



Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Summary

Nonparametric Function Estimation Using Overcomplete Dictionaries

Merlise Clyde¹

Institute of Statistics and Decision Sciences
Duke University

Graybill Conference on MultiScale Methods
June 12, 2006

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



Problem Setting

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

- ▶ Model:
 - ▶ Observe data $\{Y_i, \mathbf{x}_i\} \quad i = 1, \dots, n$
 - ▶ $E[Y | \mathbf{x}] = f(\mathbf{x}), \quad \mathbf{x} \in \mathcal{X}$
 - ▶ Goal: inference about unknown $f(\mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}$
- ▶ 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions

Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions

Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

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?

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Expansions

Overcomplete

Dictionaries

Lévy Random

Field Priors

Examples

Summary

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:

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



Which Basis Expansion?

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Expansions

Overcomplete

Dictionary

Lévy Random

Field Priors

Examples

Summary

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)

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

- ▶ 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

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



Lévy Adaptive Regression Kernels

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

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

$$g(x, \omega_j) = \exp(-\lambda_j |x - \chi_j|^{\rho_j})$$

$$f(x) = \sum_{j \leq J} g(x; \omega_j) \beta_j \equiv \int_{\Omega} g(x; \omega) \mathcal{L}(d\omega)$$

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

- ▶ 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

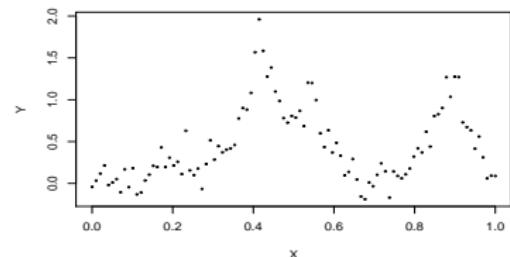
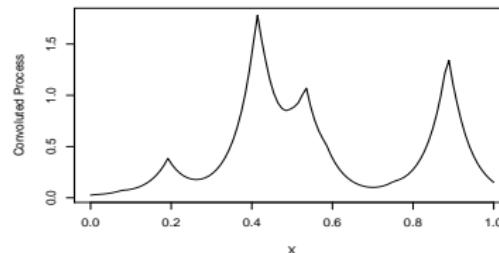
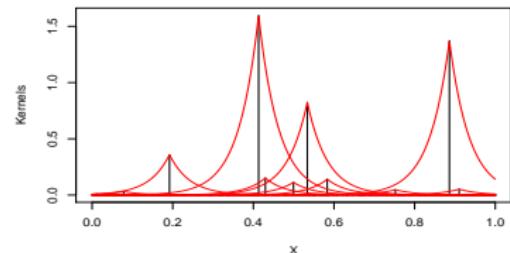
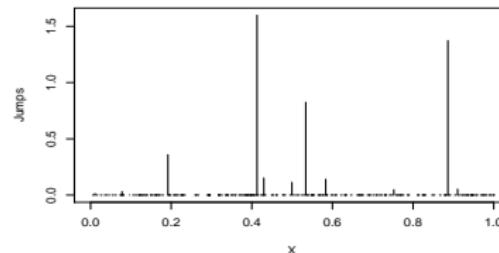
Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary





Why Overcomplete Dictionaries?

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries

Lévy Random
Field Priors

Examples

Summary

- + 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Expansions

Overcomplete
Dictionaries

Lévy Random
Field Priors

Examples

Summary

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





Lévy Random Fields

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

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

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

where ν satisfies integrability condition:

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

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



Poisson Representation

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

Goal: $f(x) = \sum_{j < J} g(\mathbf{x}, \omega_j) \beta_j$

Sufficient condition:

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

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

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

- ▶ Finite number of “big” coefficients $|\beta_j|$
- ▶ Possibly infinite number of $\beta \in [-\epsilon, \epsilon]$
- ▶ Jumps $|\beta_j|$ are absolutely summable²

²need 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

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

$$\beta_j \stackrel{iid}{\sim} \text{Ga}(0, \tau)$$

- ▶ Symmetric Gamma:

$$\nu(d\beta, d\omega) = |\beta|^{-1} \exp(-\tau|\beta|) d\beta \gamma(d\omega)$$

- ▶ Cauchy: $\nu(d\beta, d\omega) = |\beta|^{-2} d\beta \gamma(d\omega)$

$$\beta_j | \lambda_j \stackrel{ind}{\sim} \text{N}(0, 1/\lambda_j) \quad \lambda_j \stackrel{iid}{\sim} \text{Ga}(1/2, 0)$$

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

$$\beta_j | \lambda_j \stackrel{ind}{\sim} \text{N}(0, 1/\lambda_j) \quad \lambda_j \stackrel{iid}{\sim} \text{Ga}(\alpha/2, 0) \quad 0 < \alpha < 2$$

Provides a generalization of [Generalized Ridge Priors](#) to infinite dimensional case



Approximating Lévy Random Fields

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

In practice, cannot use infinite expansion

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

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

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

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



Truncated Cauchy

Nonparametric
Function
Estimation

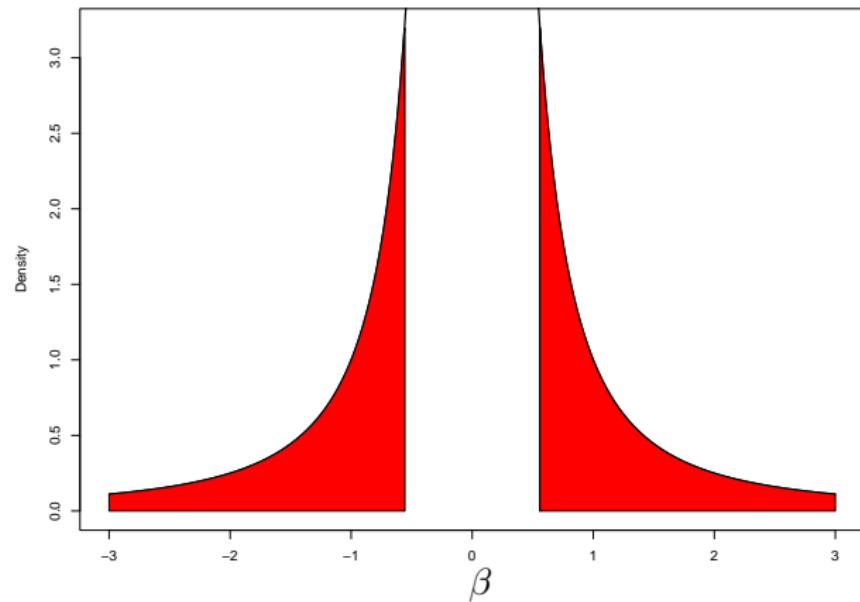
M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

Examples

Summary

Restriction $|\beta| > \epsilon$





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

Nonparametric
Function
Estimation

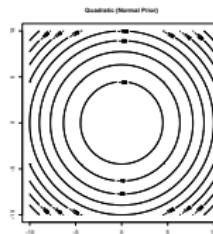
M. Clyde

Nonparametric
Regression
Expansions
Overcomplete
Dictionaries
Lévy Random
Field Priors

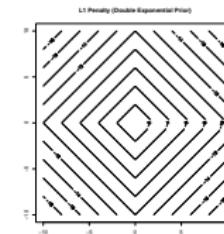
Examples

Summary

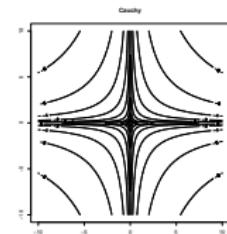
Normal



DE



Cauchy



Penalized Likelihood:

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



Wavelet Test Functions (SNR = 7)

