Multiscale Generalized Linear Models for Nonparametric Function Estimation

Eric D. Kolaczyk

Department of Mathematics and Statistics

Boston University, Boston, MA, 02215

kolaczyk@math.bu.edu

Robert D. Nowak

Department of Electrical and Computer Engineering

Rice University, Houston, TX, 77005-1892

nowak@rice.edu

February, 2003

Abstract

We present a method for extracting information on both the scale and trend of local components of an inhomogeneous function in a nonparametric generalized linear model (NP-GLM). Our multiscale framework combines recursive partitions, which allow for the incorporation of scale in a natural manner, with systems of piecewise polynomials supported on the partition intervals, which serve to summarize the smooth trend within each interval. Our estimators are formulated as solutions of complexity-penalized likelihood optimizations, where the penalty seeks to limit the number of intervals used to model the data. The actual calculation of the estimators may be accomplished using standard GLM software routines, within the context of efficient, tree-based, polynomial-time algorithms. A risk analysis shows that these estimators achieve the same rates in the NP-GLM model as the classical wavelet-based estimators in the Gaussian "function plus noise" model, for suitably defined ranges of Besov spaces. Illustrations of the method to gamma-ray burst data in astronomy and packet loss data in computer network traffic analysis confirm its practical relevance.

Key words: Astronomy, computer network traffic, minimax, piecewise polynomial, recursive partitioning, wavelets.

1 Introduction

Nonparametric curve estimation (e.g., Efromovich 1999) has become a fundamental part of data analysis, arising in applications that range across the spectrum of scientific disciplines. And multiscale methods, particularly those based on wavelets and related paradigms, have become a leading choice when the curve to be estimated is inhomogeneous in nature. A large proportion of these methods assume some variant of the traditional "function plus noise" model i.e., $y_i = \theta_i + z_i$, where the θ_i are samples of some unknown function f and the z_i are an additive noise, often independent and identically distributed Gaussian. The goal then is to produce an accurate estimate of $\theta = (\theta_1, \ldots, \theta_n)$, particularly in a manner that does not obscure the inhomogeneous structure in f (e.g., jumps, cusps, etc.) in the process.

In this paper we are interested in problem contexts that differ from that just described in two ways. First, a "function plus noise" model may not be tenable – the data may take the form of counts or proportions, for example. Second, the function f arguably may be itself a composite of simpler, localized homogeneous functions, and accurate indication of both the scale (extent) and overall trend of each component is felt to be useful in addressing the scientific questions accompanying the data.

In Section 4, two examples of such situations are examined. The first example concerns photon arrival-time data from an astronomical gamma-ray burst, modeled as a realization of a Poisson counting process, such as that shown in Figure 1(a). Accurate characterization of the number, location, and extent of component pulses underlying such bursts is a fundamental part of analyzing such data in the field of high-energy astrophysics (e.g., Norris *et al.* 1996). The second example concerns the monitoring of computer network traffic and, more specifically, the loss ('dropping') of packets in transmitting information between two locations, such as shown in Figure 1(b). Packet loss data takes the form of a Bernoulli process which, with appropriate sub-sampling and aggregation, may be modeled as a binomial time series. An understanding of the underlying characteristics and patterns in packet loss rates is fundamental to successful network monitoring and maintenance, particularly where multimedia applications are concerned

(e.g., Yajnik *et al.* 1999).

Motivated by problems such as these, we present here a method for extracting information on both the scale and trend of local components of an inhomogeneous function in a nonparametric generalized linear model (NP-GLM). Our method combines recursive partitions, which allow for the incorporation of scale in a natural manner, with systems of piecewise polynomials supported on the partition intervals, which serve to summarize the smooth trend within each interval. Our estimators are formulated as solutions of complexity-penalized likelihood optimizations, where the penalty seeks to limit the number of intervals used to model the data. The actual calculation of the estimators may be accomplished using standard GLM software routines, within the context of efficient, tree-based, polynomial-time algorithms. A risk analysis, based on squared Hellinger loss, shows that these estimators achieve the same rates in the NP-GLM model as the classical wavelet-based estimators in the Gaussian "function plus noise" model, for suitably defined ranges of Besov spaces.

At the time of this writing, we currently are aware only of a handful of related work in this direction. For example, Sardy *et al.* (2002) introduce a class of wavelet-based complexity-penalized likelihood estimators for NP-GLM's, where the penalty is on the ℓ_1 norm of the wavelet coefficients. However, the optimization involved in calculating these estimators is expensive, requiring the use of interior-point methods for the solution of the corresponding Kuhn-Tucker equations. Additionally, there is the work of Antoniadis and Sapatinas (2001) and Antoniadis *et al.* (2001), in which, adopting a mean-squared error criterion, the wavelet shrinkage paradigm is extended to natural exponential families. In this case the calculations are computationally efficient, and optimal risk rates are proven, although for estimation in the more classical setting of Sobolev spaces. Additionally, the results pertain only for natural exponential families with quadratic or cubic variance functions. These do, however, include the most widely used cases of Gaussian, Poisson, and binomial models, for example.

