Lévy-based Nonparametric Bayesian Models and their Applications

Robert L Wolpert ’72

Duke University

rlw@duke.edu

2019 September 06
Motivation

LARK Models

Examples

Conclusion

The Theme...

We teach our students about ARMA, ARIMA, Diffusions, and such, featuring

- Nicely behaved sample paths,
- Tame tail behavior,
- Regularly-spaced observations;

Then they graduate and face data with

- Jumps,
- Heavy tails,
- Spikiness,
- Irregularly-spaced observations &/or missing data.
The Theme...

We teach our students about ARMA, ARIMA, Diffusions, and such, featuring
- Nicely behaved sample paths,
- Tame tail behavior,
- Regularly-spaced observations;

Then they graduate and face data with
- Jumps,
- Heavy tails,
- Spikiness,
- Irregularly-spaced observations &/or missing data.
Time-series Data 1: Rockfalls at Soufrière Hills Volcano

Number of rockfalls per day
(Rockfall data: 12/12/1995 – 06/13/2007)

$^0$SHV on island of Montserrat, BOT in Lesser Antilles, Caribbean.
Time-series Data 2: Proteomics (MALDI-ToF)
Point Process Data 3: Forest Ecology (Spatial Biodiversity)

Oak Trees

Hickories

Data from 140m $\times$ 140m Borman plot in Duke Forest

Robert L Wolpert ’72

Lévy NPB Models & Applications
Time-series Data 4: GRB Light Curves from BATSE

Burst & Transient Source Experiment on Compton GRO
One approach: Lévy Adaptive Regression Kernels

- General goal: inference on unknown function \( f(\cdot) \)
- Usual Kernel regression approximates unknown function with weighted sum of functions
- Adaptive kernel regression infers the kernel shape locally:

\[
 f(x) \approx \sum_j u_j K(x \mid s_j, \theta_j)
\]

where \( x, \{s_j\} \subset S \) are times, locations, etc., and \( \{\theta_j\} \subset \Theta \) determine the kernel shapes.

- Good things happen if we take \( \{(u_j, s_j, \theta_j)\} \) to be \( \text{spt}(H) \) for a Poisson random measure \( H \sim \text{Po}(\nu(du \, ds \, d\theta)) \).
One approach: Lévy Adaptive Regression Kernels

- General goal: inference on unknown function $f(\cdot)$
- Usual Kernel regression approximates unknown function with weighted sum of functions
- Adaptive kernel regression infers the kernel shape locally:

$$f(x) \approx \sum_{j} u_j K(x \mid s_j, \theta_j)$$

where $x, \{s_j\} \subset \mathcal{S}$ are times, locations, etc., and $\{\theta_j\} \subset \Theta$ determine the kernel shapes.

- Good things happen if we take $\{(u_j, s_j, \theta_j)\}$ to be $\text{sp}\mathcal{T}(H)$ for a Poisson random measure $H \sim \text{Po}(\nu(du \, ds \, d\theta))$. 
LARK as a Stochastic Integral

\[ f(x) = \sum_j u_j K(x | s_j, \theta_j) = \int_{\mathbb{R} \times S \times \Theta} u K(x | s, \theta) H(du \, ds \, d\theta) \]

- Infinitely-many terms if \( \nu(\mathbb{R} \times S \times \Theta) = \infty \)
- But \( f(x) < \infty \) a.s. if \( u K(x | s, \theta) \) is in the Musielak-Orlicz space of functions that satisfy

\[ \int_{\mathbb{R} \times S \times \Theta} \left( 1 \wedge |u K(x | s, \theta)| \right) \nu(du \, ds \, d\theta) < \infty \]
Features of LARK Models

\[ f(x) = \int_{\mathbb{R} \times S \times \Theta} u K(x \mid s, \theta) H(du \, ds \, d\theta) \]

- Marginal dist’ns of \( f(x) \) are ID (Infinitely-Divisible);
- Any ID dist’n can be attained with suitable Lévy Measure \( \nu(du \, ds \, d\theta) \): Po, Ga, \( \alpha \)St, IG, NB, No, ...
- Theorem: Any Stationary Moving Average process is LARK with kernel \( K(x \mid s, \theta) = b_\theta(x - s) \) (plus Wiener integral)

\[ f(x) = \int_{\mathbb{R}^n} b_\theta(x - s) \zeta(ds \, d\theta) + \int_{\mathbb{R}^n} \ldots \, W(ds) \]
Bayesian Inference for LARK Models

