Multiscale Generalized Linear Models w/ Applications to Poisson Time Series

Eric D. Kolaczyk

kolaczyk@math.bu.edu

Dept of Mathematics and Statistics, Boston University

(Visiting Harvard, Spr '05)

Credit, Support, Citation

1. Joint work with Rob Nowak (UWisc).
2. Work supported by ARO, NSF, and ONR.
3. Coming this Spring as . . .

   Kolaczyk, E.D. and Nowak, R.D. (2005). Multiscale generalised linear models for nonparametric function estimation. *Biometrika*, **92,** xxx-yyy.

   (http://math.bu.edu/people/kolaczyk/pubs/ msglm_rev_final.pdf)

4. Software available at

   http://math.bu.edu/people/kolaczyk/software.html

Overview: Basic Setup

- Assume 1D, non-parametric generalized linear model (GLM)
- Goal is to characterize inhomogeneous regression function from discrete observations.

  In particular, we wish to extract information on
  - scale of local components, and
  - trend of local components.

Overview: Proposed Method

Our framework combines

- Recursive partitioning of the dataspace.

  $\Rightarrow$ *Captures scale of local components.*

- Piecewise polynomials, w/ support on partition intervals.

  $\Rightarrow$ *Captures smooth trends in local components.*

- Model selected by complexity penalized likelihood.

  $\Rightarrow$ *Insures parsimony of representation.*

Characteristics of Method

- Calculations may be performed using efficient, polynomial-time algorithms.
- Estimators of regression function have properties of near-optimality and adaptivity.

$\Rightarrow$ *Constitutes an extension of wavelet-based methods for Gaussian data to context of non-parametric GLMs.*

Wavelets and GLMs

- Posed as “challenge problem” by Bernard Silverman in his Special Invited Paper presentation at JSM 1999.

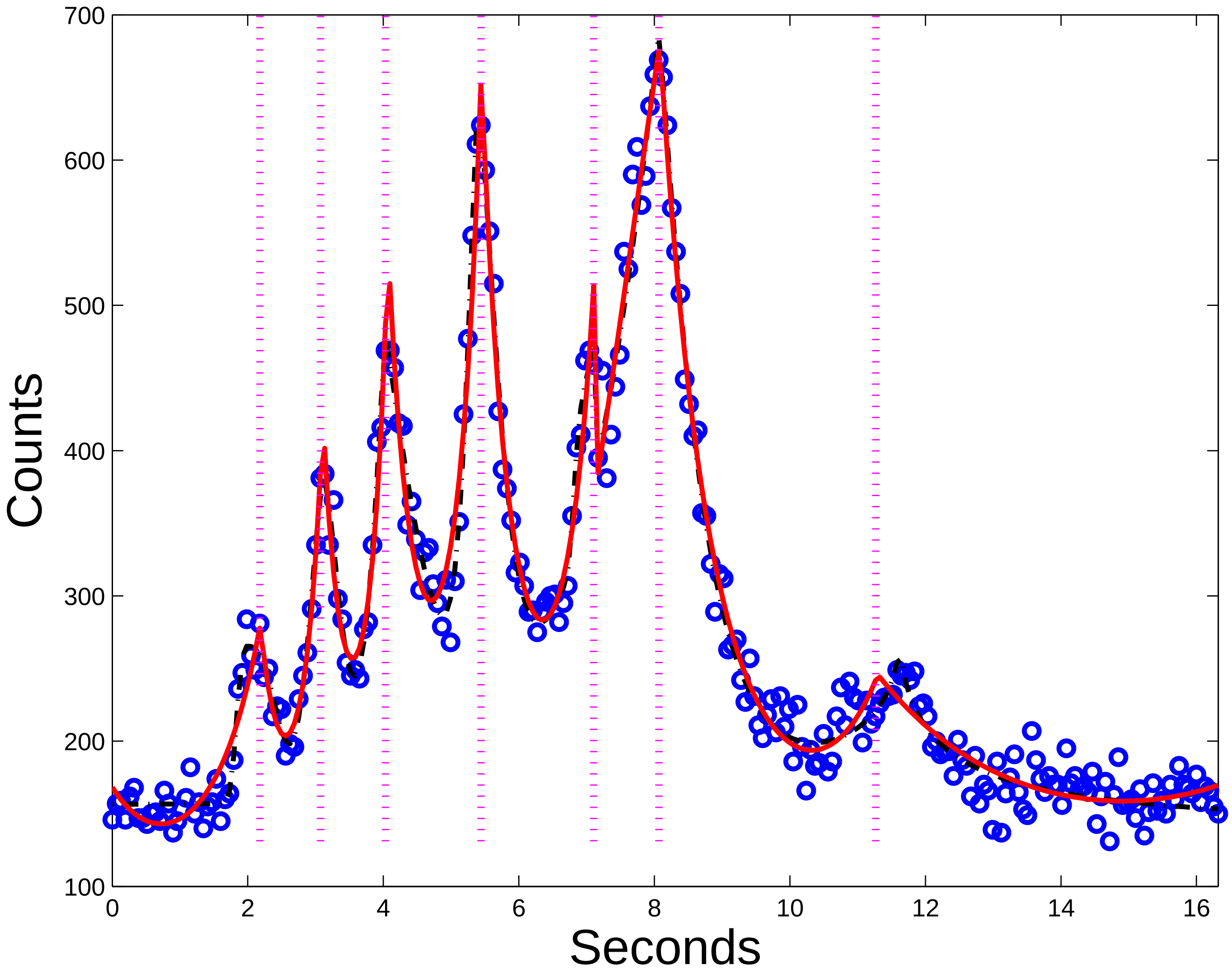
- Previous answers to this challenge consist of:
  - Antoniadis and Sapatinas (2002).

