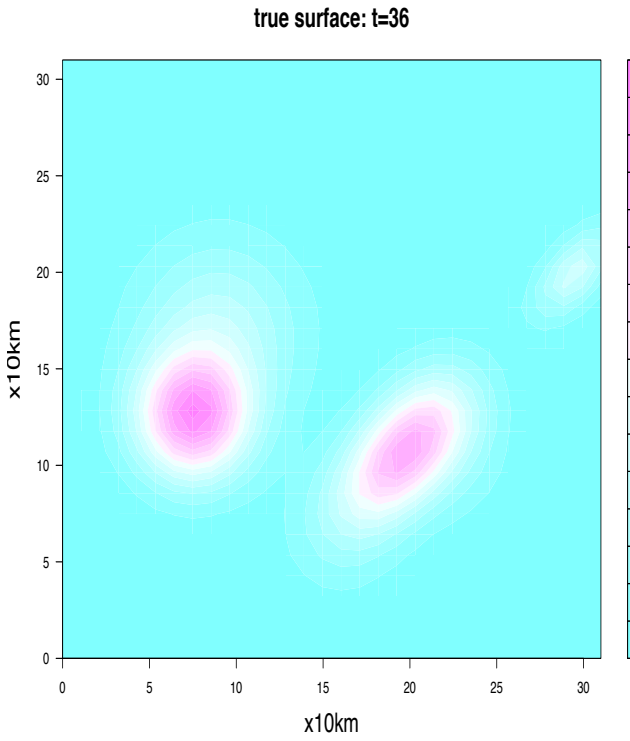


Model Fit at Nine Locations, All Times:

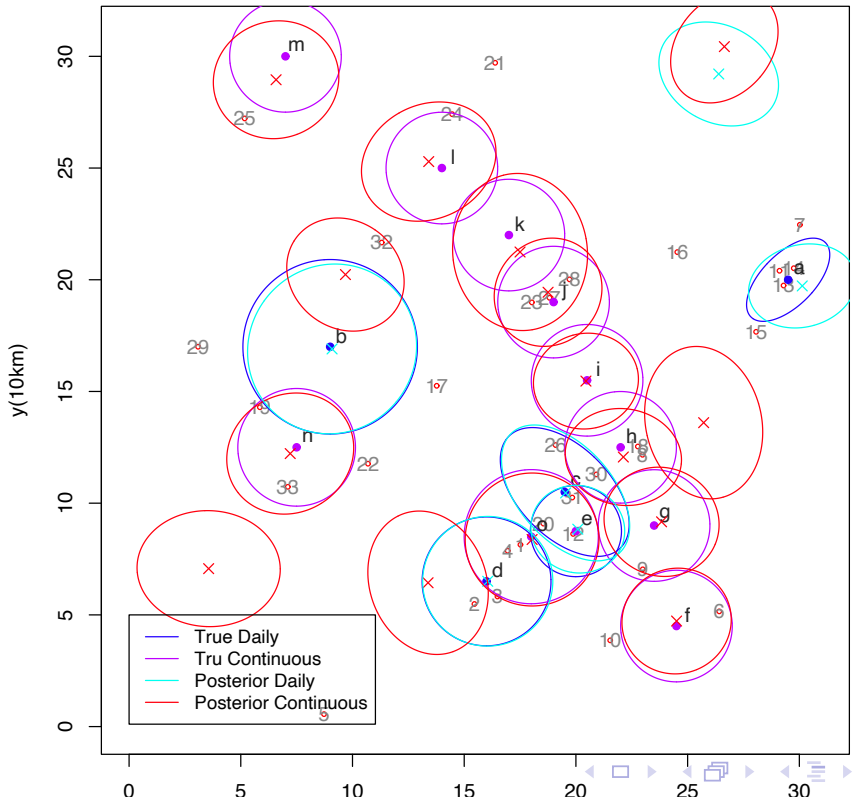


Model Fit at All Locations, One Time:

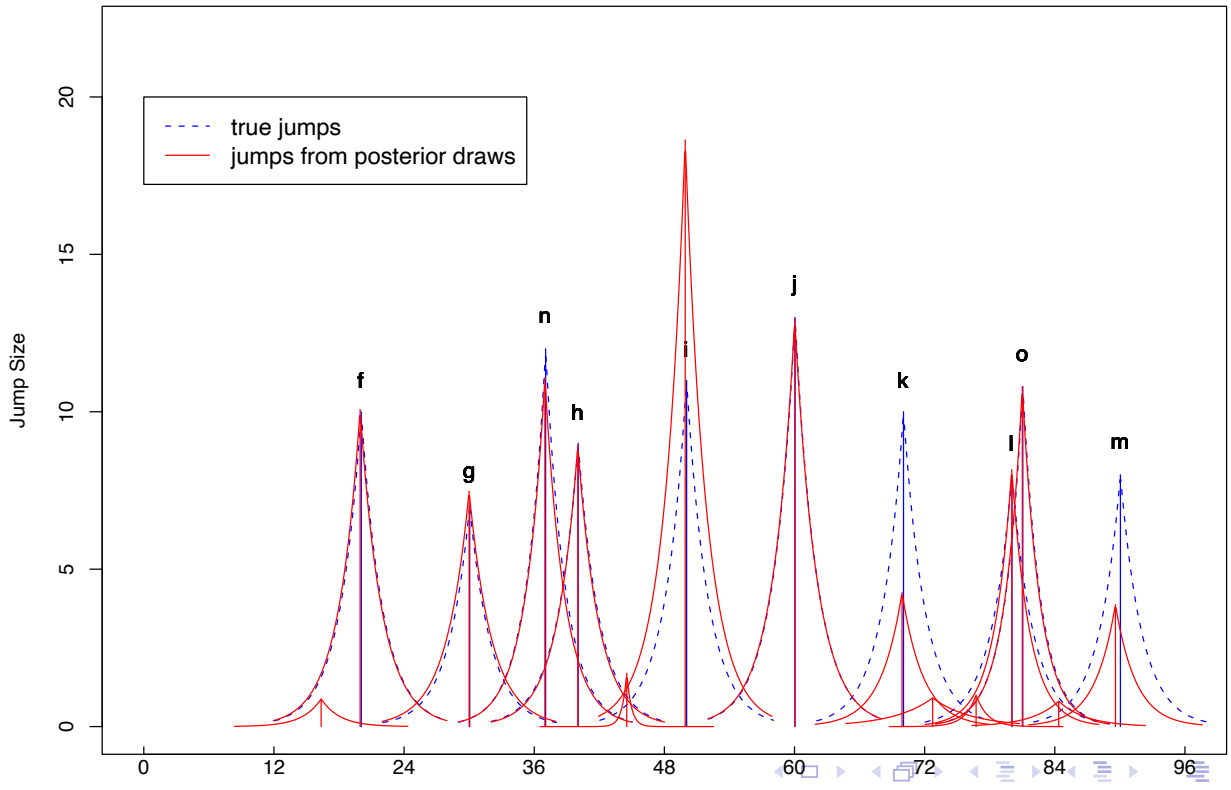




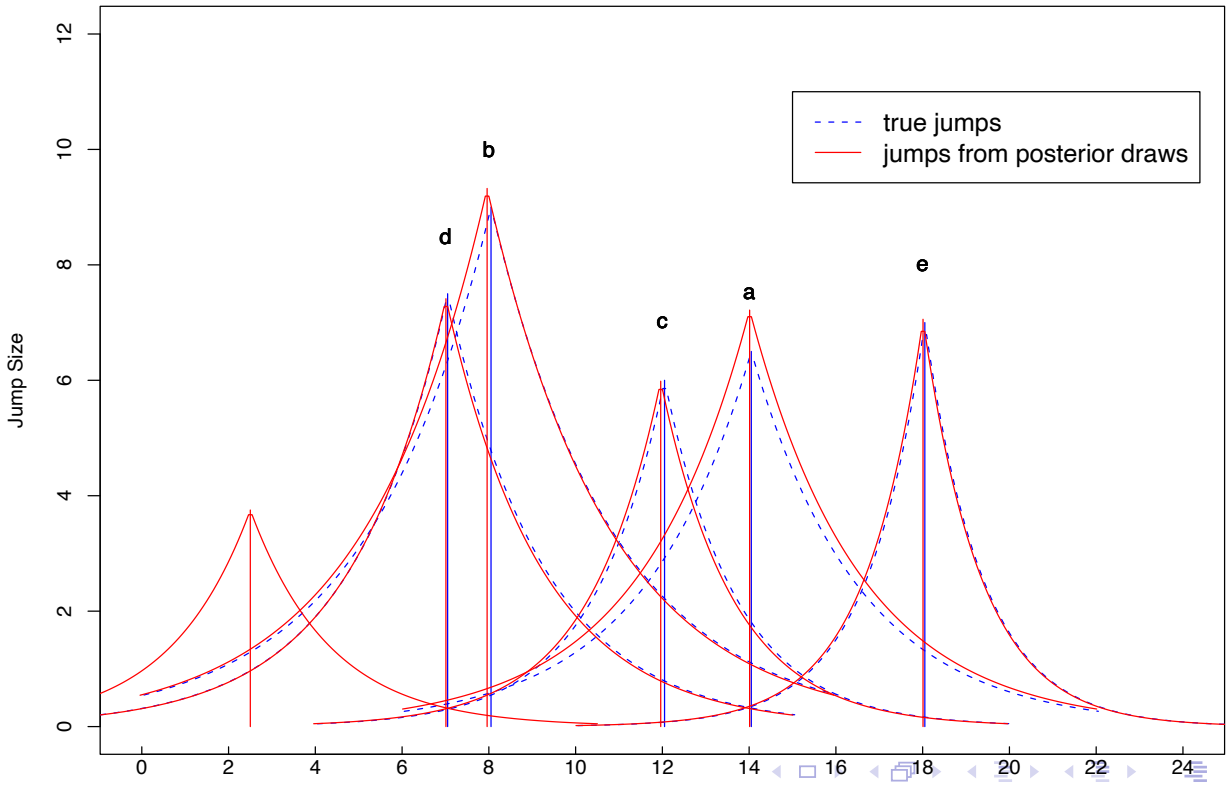
Posterior Estimate in a Movie: MCMC It'ns, Loc's