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

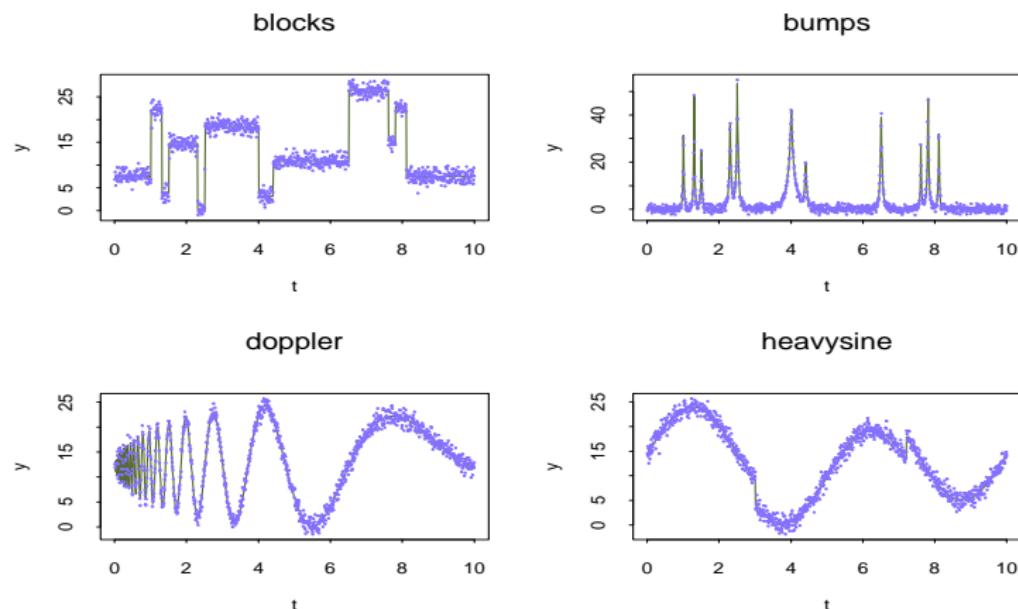
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Kernel Functions

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

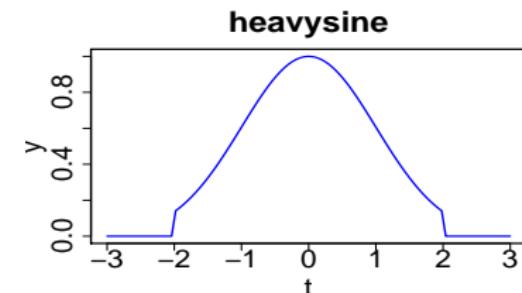
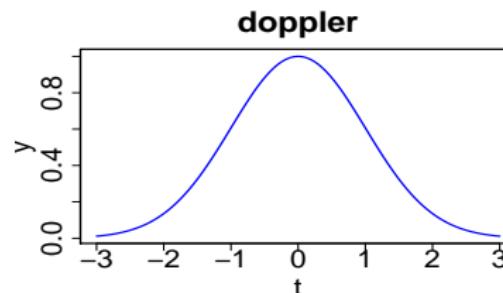
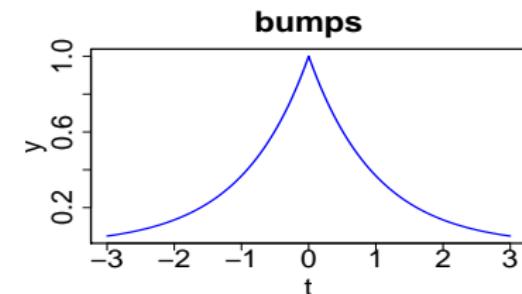
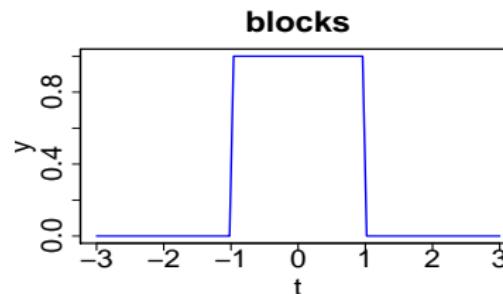
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Comparisons of OCD Methods

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

- ▶ 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

Multidimensional
Time Series

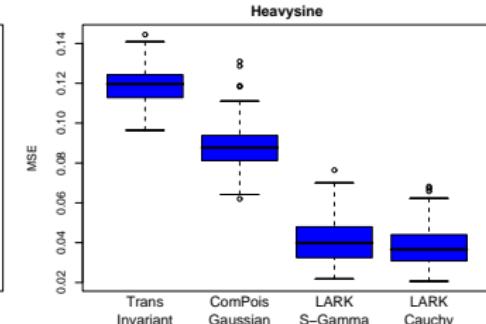
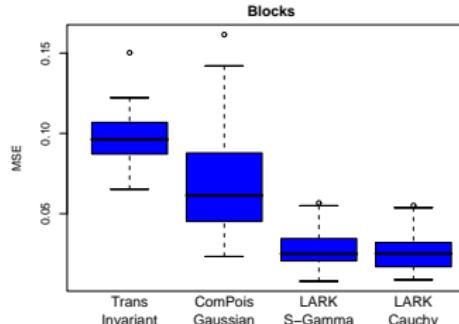
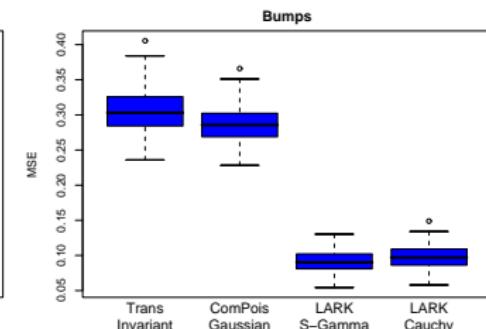
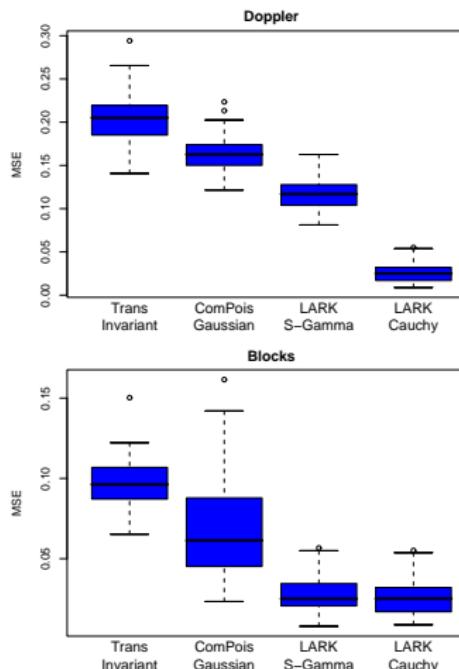
LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

100 realizations of each function





Motorcycle Crash Data

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

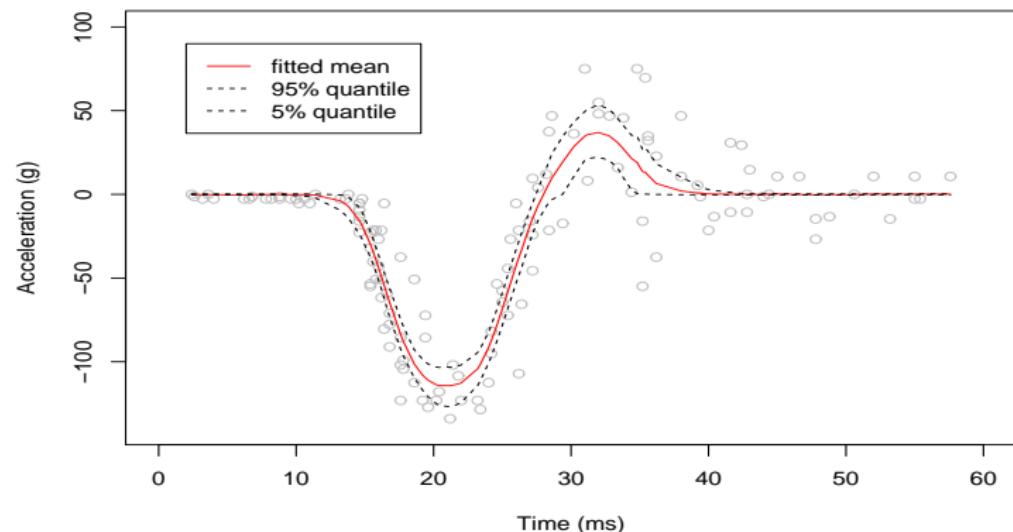
LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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