    Extension of wavelet shrinkage to NEFs with quadratic variance functions; risk theory only for Sobolev spaces.

  - Sardy, Antoniadis, and Tseng (2004).

    Wavelet reparameterization of ‘natural’ parameter; L1-penalized likelihood procedure; computationally intensive; no risk theory.

A Peek at the End Product: Gamma-Ray Burst 1425



Nonparametric GLM's.

- Independent observations $y_1, \dots, y_n$
- Each y_i has exponential family distribution

$$p_\theta(y_i) = \exp\left\{ \frac{y_i\,\theta_i - b(\theta_i)}{\tau} + c(y_i, \tau) \right\} \quad ,$$

- Natural parameters $\theta_1, \dots, \theta_n$ related through an unknown function $\theta(t) \in \Theta, t \in [0, 1]$.
- Dispersion parameter τ considered fixed and known.
- GOAL: Estimate mean vector $\mu_i \equiv G^{-1}(\theta_i)$.

Examples: Poisson and Binomial

1. Poisson:

$$\Pr(Y = y) = \frac{e^{-\mu}\mu^y}{y!} \propto \exp\{y \, \log \mu - \mu\}$$

$$\Rightarrow \quad \theta = G(\mu) \equiv \log \mu.$$

2. Binomial

$$\begin{aligned} \Pr(Y = y) &= \binom{m}{y} \left(\frac{\mu}{m}\right)^y \left(\frac{m-\mu}{m}\right)^{m-y} \\ &\propto \exp\{y \log[\mu/(m-\mu)] - m \log(m-\mu)\} \end{aligned}$$

$$\Rightarrow \quad \theta = G(\mu) \equiv \log[\mu/(m-\mu)].$$

Basic Model Class: RDP's and PP's.

- Let $\mathcal{P}^*_{Dy}$ be a complete recursive dyadic partition (C-RDP) of the interval $(0,1]$, composed of n equi-length intervals I_i.

$$[0,1) \;\rightarrow\; \{\,(0,0.5]\,,\,(0.5,1]\,\} \;\rightarrow\; \cdots \;\rightarrow\; \mathcal{P}^*_{Dy} = \{I_i\}_{i=1}^n$$

- Let $\mathcal{P} \preceq \mathcal{P}^*_{Dy}$ denote an intermediate, recursive dyadic partition (RDP) encountered between $(0,1]$ and $\mathcal{P}^*_{Dy}$.

- Model $\theta(\cdot)$ by members of the class

$$PP(\mathcal{P};D) \equiv \{\,\text{Piece-wise polynomials, of order D, with components supported on } I \in \mathcal{P}\,\}\;\;.$$

Example: Piecewise exponential models

Consider $y_1, \ldots, y_n$ a Poisson time series. If for $t_i \in I \in \mathcal{P}$

$$\theta_i = \log(\mu_i) = \sum_{k=1}^{K} \alpha_k (t_i - t_I)^k$$

Then

$$\Rightarrow \mu_i \approx \exp\{\alpha_K (t_i - t_I)^K)\}$$

Result: A class of piecewise exponential models.

Compare

- Norris *et al.* (1996)
- Connors (2003)

Model Selection

- Let

$$\ell(\theta) \equiv \log p_\theta(\mathbf{y}) \equiv \sum_{i=1}^{n} \log p_\theta(y_i) \quad ,$$

$$\#(\mathcal{P}) \equiv \text{Number of intervals } I \in \mathcal{P} \quad ,$$

and

$$\lambda = \lambda(D; n) \equiv (D/2) \log(n) \quad .$$

Estimate θ by the complexity-penalized likelihood estimator

$$\hat{\theta}_{RDP} \equiv \arg \max_{\mathcal{P} \preceq \mathcal{P}^*_{Dy}} \max_{\theta' \in PP(\mathcal{P};D)} \{ \ell(\theta') - 2\lambda \#(\mathcal{P}) \} \quad .$$

What's up with the penalty?