Organization of this paper is as follows. In Section 2 we present the basic framework of our class of multiscale generalized linear models (MS-GLM) and define certain estimators. Properties of these estimators, relating to algorithmic complexity and risk rates, are then detailed

in Section 3. In Section 4 we present background and results of the analysis of the gamma-ray burst and packet loss data mentioned above. Finally, some closing remarks are gathered in Section 5. Proofs of results stated in the body of the paper may be found in the appendix.

2 Multiscale Generalized Linear Models.

Let $y = (y_1, \ldots, y_n)$ be independent observations from a one-dimensional nonparametric generalized linear model, where the y_i share a common natural exponential family (NEF) form i.e.,

$$p_\theta(y_i) = \exp\left\{ \frac{y_i \cdot \theta_i - b[\theta_i]}{\tau} + c(y_i, \tau) \right\} \quad , \tag{1}$$

up to the values θ_i. The elements of the vector $\theta = (\theta_1, \ldots, \theta_n)$ will be assumed to be related through equi-spaced, average sampling of an unknown function f i.e., $\theta_i \equiv n \int_{I_i} f(t)dt$ and $I_i \equiv ((i-1)/n, i/n]$, a common model for many instrumental measurement devices. The function $f(\cdot)$ will be assumed to be a member of some appropriately defined class of functions $\mathcal{F}([0,1])$ with support on the unit interval. The dispersion parameter τ will be taken to be fixed and known. Interest will focus on the estimation of θ, from which estimates of the mean $\mu = (\mu_1, \ldots, \mu_n)$ can be obtained, if desired, using the inverse $G^{-1} \equiv db/d\theta$ of the canonical link function $G(\cdot)$ i.e., $\mu_i = G^{-1}(\theta_i)$.

Now suppose that $\mathcal{F}([0,1])$ is a class of inhomogeneous functions, such as a Besov space. In this case the approximation theoretic properties of wavelets recommend the development of an appropriate wavelet-based method for estimating θ. However, their beauty and elegance aside, orthonormal wavelet bases can be said to consist essentially of two key elements:

1. a nested hierarchy of recursive, dyadic partitions;

2. an exact, efficient representation of polynomial smoothness.

The first element allows one to describe isolated singularities in a concise manner i.e., with roughly $\log_2(n)$ wavelet coefficients, where singularities may be considered structure in the function beyond a certain degree of polynomial smoothness. The second element then states

that this remaining smoothness is represented concisely as well i.e., through a handful of so-called scale coefficients.

In light of these observations, we propose an approach to modeling θ (and hence μ) that uses these two elements explicitly, through the use of piecewise polynomials, as opposed to using them implicitly through an approach based on wavelets. This decision yields estimators $\hat{\theta}$ of θ (or $\hat{\mu} \equiv \mu_{\hat{\theta}}$ of μ) that are simple to define, easily interpreted, and produced by efficient computational algorithms. Yet they also attain the type of minimax risk rates for which wavelet-based methods have come to be known. We first introduce a basic version of our framework, based on the use of a single hierarchy of partitions, and then describe an extension of that method using a full library of such hierarchies. Additionally, we discuss how moment-interpolating refinement (Donoho *et al.* 2000) may be used as a simple post-processing procedure for smoothing out the discontinuities in our piecewise estimators, when this is felt desirable, without jeopardizing the relevant theoretical properties.

2.1 *The Basic Method: Recursive Dyadic Partitioning.*

Consider the class of recursive dyadic partitions (RDP) of the unit interval $(0, 1]$. Assume that $n = 2^J$ is a power of two (this assumption will be relaxed in the next section). Beginning with the unit interval, we picture partitioning that in a recursive fashion, each time splitting some previously resulting interval exactly in half. This process is repeated until the complete recursive dyadic partition (C-RDP) $\mathcal{P}^*_{Dy} \equiv \{I_i\}_{i=1}^n$ is achieved. Partitions encountered intermediate to $(0, 1]$ and $\mathcal{P}^*_{Dy}$ will be denoted $\mathcal{P}$, and the partial ordering induced by the operation of successive refinement will be expressed using the symbols '$\prec$' and '$\preceq$' (e.g., $\mathcal{P} \prec \mathcal{P}^*_{Dy}$).

Next, for a given partition $\mathcal{P} \preceq \mathcal{P}^*_{Dy}$, define $PP(\mathcal{P}; D)$ to be the class of piecewise-polynomial functions of order D on $(0, 1]$, such that the component pieces of these functions are restricted in number and support in one-to-one correspondence with the intervals $I \in \mathcal{P}$. For example, if $\mathcal{P} \equiv \{\, (0, 0.5]\,,\ (0.5, 1]\,\}$, and $D = 2$, then $PP(\mathcal{P}; D)$ is the set of all two-component piecewise-linear functions on $(0, 1]$, with the first component restricted to the sub-interval $(0, 0.5]$, and the second, to the sub-interval $(0.5, 1]$.