Posterior on Kernel Power

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

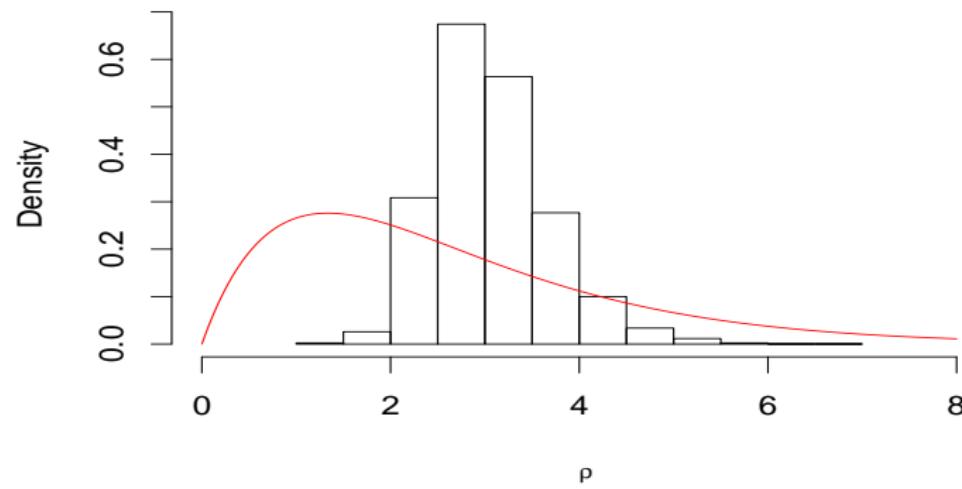
LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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





MALDI-TOF Mass Spectroscopy

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

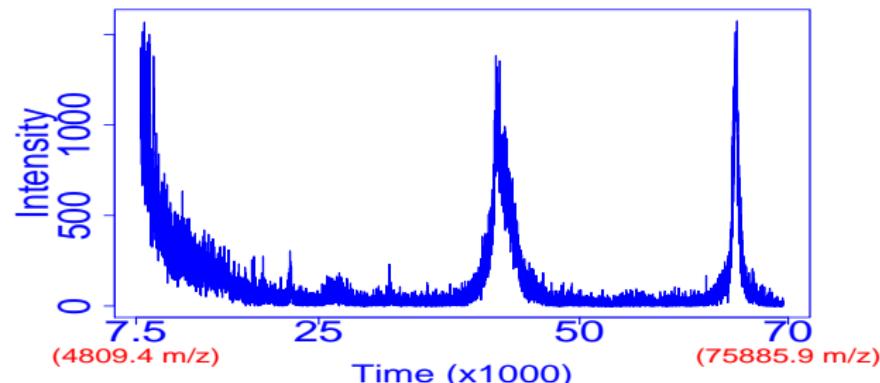
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary



Proteins correspond to peaks in the spectrogram



Single Spectrum Likelihood

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

Model for Measured Intensity at TOF t

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

$$Y(t) - b \stackrel{ind}{\sim} \text{Ga}(f(t) \cdot \phi, \phi)$$

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

$$f(t) = \sum_j^J k(t, \theta_j) \beta_j = \int_{\Theta} k(t, \theta) \Gamma(d\theta)$$



Natural Basis Functions

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK

Space-Time
Models

Fitting

Simulated Data

Fitting Real
Data

Summary

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

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

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

Both τ and λ may vary from peak to peak



Induced Prior on f

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results
Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

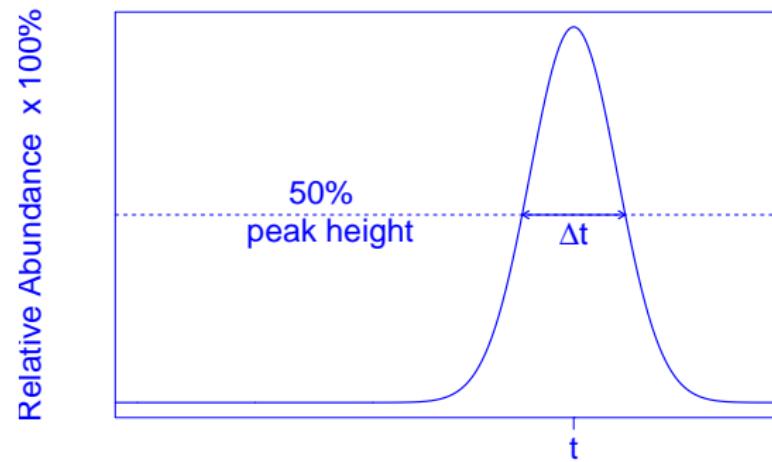
Fitting Real
Data

Summary

- ▶ $f(t) = \sum_j^J k(t, \theta_j) \beta_j$
- ▶ 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)
- ▶ $\theta = (\tau, \lambda)$
 - ▶ $\{\tau_j\}$ Expected TOF of protein (unknown) (Uniform)
 - ▶ $\{\lambda_j\}$ peak width – determined by prior on resolution



Mass Resolution



$$\text{Resolution } \varrho \equiv \Delta t / \tau$$

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



Hierarchical Prior on Resolution

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

- ▶ Individual Peak Resolution

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

- ▶ 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

- ▶ 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 $E[f(\cdot)]$



Estimated Spectrum

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

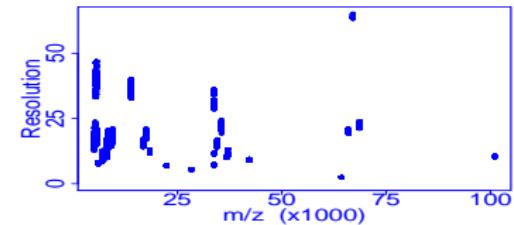
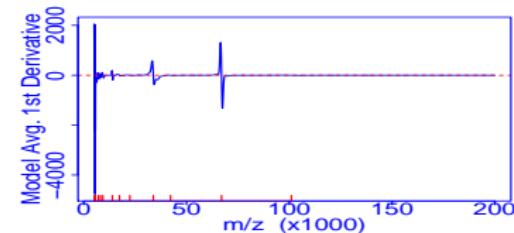
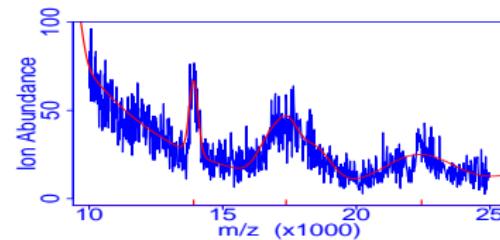
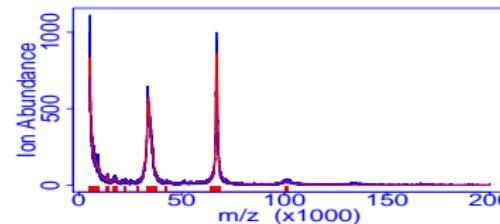
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Features

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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 PM₁₀ concentration in Maricopa County, AZ

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK

Space-Time
Models

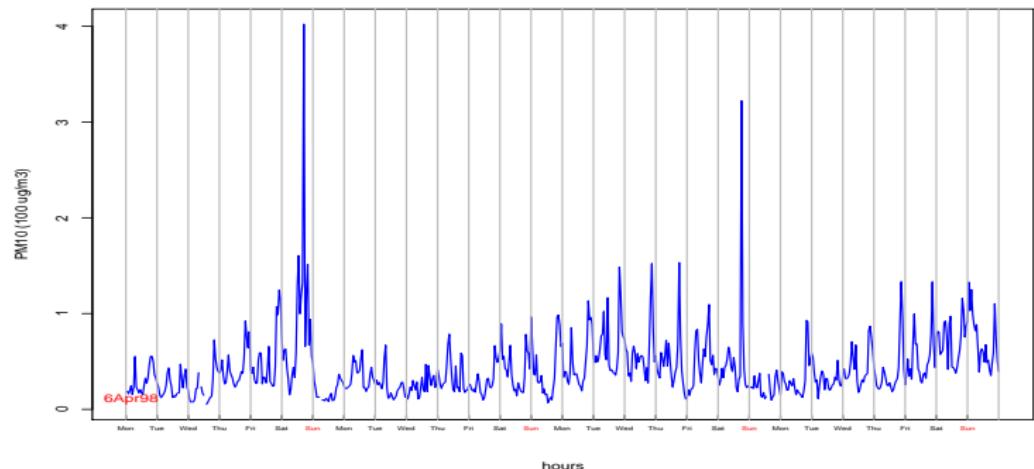
Fitting

Simulated Data

Fitting Real
Data

Summary

April 1998



The “Spikey” concentration profiles don’t fit ARMA well.
Semi-periodic with possible daily and meteorologically driven
patterns.