Posterior Estimate in a Movie: MCMC It'ns, Aper Ker's



Posterior Estimate in a Movie: MCMC It'ns, Dy Ker's



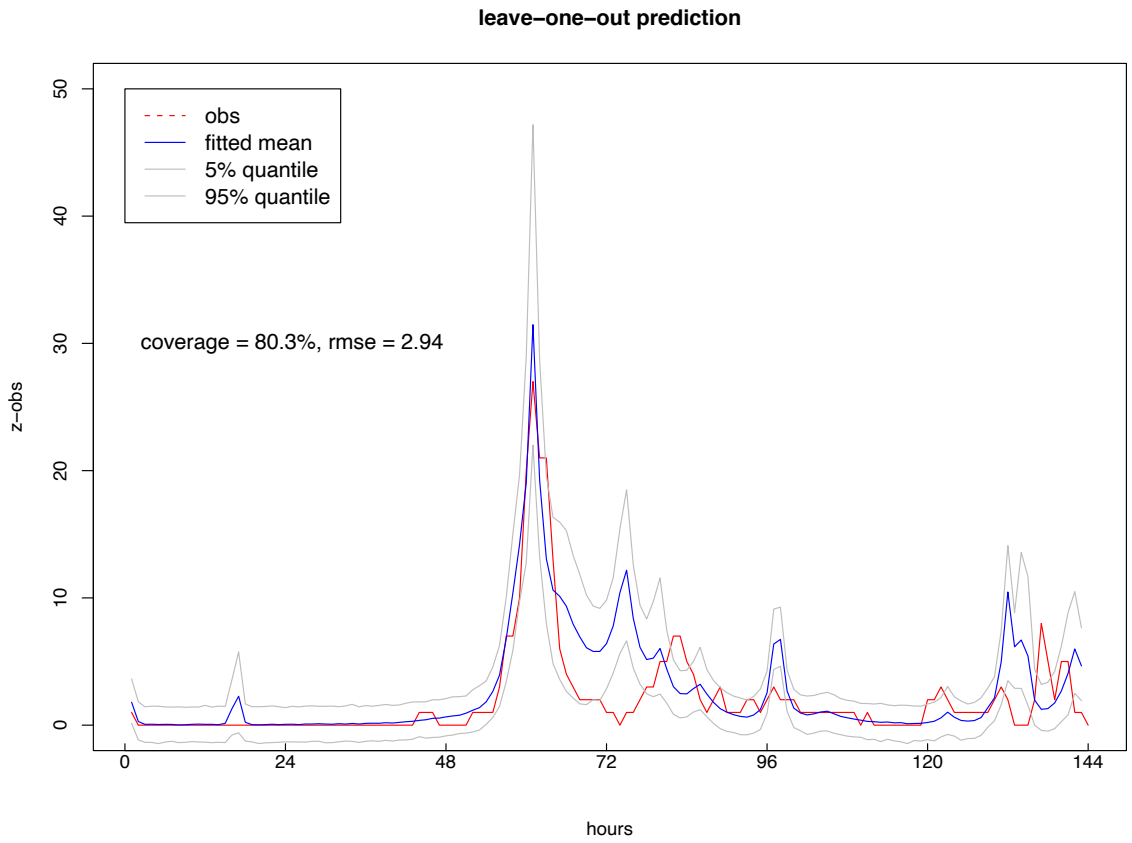
Posterior Estimate in a Movie: Evolution through time

(Loading *movie-surface.avi*)

Posterior Estimate in a Movie: Evolution through time

(Loading *movie-kernel.avi*)

Real data auto-validation: Leave-one-out (Site 31):



Lévy Adaptive Regression Kernel Features

- Limit of finite dimensional priors (GRP & SVSS)
- Flexible generating functions (nonparametric)
- Kernel locations and shapes both adaptive
- Sparse representations
- Modelling Dependence

within: Kernel "decay"
across: Shared jumps

- Expressions for Means & Covariances
- Nonstationarity

Features

- Interpretation of model parameters
- Computationally tractable as coefficients updated individually or in small blocks
- Non-Gaussian prior and likelihoods
- Missing observations, irregular space and time

Ongoing Work & Extensions

- Other Lévy processes (α-Stable)
- Multivariate processes
- Functional Data Analysis
- Arbitrary $\mathcal{X}$ (higher dimensions)
- Theoretical properties (consistency)

Thanks!

More details and related work are available at

www.stat.duke.edu/∼clyde/

or on request from

clyde@stat.duke.edu.

Many thanks to Jenhwa Chu, Leanna House, Zhi Ouyang, Chong Tu!