The penalty

$$\mathrm{pen}(\mathcal{P}) = (D/2)\log(n) \ \times \ \#(\mathcal{P})$$

is chosen to satisfy the condition

$$\sum_{\mathcal{P} \preceq \mathcal{P}^*_{Dy}} \exp\{-\mathrm{pen}(\mathcal{P})\} \leq 1 \ \ .$$

Derives from role in underlying risk theory; connections with coding theory; essentially an unnormalized prior.

Implementation

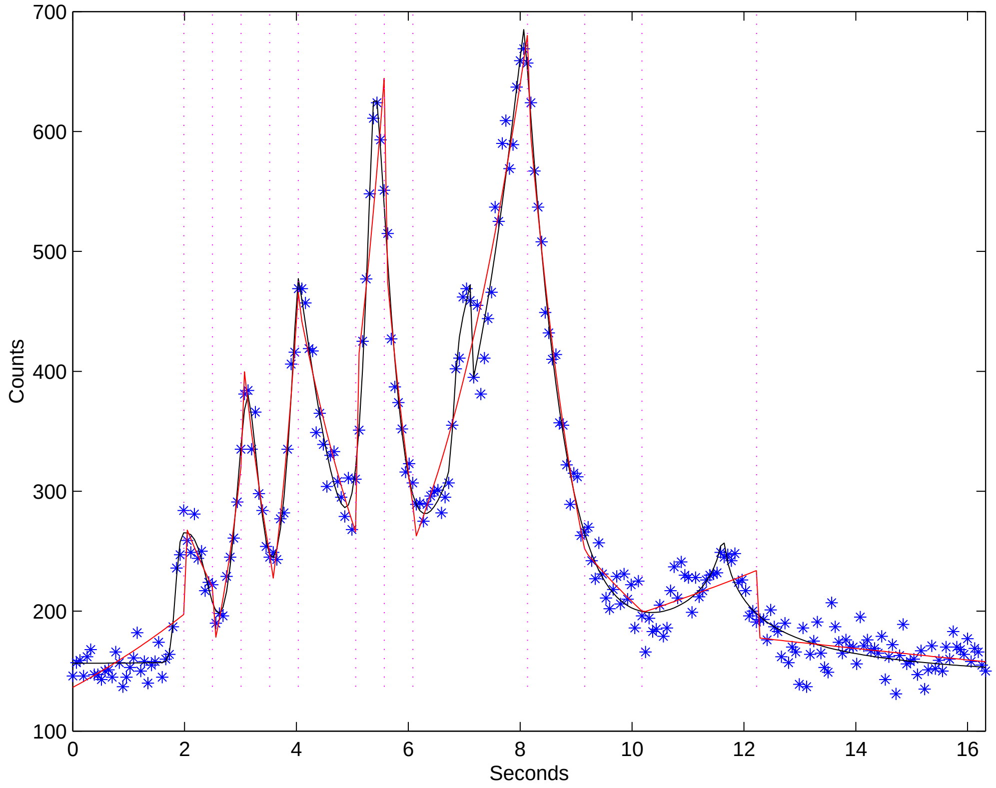
- For a given candidate partition $\mathcal{P}$, and fixed $I \in \mathcal{P}$, the polynomial piece of $\hat{\theta}_{RDP}$ can be fit using standard GLM software.

- Partitions $\mathcal{P} \preceq \mathcal{P}^*$ possess certain
  - redundancies
  - hereditary properties

  that enable $\hat{\theta}_{RDP}$ to be computed using an $O(n)$ bottom-up, optimal tree-pruning algorithm (e.g., as in CART).
- Run time for $n = 256$ on 3.2GHz machine: ~ 2 seconds.

Illustration: $\hat{\theta}_{RDP}$ w/ GRB's

- $\theta(t) \approx \log(\mu(t))$ piece-wise linear in t.
- Reasonably good match between estimated regression function (red) and fitted FRED model (black).



Extension to Arbitrary C-RP's

- Let $\mathcal{L}$ be the library of all $(n-1)!$ possible complete recursive partitions (C-RP) $\mathcal{P}^* = \{I_i\}_{i=1}^n$.

- Define $PP(\mathcal{P}; D)$ as before, for all $\mathcal{P} \preceq \mathcal{P}^*$, and all $\mathcal{P}^* \in \mathcal{L}$.