Marked Lévy model

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

$$Y_{t_i} = f(t_i) + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

$$f(t) = b_0 + \int_{\Omega} k(t; \omega) \Gamma(d\omega) \quad \checkmark \text{ Marks}$$

$$\Omega = [0, 720] \times \mathbb{R}_+ \times \{0, 1\}$$

$$k(t; \omega) = k(t; (\tau, \lambda, a))$$

$$= \begin{cases} e^{-\lambda|t-\tau|} & a = 0, \\ e^{-\lambda|(t-\tau) \pmod{24}|} & a = 1, \end{cases} \quad \begin{array}{l} \text{Aperiodic part} \\ \text{Daily part} \end{array}$$



The Prior Distribution:

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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

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

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



The Fitted Model

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

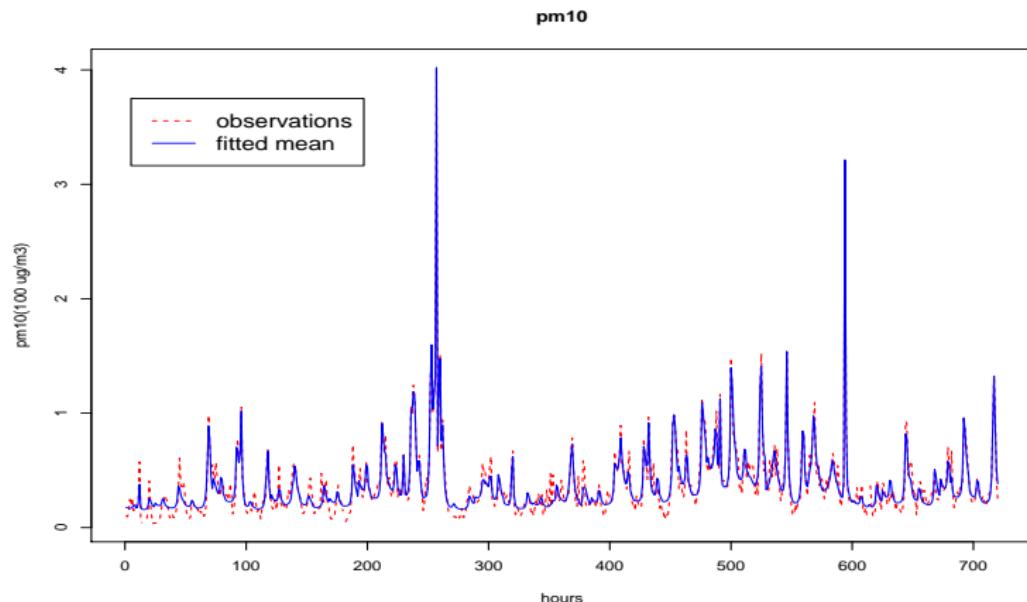
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





The Fitted Model: Aperiodic Part

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK

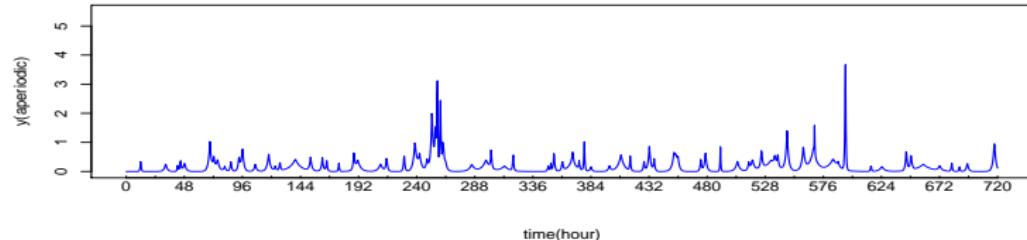
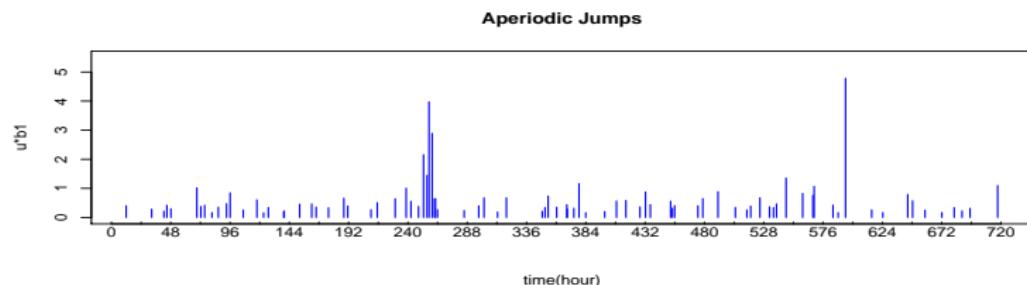
Space-Time
Models

Fitting

Simulated Data

Fitting Real

Data





The Fitted Model: Daily Part

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics
Results

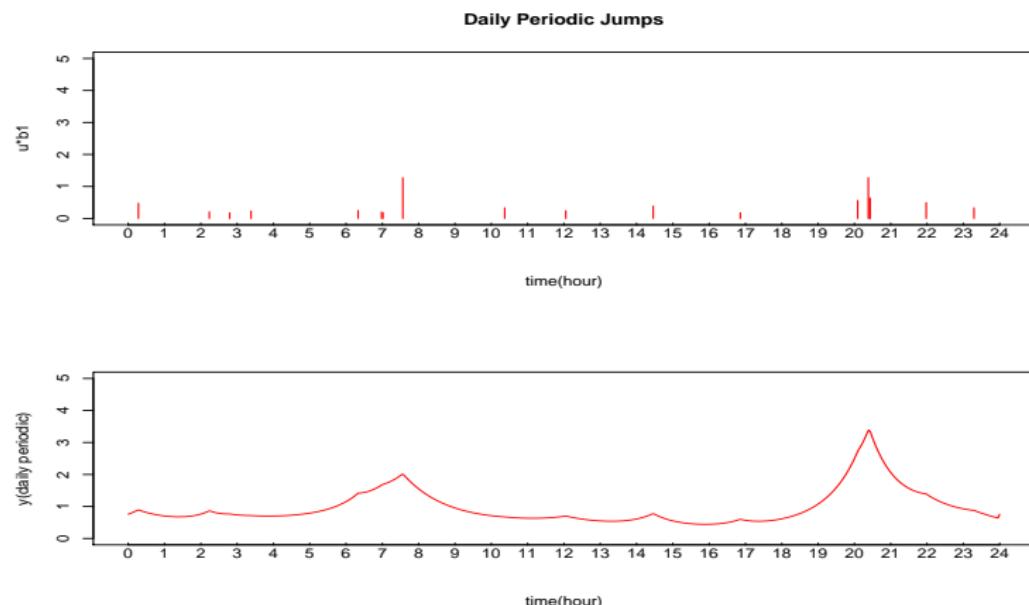
Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data
Fitting Real
Data

Summary





The Fitted Model: Predictions

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

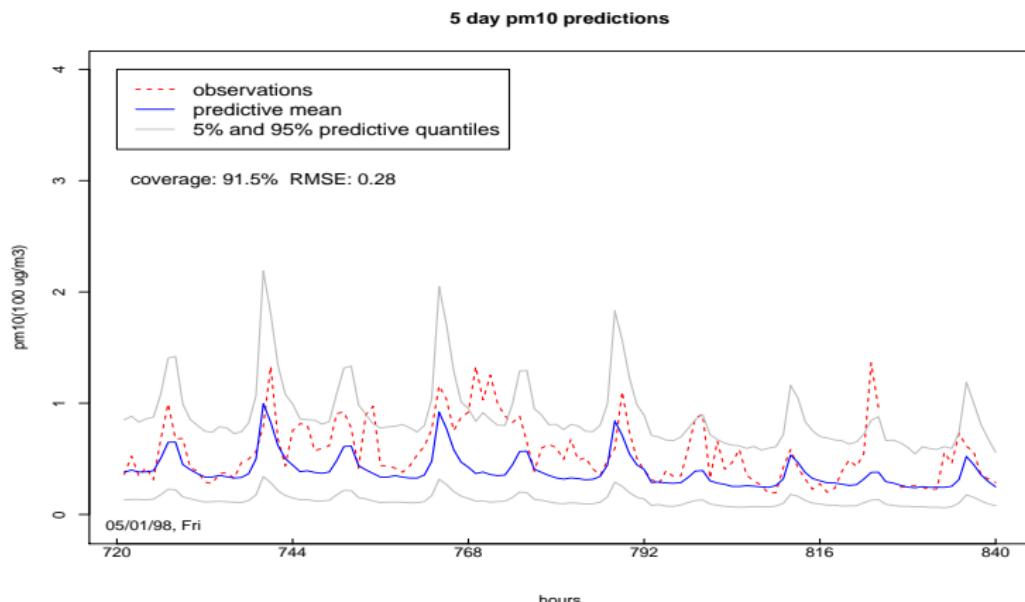
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Two Pollutants: PM₁₀ and CO