More important: Bayesian Inference is straightforward:

\[ f(x) = \sum_{j} u_j K(x \mid s_j, \theta_j) \]

1. Find Likelihood Function describing how badly \( f(x) \) fits data;
2. Truncate to a finite sum with (random?) \( J \in \mathbb{N} \) terms;
3. Wiggle \( J \) and the \( \{(u_j, s_j, \theta_j)\} \) in a RJ-MCMC scheme;
4. Generate posterior samples of anything you like.
Bayesian Inference for LARK Models

More important: **Bayesian Inference** is straightforward:

\[ f(x) = \sum_{j} u_j K(x \mid s_j, \theta_j) \]

1. Find **Likelihood Function** describing how badly \( f(x) \) fits data;
2. Truncate to a finite sum with (random?) \( J \in \mathbb{N} \) terms;
3. Wiggle \( J \) and the \( \{(u_j, s_j, \theta_j)\} \) in a RJ-MCMC scheme;
4. Generate posterior samples of anything you like.
Example 1: Biomass and Biodiversity

We construct a moving-average Cox model, with:

- Inhomogeneous Poisson random field for trees;
- Intensity is moving average of latent Gamma random field (Poisson/Gamma conjugacy lends computational advantages);
- Posterior mean of Poisson intensity is NPB estimate of tree density;
- Simultaneous estimation for eight species leads to spatial biodiversity index.
Example 1: Biomass and Biodiversity

We construct a moving-average Cox model, with:

- Inhomogeneous Poisson random field for trees;
- Intensity is moving average of latent Gamma random field (Poisson/Gamma conjugacy lends computational advantages);
- Posterior mean of Poisson intensity is NPB estimate of tree density;
- Simultaneous estimation for eight species leads to spatial biodiversity index.
Over-story Trees \((D > 25\text{cm})\) in Bormann Plot

Eight species of large trees in Duke Forest
Motivation

LARK Models

Examples

Conclusion

Moving-Average Cox Model for Oak Density

Trees: \( N(dx) \sim \text{Po}(\Lambda(x) \, dx) \)

Intensity: \( \Lambda(x) = \int_{S \times \Theta} k(x - s \mid \theta) \zeta(ds \, d\theta), \quad x \in S \)

\[ = \int_{\mathbb{R} \times S \times \Theta} k(x - s \mid \theta) \, u \, H(du \, ds \, d\theta) \]

Innovation: \( \zeta(ds \, d\theta) \sim \text{Ga}(\alpha(ds \, d\theta), \beta(s, \theta)) \)

Poisson Rep’n: \( H(du \, ds \, d\theta) \sim \text{Po}(\alpha(ds \, d\theta) \, u^{-1} e^{-\beta(s, \theta) u} \, du) \)

Kernel: \( k(x - s \mid \theta) = e^{-(x-s)' \Lambda(x-s)/2}, \quad \theta = (\Lambda, \ldots) \)

Features of \( \alpha(ds \, d\theta), \, \beta(s, \theta), \, k(x - s \mid \theta) \) may be treated as uncertain, with joint prior distributions.
Posterior Image Estimate of Oak Density
Posterior Contour Estimate of Oak Density
Oaks and Hickories

Note Oaks are under-dispersed, Hickories over-dispersed.
Posterior Estimates of Hickory Density
Spatial Biodiversity

How Much Diversity? One Species or Two?

Each figure has same number of Red and Blue dots.
Spatial Biodiversity

How Much Diversity? One Species or Two?

Each figure has same number of Red and Blue dots.
Spatial Hill’s Index of Biodiversity

Mark Hill’s *Equivalent Number of Species* index (*Ecol.*, 1973)

\[ 1 \leq H_1 \equiv \exp(H) = \prod_{i=1}^{n} \left( \frac{1}{p_i} \right)^{p_i} \leq n \]
Example 2: Bayesian Semipar. Spatial Epidemiology

Does traffic pollution induce respiratory disease in children?

Model spatially-varying disease rate $\Lambda(s)$ (cases/100 pop) dependence on:

- Individual-level covariates (sex, parental smoking, coal);
- Spatially-varying covariates ($\text{NO}_2$ levels as surrogate);
- Unattributed spatial variation (possible clues for etiology!).