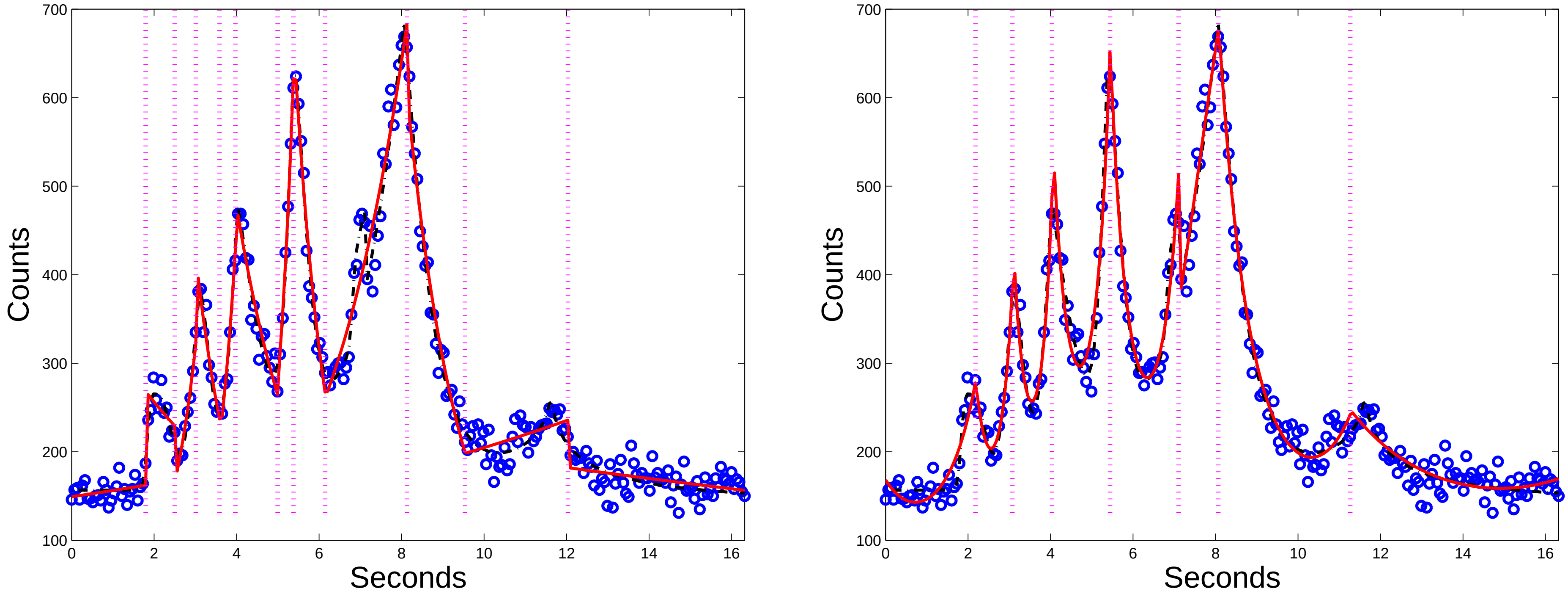
- Estimate θ by the complexity-penalized estimator

$$\hat{\theta}_{RP} \equiv \arg \max_{\mathcal{P}^* \in \mathcal{L}} \max_{\mathcal{P} \preceq \mathcal{P}^*} \max_{\theta' \in PP(\mathcal{P};D)} \{\, \ell(\theta') - 2\lambda \, \#(\mathcal{P}) \,\}$$

Practical Implications

- End result is piecewise polynomial fit with segmentation at any of the $n-1$ interior points.
- Estimator is multiscale, in that there are no restrictions on extent of segments.
- Redundancies and hereditary properties among C-RPs may be exploited to calculate $\hat{\theta}_{RP}$ in $O(n^3)$ steps.
- Run time for $n = 256$ on 3.2GHz machine: ~ 4 minutes.