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

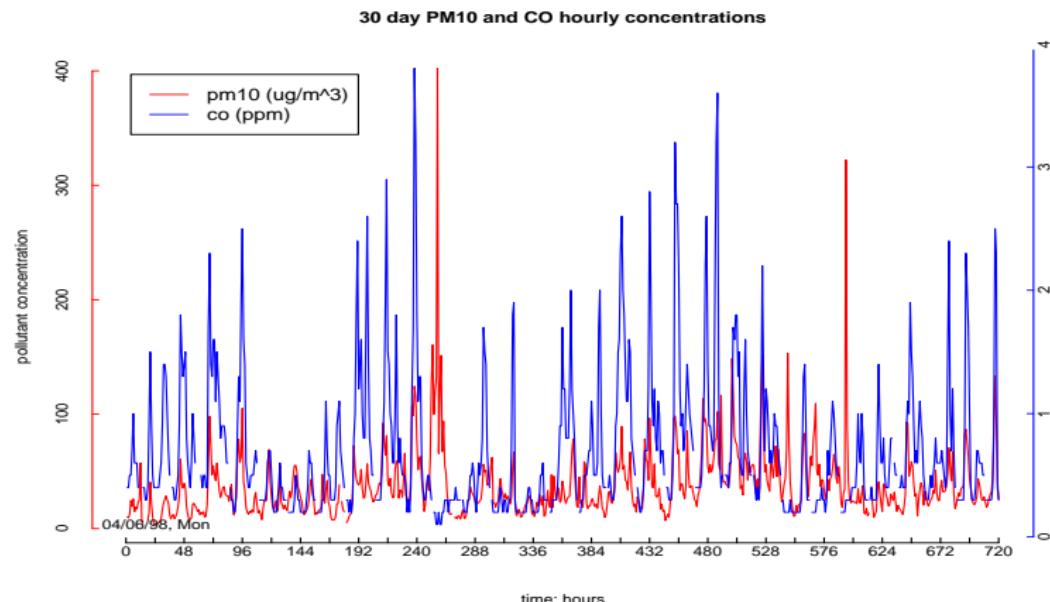
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Marks

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

six marks

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

where

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



Fits and Predictions: PM₁₀ and CO

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics
Results

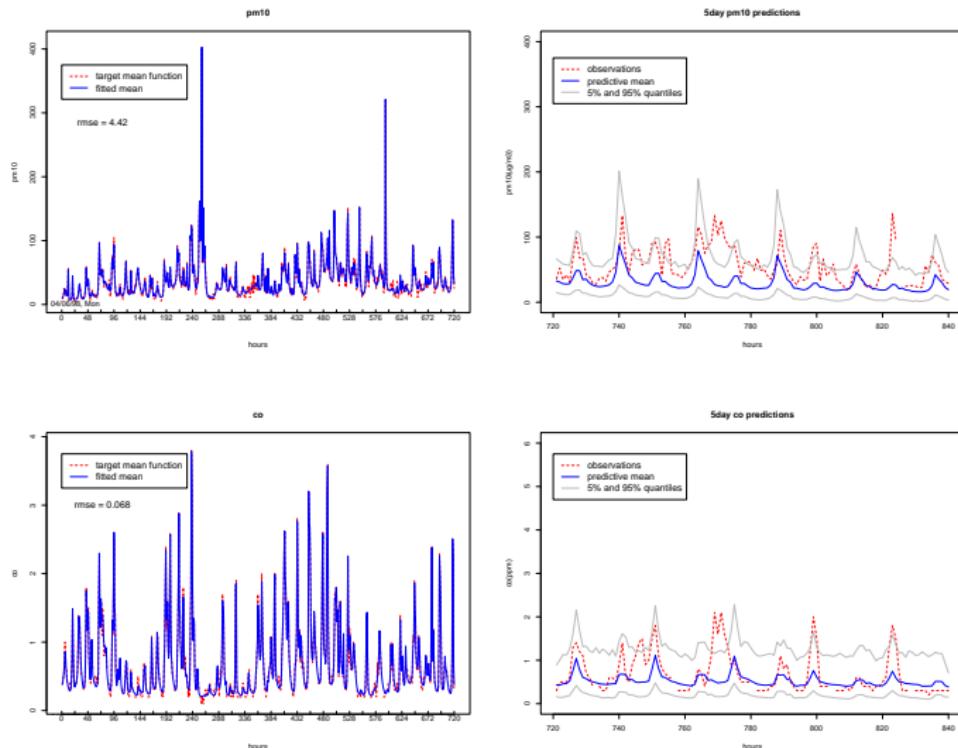
Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data
Fitting Real
Data

Summary





Decomposition, PM₁₀

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

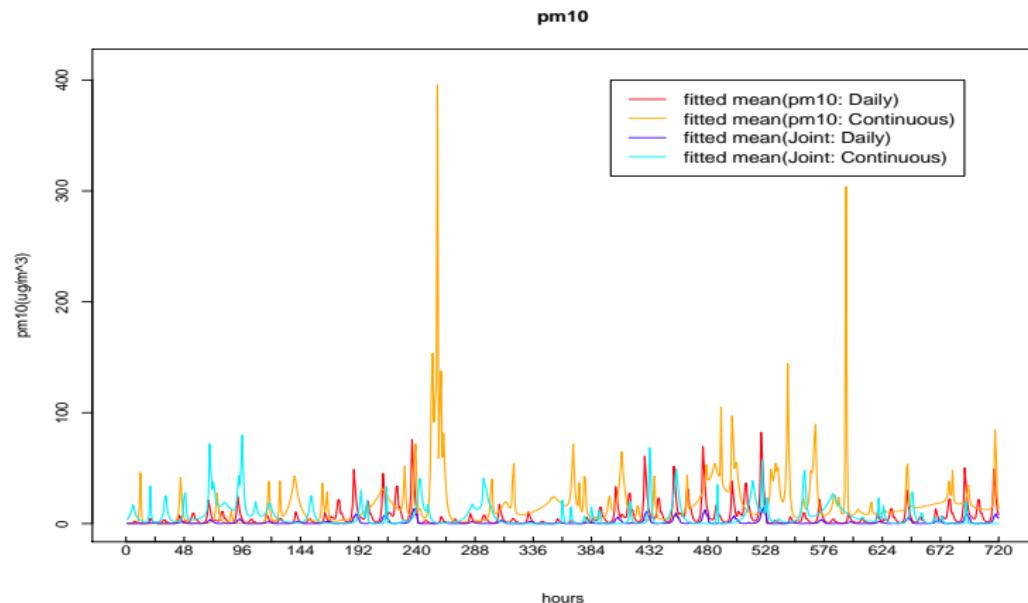
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Decomposition, CO