Use LARK to capture Unattributed spatial variation.
The Study Area: Huddersfield, UK
Health Data

Population density (shading) and case locations for **Severe Wheeze** among 7–9 year old Huddersfield school children
Modeled NO\textsubscript{2} concentrations, as surrogate for all road pollution (PM\textsubscript{10}, PM\textsubscript{2.5}, SO\textsubscript{4}, CO, NO, ...) Note A\textbf{62} (SW–NE), A\textbf{629} (NW–SE), A\textbf{640} (W), A\textbf{616} (S).
Non-nested Spatial Scales

Three non-nested spatial scales: **postcode** centres, 250m **grid**, **EDs**.
Question:

- Is severe wheeze incidence within 7–9 year old school children in Huddersfield, UK associated with traffic pollution?
- Note: Could be any disease, any specified population, any spatially varying risk factor.

Goal: Analyze point count intensities regressing on spatial covariates and individual marks, all at their natural levels of aggregation, using a single class of marked point process models.
Usual Approaches Fail

**SAS:** Small Area Statistics (averaging data over EDs) don’t reflect individual risk factors: e.g., about 52% are “boys” in every ED;

**LR:** Logistic Regression doesn’t reflect spatial exposure patterns
Latent Spatial Effect, in Pictures:
**Results:**

<table>
<thead>
<tr>
<th>Risk factor</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dampness</td>
<td>8.1% (0.18)</td>
</tr>
<tr>
<td>Tobacco</td>
<td>3.5% (0.08)</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>4.4% (0.10)</td>
</tr>
<tr>
<td>Intercept</td>
<td>12.8% (0.28)</td>
</tr>
<tr>
<td>Latent term</td>
<td>71.3% (1.57)</td>
</tr>
<tr>
<td>RR Boys:Girls</td>
<td>2.96 : 1</td>
</tr>
</tbody>
</table>

**Conclusion:**

Traffic pollution doesn’t cause Severe Wheeze. Population does.
Last Example: Gamma Ray Burst Light Curves

Photo credit: NASA Goddard Space Center [via M. E. Broadbent]
GRB Pulse Number and Shape

- **Figure:** A dozen GRB photon rate time series (known as light curves)
- **Timescale:** 0.5 to 100s
- **Number of pulses:** 1 to 5 or more
- **Just one (lowest) of four energy channels**
- **Higher energy photons arrive sooner. Why?**
Norris ("Fred") Kernels

Norris Kernel: $\theta = (A, t_0, \tau_1, \tau_2)$,

$$K_N(t \mid \theta) \propto A \exp\{-\tau_1/(t - t_0) - (t - t_0)/\tau_2\}1_{\{t > t_0\}}.$$
Norris & GiG Kernels

Generalized Inverse Gaussian ("GiG") Kernel: $\theta = (A, t_0, \tau_1, \tau_2, \rho)$,

$$K_G(t \mid \theta) \propto A(t - t_0)^{p-1} \exp\left\{ -\tau_1/(t - t_0) - (t - t_0)/\tau_2 \right\} \mathbf{1}_{\{t > t_0\}}.$$
GRB 2571: Six Posterior Samples

**Figure:** Posterior samples for the mean for GRB 2571.
GRB 2571: How many pulses with $u > \epsilon$?

**Figure:** Posterior histogram for # of pulses comprising GRB 2571.
Four Energy Channels

- **Figure:** Photons are sorted into 4 energy channels, based on energy deposited (not *incident* energy, alas)
- Channel 1 is lowest energy; Channel 4 is highest
- Energy deposited is less than incident energy
- Scientific interest is in incident space.
New Challenges for The GRB Application

- **Heavy Tails** ⇒
  Switched from ID Gamma process with Lévy measure $\nu(u) \propto u^{-1} e^{-\beta u}$ to $\alpha$-Stable with $\alpha = 3/2$, and Lévy measure $\nu(u) \propto u^{-5/2}$ to match photon fluence decay rate;

- **Sticky MCMC** ⇒
  Parallel Thinning (a new variant of Parallel Tempering), exploiting ID property of LARK;

- **Overdispersion** ⇒
  NB modeling of bin counts instead of Po, exploiting Gamma mixture property.
New Challenges for The GRB Application