Illustration: GRBs

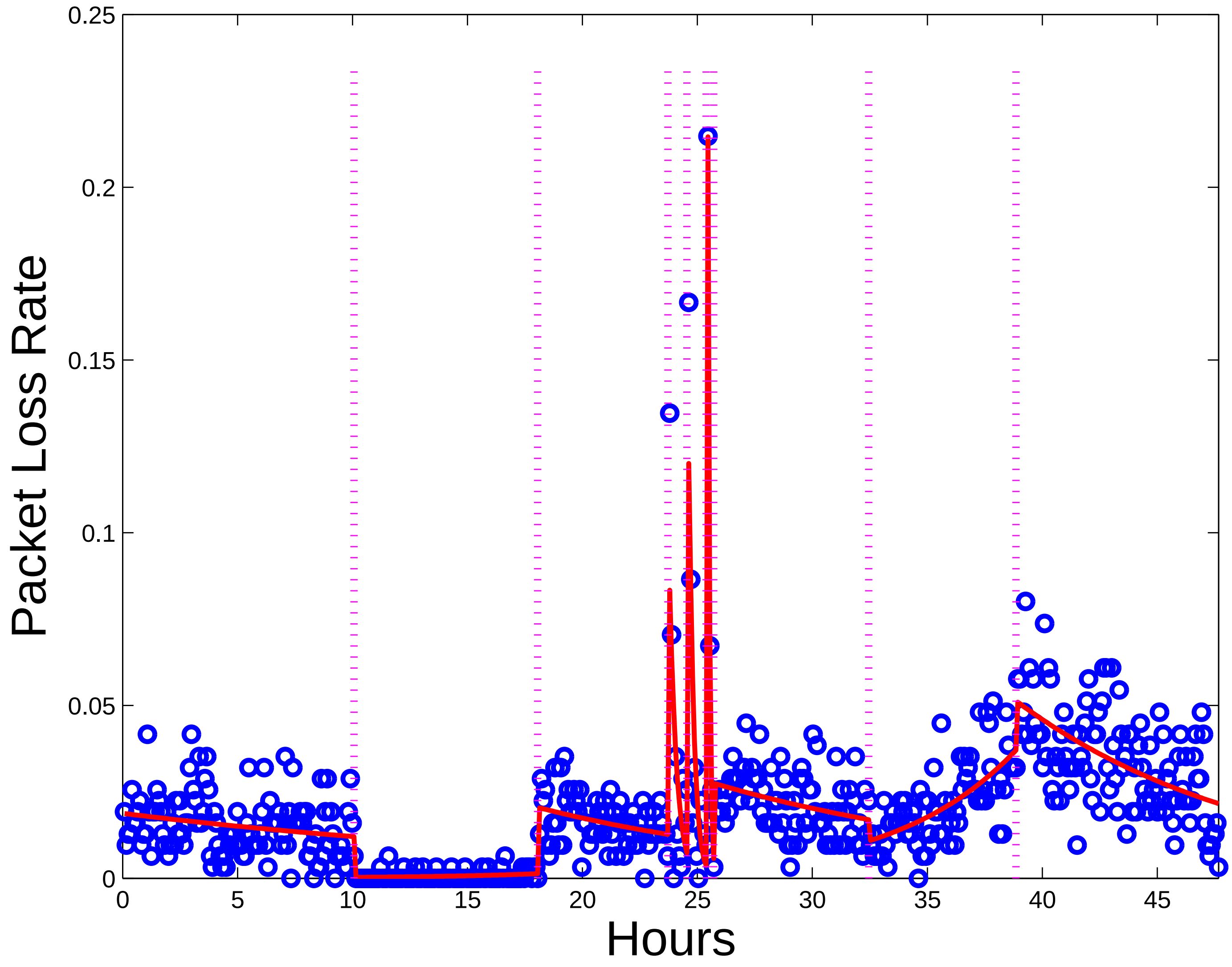


[Left: Linear; Right: Quadratic]

Illustration: Packet Loss Data.

- Packets transmitted every 160ms from UMass-Amherst to Sweden.
- Interest in estimating packet loss rates.
- $0/1$ loss data subsampled at 1000ms, and binned over 5 minute intervals.
- Modeled as binomial time series.
- $\theta(t) \approx \log[p(t)/(1 - p(t))]$ modeled as piecewise linear in t.

Packet Loss Data: Results



Risk Theory.

Define loss in estimation through (squared) Hellinger distance i.e.,

$$H_n^2\left(p_\theta, p_{\hat{\theta}}\right) = \int \left\{ \sqrt{p_\theta(\mathbf{y})} - \sqrt{p_{\hat{\theta}}(\mathbf{y})} \right\} \nu_n(\mathbf{y}) \ .$$

and measure risk as
$R_n \equiv E\left[H_n^2\left(p_\theta, p_{\hat{\theta}}\right)\right]$.

Let $B^\alpha_{p,q}$ be a Besov space with parameters $0 < \alpha < D$ and $1 \leq p < \infty$ such that $1/p < \alpha + 1/2$, and $q > 0$.