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK

Space-Time
Models

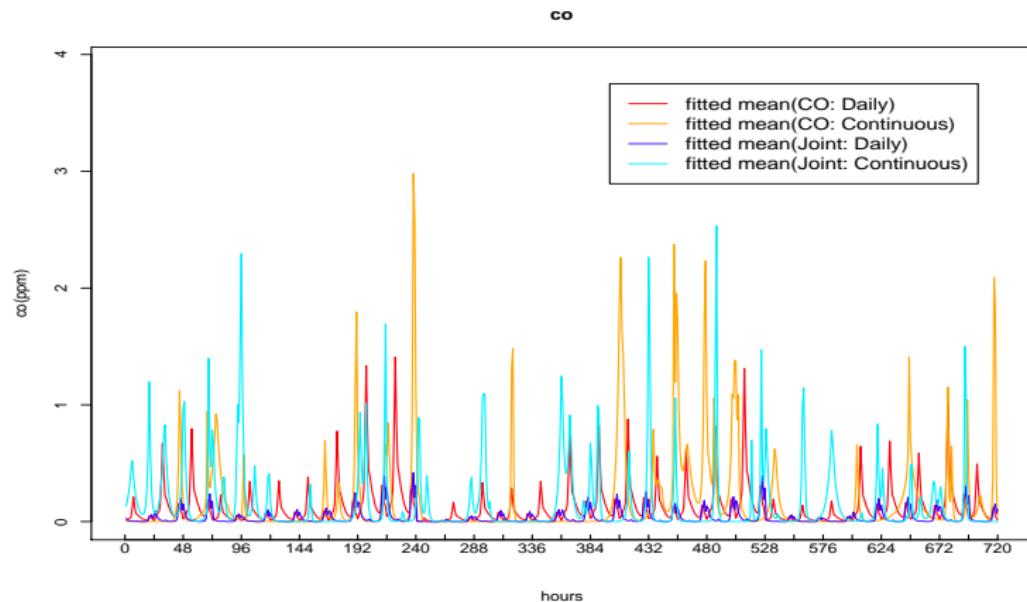
Fitting

Simulated Data

Fitting Real

Data

Summary





Benefits of the Method

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

- ▶ Non-negative data (here, $[PM_{10}]$ and $[CO]$) 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

- ▶ Non-negative data (here, $[PM_{10}]$ and $[CO]$) 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

- ▶ Non-negative data (here, $[PM_{10}]$ and $[CO]$) 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

- ▶ Non-negative data (here, $[PM_{10}]$ and $[CO]$) 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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 ,

$$Z_i = f(s_i, t_i) + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \text{No}(0, \sigma^2)$$

$$f(s, t) = b_0 + \int_{\Omega} k(s, t; \omega) \Gamma(d\omega)$$

$$\Omega = \mathcal{S} \times \mathcal{T} \times \mathbb{R}_+ \times \mathcal{M}_+ \times \mathcal{A}$$

$$k(s, t; \omega) = k(s, t; (\sigma, \tau, \lambda, \Lambda, a))$$

$$= \begin{cases} e^{-\lambda|t-\tau| - (s-\sigma)' \Lambda (s-\sigma)/2} & a = 0 \text{ (Aper)} \\ e^{-\lambda|(t-\tau) \pmod{24}| - (s-\sigma)' \Lambda (s-\sigma)/2} & a = 1 \text{ (Daily)} \end{cases}$$

$(\sigma, \tau) \in \mathcal{S} \times \mathcal{T}$ is space-time center (random, adaptive)

$\lambda \in \mathbb{R}_+$ is Temporal Decay rate (random, adaptive)

$\Lambda \in \mathcal{M}_+$ is Spat dispers'n matrix (random, adaptive)



Spatial-Temporal Models

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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 ,

$$Z_i = f(s_i, t_i) + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \text{No}(0, \sigma^2)$$

$$f(s, t) = b_0 + \int_{\Omega} k(s, t; \omega) \Gamma(d\omega)$$

$$\Omega = \mathcal{S} \times \mathcal{T} \times \mathbb{R}_+ \times \mathcal{M}_+ \times \mathcal{A}$$

$$k(s, t; \omega) = k(s, t; (\sigma, \tau, \lambda, \Lambda, a))$$

$$= \begin{cases} e^{-\lambda|t-\tau| - (s-\sigma)' \Lambda (s-\sigma)/2} & a = 0 \text{ (Aper)} \\ e^{-\lambda|(t-\tau) \pmod{24}| - (s-\sigma)' \Lambda (s-\sigma)/2} & a = 1 \text{ (Daily)} \end{cases}$$

$(\sigma, \tau) \in \mathcal{S} \times \mathcal{T}$ is space-time center (random, adaptive)

$\lambda \in \mathbb{R}_+$ is Temporal Decay rate (random, adaptive)

$\Lambda \in \mathcal{M}_+$ is Spat dispers'n matrix (random, adaptive)



Spatial-Temporal Models

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

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 ,

$$Z_i = f(s_i, t_i) + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} \text{No}(0, \sigma^2)$$

$$f(s, t) = b_0 + \int_{\Omega} k(s, t; \omega) \Gamma(d\omega)$$

$$\Omega = \mathcal{S} \times \mathcal{T} \times \mathbb{R}_+ \times \mathcal{M}_+ \times \mathcal{A}$$

$$k(s, t; \omega) = k(s, t; (\sigma, \tau, \lambda, \Lambda, a))$$

$$= \begin{cases} e^{-\lambda|t-\tau| - (s-\sigma)' \Lambda (s-\sigma)/2} & a = 0 \text{ (Aper)} \\ e^{-\lambda|(t-\tau) \pmod{24}| - (s-\sigma)' \Lambda (s-\sigma)/2} & a = 1 \text{ (Daily)} \end{cases}$$

$(\sigma, \tau) \in \mathcal{S} \times \mathcal{T}$ is space-time center (random, adaptive)

$\lambda \in \mathbb{R}_+$ is Temporal Decay rate (random, adaptive)

$\Lambda \in \mathcal{M}_+$ is Spat dispers'n matrix (random, adaptive)



Sulfur Dioxide Concentrations in PA, MD, NJ:

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

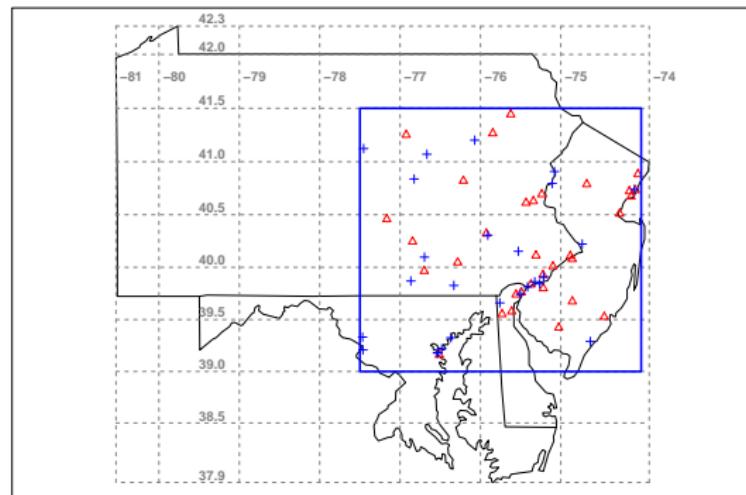
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary



EPA SO₂ Monitoring stations (real ones!) are shown as Δ 's,
Point Sources (power plants, etc.) as $+$'s



Simulated Data

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

Simulated data in a Table
 $J_0 = 10, J_1 = 5, \{(u_j, \omega_j)\}_{j \leq 15} =$

	u	(σ_x, σ_y)	τ	λ_t	(λ_x, λ_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:

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

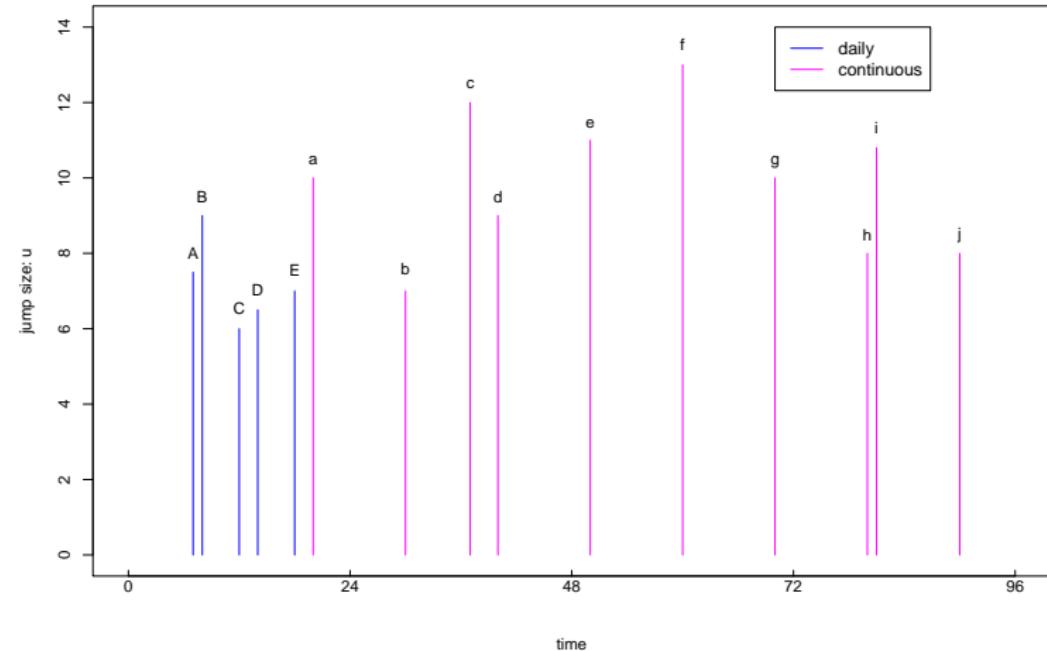
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Simulated data in a Picture:

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

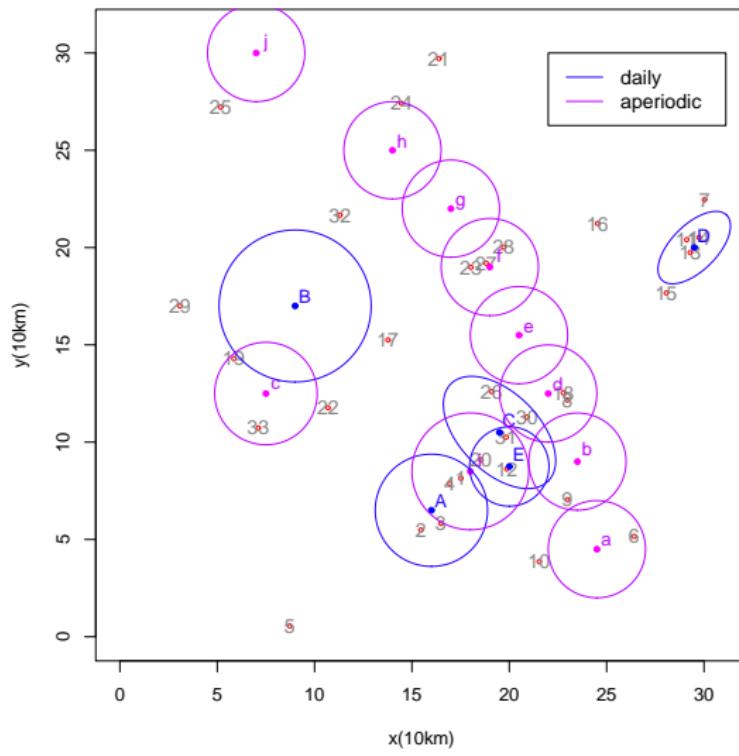
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Model Fit at Nine Locations, All Times:

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

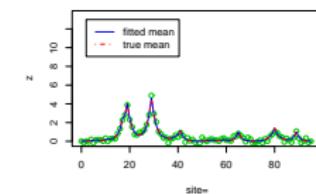
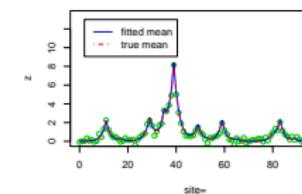
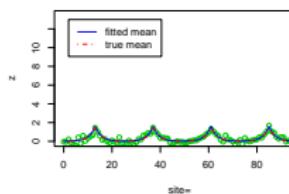
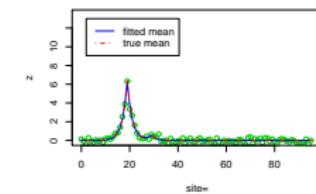
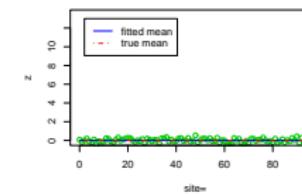
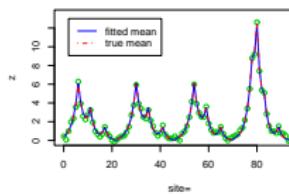
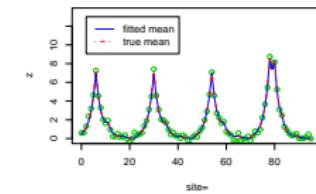
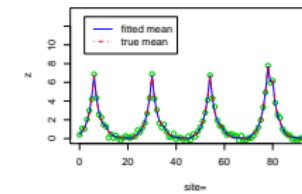
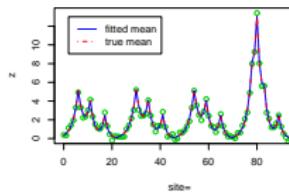
LARK

Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary



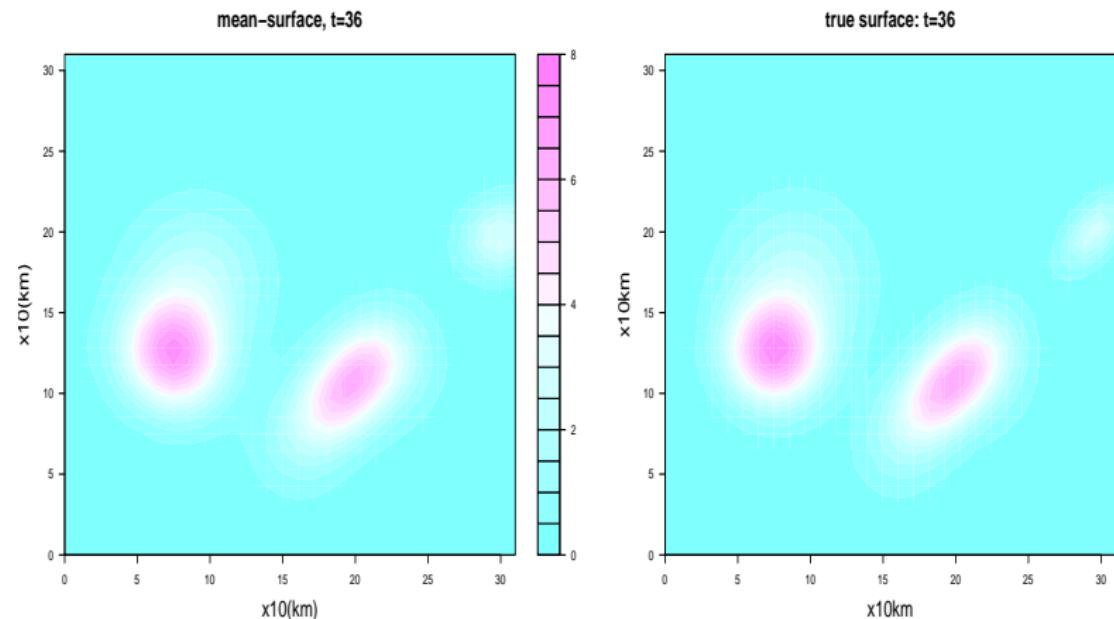


Model Fit at All Locations, One Time:

Nonparametric Function Estimation

M. Clyde

Fitting Simulated Data





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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

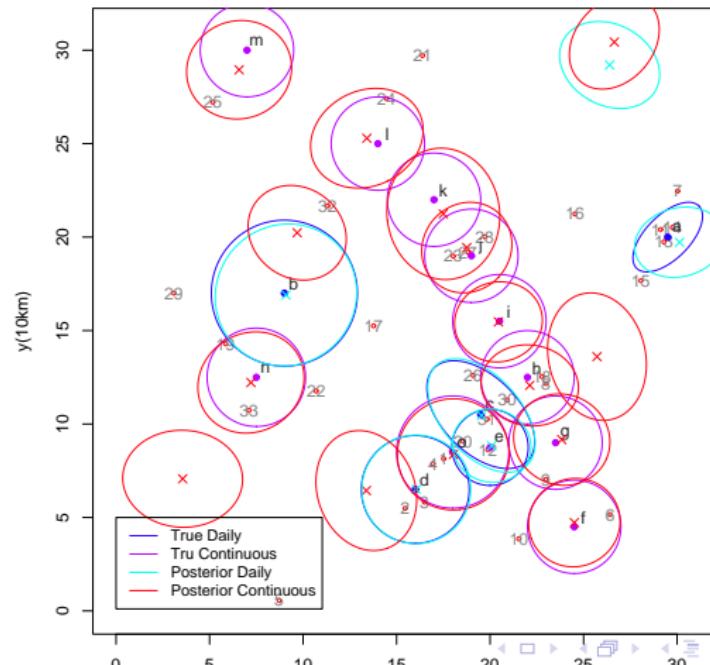
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle

Crash Data

Proteomics

Results

Time Series
Models

Multidimensional
Time Series

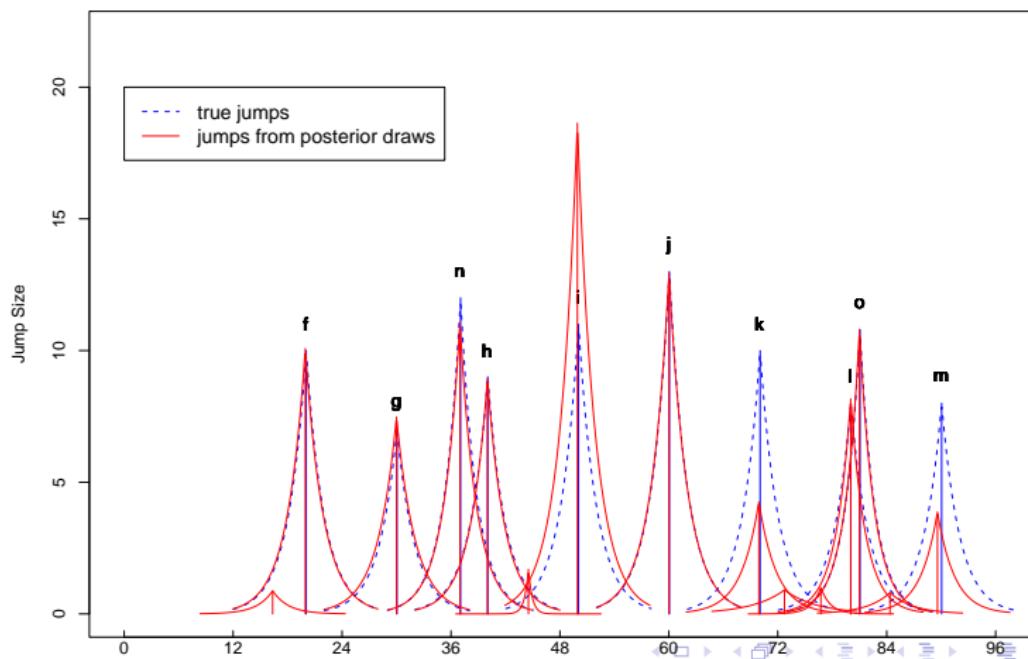
LARK

Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Posterior Estimate in a Movie: MCMC It's, Dyker's

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics
Results

Time Series
Models

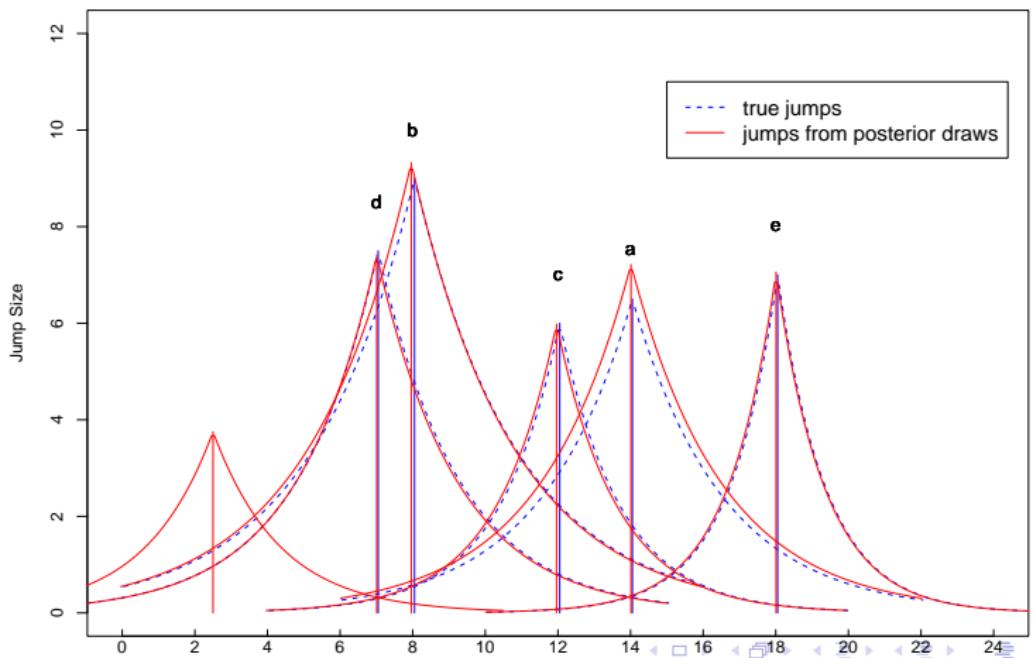
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Posterior Estimate in a Movie: Evolution through time

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data
Proteomics

Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

(Loading *movie-surface.avi*)



Posterior Estimate in a Movie: Evolution through time

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics
Results

Time Series
Models

Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary

(Loading *movie-kernel.avi*)



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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Wavelet Test
Functions

Motorcycle
Crash Data

Proteomics

Results

Time Series
Models

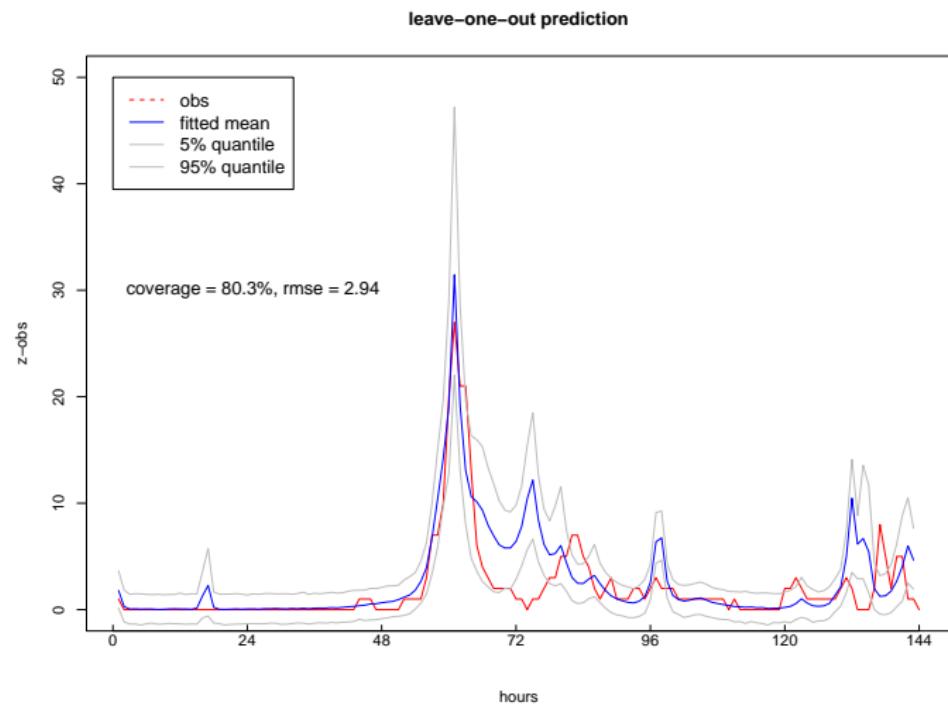
Multidimensional
Time Series

LARK
Space-Time
Models

Fitting
Simulated Data

Fitting Real
Data

Summary





Lévy Adaptive Regression Kernel Features

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Summary

- ▶ 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Summary

- ▶ 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

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Summary

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



Thanks!

Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Summary

More details and related work are available at

www.stat.duke.edu/~clyde/

or on request from

[clyde@stat.duke.edu.](mailto:clyde@stat.duke.edu)



Nonparametric
Function
Estimation

M. Clyde

Nonparametric
Regression

Examples

Summary

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