- **Heavy Tails** ⇒
  Switched from ID Gamma process with Lévy measure $\nu(u) \propto u^{-1} e^{-\beta u}$ to $\alpha$-Stable with $\alpha = 3/2$, and Lévy measure $\nu(u) \propto u^{-5/2}$ to match photon fluence decay rate;

- **Sticky MCMC** ⇒
  Parallel Thinning (a new variant of Parallel Tempering), exploiting ID property of LARK;

- **Overdispersion** ⇒
  NB modeling of bin counts instead of Po, exploiting Gamma mixture property.
New Challenges for The GRB Application

- **Heavy Tails** ⇒
  Switched from ID Gamma process with Lévy measure $\nu(u) \propto u^{-1} e^{-\beta u}$ to $\alpha$-Stable with $\alpha = 3/2$, and Lévy measure $\nu(u) \propto u^{-5/2}$ to match photon fluence decay rate;

- **Sticky MCMC** ⇒
  Parallel Thinning (a new variant of Parallel Tempering), exploiting ID property of LARK;

- **Overdispersion** ⇒
  NB modeling of bin counts instead of Po, exploiting Gamma mixture property.
New Challenges for The GRB Application

- **Heavy Tails** ⇒
  Switched from ID Gamma process with Lévy measure $\nu(u) \propto u^{-1} e^{-\beta u}$ to $\alpha$-Stable with $\alpha = 3/2$, and Lévy measure $\nu(u) \propto u^{-5/2}$ to match photon fluence decay rate;

- **Sticky MCMC** ⇒
  Parallel Thinning (a new variant of Parallel Tempering), exploiting ID property of LARK;

- **Overdispersion** ⇒
  NB modeling of bin counts instead of Po, exploiting Gamma mixture property.
GRB 501 Results: CIs

Figure: 95% Credible Interval for Mean, GRB 501
GRB 501 Results: PPIs

Figure: 95\% Posterior Predictive Intervals for GRB 501
Benefits

Benefits of the **LARK** Method

- Nonnegative data (like $[\text{PM}_{10}]$ and $[\text{CO}]$ concentrations) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- Unequally spaced data okay
- 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 auto-correlations
Benefits of the LARK Method

- Nonnegative data (like $[PM_{10}]$ and $[CO]$ concentrations) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- Unequally spaced data okay
- 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 auto-correlations
Benefits

Benefits of the **LARK** Method

- Nonnegative data (like $[\text{PM}_{10}]$ and $[\text{CO}]$ concentrations) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- Unequally spaced data okay
- 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 auto-correlations
Benefits (??)

**Un-Benefits of the LARK Method**

- Nonnegative data (like $[\text{PM}_{10}]$ and $[\text{CO}]$ concentrations) modeled directly, w/o transformations
- Non-stationary, non-Gaussian okay
- Unequally spaced data okay
- 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 auto-correlations
Thanks, Cornell!

LARK: Lévy Adapted Regression Kernels
A general framework for NPB function estimation
Thanks, Cornell!

**LARK:** Lévy Adapted Regression Kernels
A general framework for NPB function estimation

It’s Good to be Back!

Robert L Wolpert ’72

Lévy NPB Models & Applications
And thanks to my LARK collaborators—

- Nicky Best
- Merlise Clyde
- Leanna House
- Katja Ickstadt
- Ksenia Kyzyurova
- Danilo Lopes
- Thomas Loredo
- Zhi Ouyang
- Natesh Pillai
- Andrew Thomas
- Chong Tu
- Jianyu Wang
Memorable Math Faculty

- Jack Kieffer (freshman advisor)
- Larry Brown (first statistics course: Decision Theory)
- Roger Farrell (multivariate)
- Jacob Wolfowitz (Math. Statistics, from Cramer’s book)
- Frank Spitzer (Probability, from Chung’s book)
- Harry Kesten (Real & Complex, from Green Rudin)
- Kiyoshi Itô (Stochastic Processes)
- Anil Nerode (logic)
- * Murad Taqqu
- * Iain Johnstone
- * George Casella
More Memorable Faculty, Outside Math

- Hans Bethe (Cambridge to London train)
- Carl Sagan (wouldn’t let me take his seminar)
- David Mermin, freshman Physics (?)
- Robert Kaske, Icelandic Lit
- Carol Kaske, Divine Comedy (class met in our room)
- Walter LaFeber, History of American Foreign Relations
- Avgusta Levovna (Russian)