Risk Theory (cont)

Theorem 1 *Suppose*

1. $f \in B^{\alpha}_{p,q}([0,1])$,
2. $|f(t)| \leq C$ *for all* $t \in [0,1]$, *for* $C > 0$
3. G *and* G^{-1} *are Lipschitz.*

Then the risk of our estimators behaves like

$$R_n \sim (\log^c n/n)^{2\alpha/(2\alpha+1)} \quad ,$$

where $c = 2$ *for RDP and* $c = 1$ *for RP.*

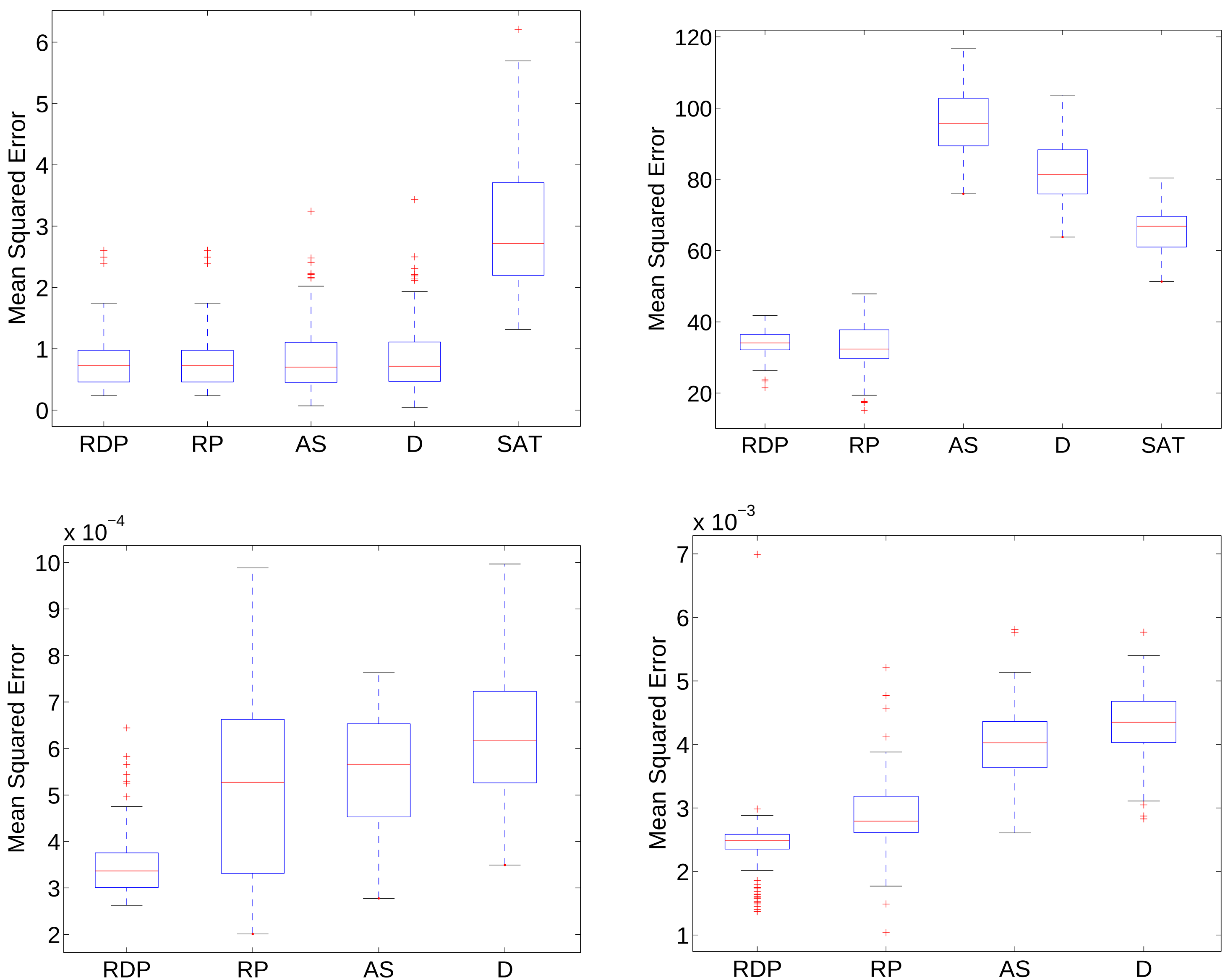
And that's important because . . .

- Optimal rates are $O\left(n^{-2\alpha/(2\alpha+1)}\right)$
- Classical wavelet-based estimators have the same properties of *near-optimality* and *adaptivity* in the standard 'signal plus noise' models.
- *Simplicity* and *performance* of our method derive simultaneously from use of piecewise polynomial system with *same* approximation theoretic properties as orthonormal wavelet systems.
- Competing methods fail to achieve either or both.

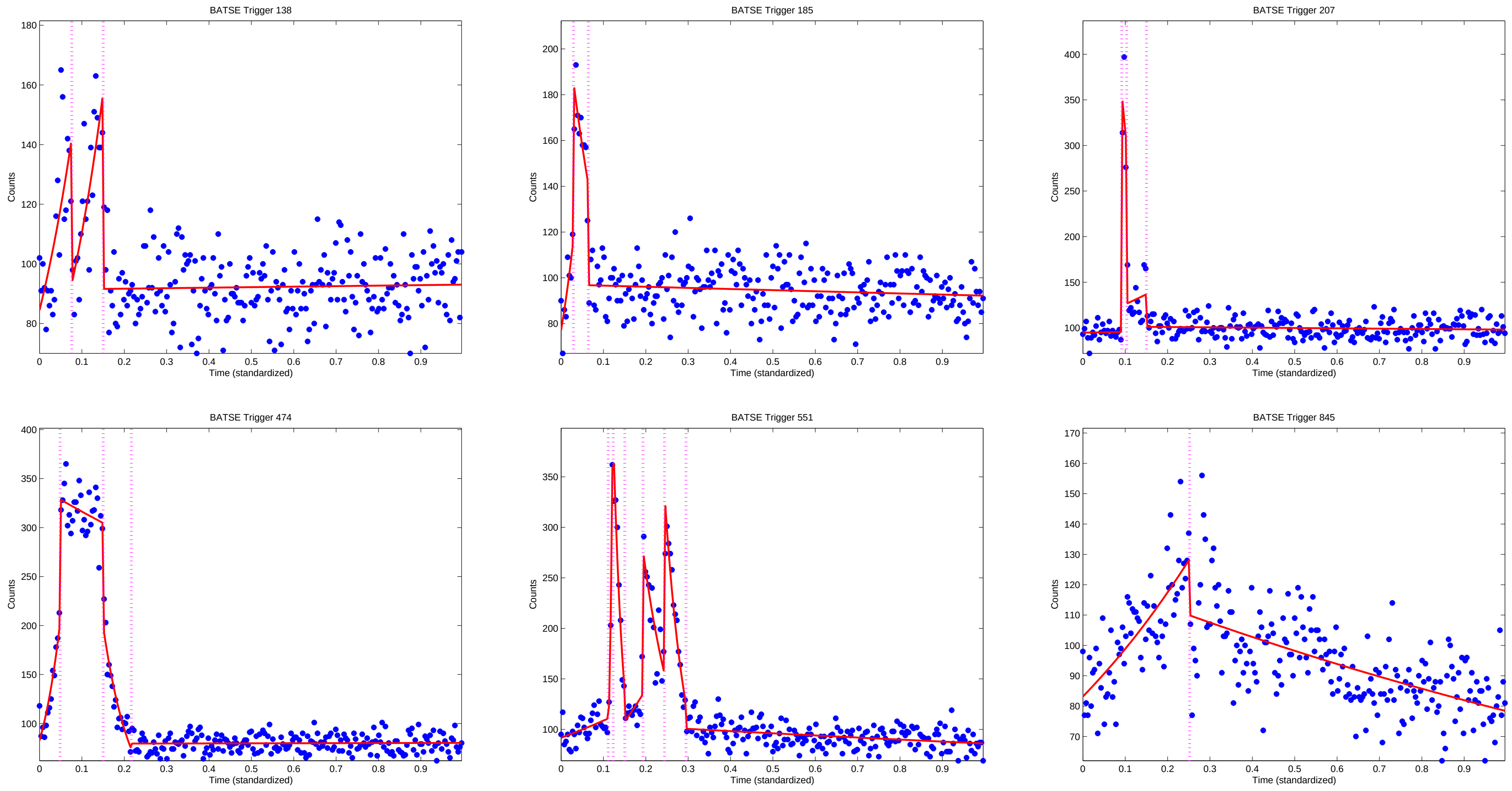
Simulation

- Simulated ‘smooth’ and ‘burst-like’ functions, using ‘medium’ SNR, for each of Poisson and binomial models.
- $M = 100$ trials for each case.
- Signals of length $n = 256$
- Compared RDP and RP methods to methods of
  - Antoniadis and Sapatinas (AS)
  - Donoho (D)
  - Sardy, Antoniadis, and Tseng (AST)

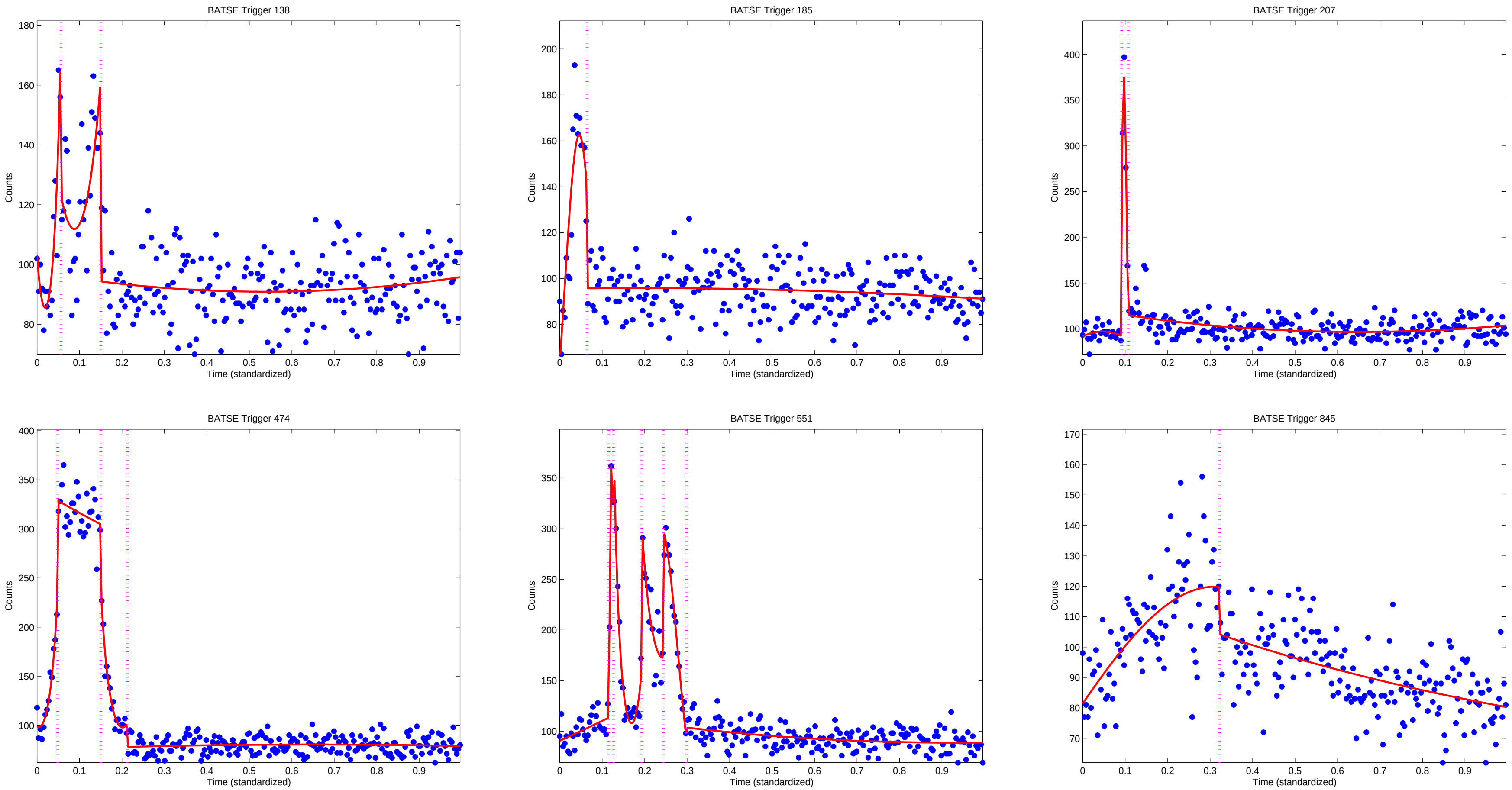
Simulation Results



Another Look at GRBs (linear)



Another Look at GRBs (quadratic)



Extensions

- Deconvolution: EMS or EM-MAP straightforward.
- Images: ‘Platelets’ of Willett & Nowak.
- Variable Degree: Easy (in practice; theory hard)
- Beyond GLMs:
  - Multiscale, multigranular image segmentation. (Kolaczyk, Ju, and Gopal)
  - Segmentation of binary genomic signals (?) (joint work Kasif and Lee)

Final Comments.

- Assumption of GLMs only necessary for underlying risk theory.
- Can show wavelet-like risk properties for estimators because
  1. general bound on Hellinger risk, due to extension of recent work of Li and Barron;
  2. piece-wise polynomials and orthonormal wavelet bases have the same approximation properties.
- Too much emphasis put on “smooth” estimators?
  (If desired, discontinuities can be smoothed without loss of properties using moment-interpolating techniques, like those of Donoho *et al.*)

Final Comments (cont.)

- Uncertainty bands would be nice!
  - Asymptotic confidence bands a possibility
    (Likely hard . . . need to extend work of Genovese & Wasserman, or Baraud.)
  - Bootstrapping possible (e.g., Young 2003)
  - Massage prior into proper form and use MCMC?