We will construct an estimate for θ by choosing, based on data y, some optimal member of the classes $PP(\mathcal{P}; D)$. Specifically, let $\ell(\theta) \equiv \sum_{i=1}^{n} \log p_\theta(y_i)$ be the log-likelihood and let $\#(\mathcal{P})$ be the number of intervals $I \in \mathcal{P}$. Then define

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

where $\lambda \equiv \lambda(n; D)$ is a smoothing parameter to be defined (see Section 3.2). This is a maximum complexity-penalized likelihood estimator of θ, where the penalty attempts to control how many polynomial components are used. The optimization is over all piecewise-polynomial fits on a given partition $\mathcal{P}$, over all partitions $\mathcal{P} \preceq \mathcal{P}^*_{Dy}$. Without any penalty, the optimal choice of partition would be $\hat{\mathcal{P}} = \mathcal{P}^*_{Dy}$, with n components, which simply returns the non-parametric MLE. In the case of $D = 1$, where piecewise-constant functions are used, the estimator $\hat{\theta}_{RDP}$ is equivalent to a type of constrained hard-thresholding with Haar wavelets (arguing as in Donoho 1997, for example), and is not unlike the class of likelihood-based, piecewise constant estimators studied in Kolaczyk and Nowak (2002).

Note that contained in $\hat{\theta}_{RDP}$ is information on both the scale (or extent) of localized components in θ, due to the implicit selection of an optimal partition $\hat{\mathcal{P}}$, and the smooth trend associated with each component. The estimator $\hat{\theta}_{RDP}$ can be calculated using a tree-based algorithm of $O(n)$ complexity, and can be shown (with appropriate choice of λ) to possess properties of near-optimality and adaptivity in terms of minimax risk. Details will be provided in Section 3.

2.2 *Extension of the Basic Method: Library of Partitions.*

Despite the advantages that accompany it, the choice of a single fixed hierarchy of (dyadic) partitions underlying $\hat{\theta}_{RDP}$ may be seen as a limitation in some contexts. For example, when it is hypothesized that separation of localized components in θ occurs at a handful of potentially non-dyadic positions i.e., changepoints, and their location and number has a meaningful interpretation, such as in the astronomy example in Section 4.1.

It is useful, therefore, to eliminate the dyadic constraints in the definition of recursive partitioning adopted above. Let n be an arbitrary positive integer i.e., not necessarily a power of 2. Consider the process whereby the unit interval again is partitioned in a recursive fashion, but now split points are constrained simply to the endpoints of the intervals I_i. Beginning with the trivial partition $(0,1]$, we split that into two pieces at one of the points $\{i/n\}_{i=1}^{n-1}$. Then, proceeding in a recursive fashion, given a partition $\mathcal{P} \prec \mathcal{P}^*$, where $\mathcal{P}^* = \{I_i\}_{i=1}^{n}$, we refine $\mathcal{P}$ by splitting one and only one of the intervals $I \in \mathcal{P}$ at one of the remaining allowable points. That is, we split at one of the unused points in the intersection of $\{i/n\}_{i=1}^{n-1}$ and the interior of I. We will call the interval I in such cases the "parent" interval and the two corresponding sub-intervals, say $I_{ch(I),l}$ and $I_{ch(I),r}$, the left and right "children" intervals. The final partition $\mathcal{P}^*$ resulting from this process will be called a complete recursive partition (C-RP).

While the basic form of estimator defined in (2) for the C-RDP $\mathcal{P}^*_{Dy}$ generalizes immediately to arbitrary fixed C-RP's $\mathcal{P}^*$, such extensions will potentially suffer from an analogous lack of flexibility in placement of changepoints. Instead, a more promising alternative involves the use of a library of C-RP's. That is, let $\mathcal{L}$ be the library of all $(n-1)!$ possible C-RP's $\mathcal{P}^*$, and define $PP(\mathcal{P}; D)$ as before, for all $\mathcal{P} \preceq \mathcal{P}^*$ and $\mathcal{P}^* \in \mathcal{L}$. Then, generalizing (2), we define

$$\hat{\theta}_{RP} \equiv \arg \max_{\mathcal{P}^* \in \mathcal{L}} \max_{\mathcal{P} \preceq \mathcal{P}^*} \max_{\theta' \in PP(\mathcal{P};D)} \{ \ell(\theta') - 2\lambda \#(\mathcal{P}) \} \quad . \tag{3}$$

This too is a maximum complexity-penalized likelihood estimator, but now defined over a much richer model space. In particular, it includes all possible partitions of $(0,1]$ into m sub-intervals, for $m = 0, 1, \ldots, n$, where the sub-intervals are disjoint unions of the I_i. As we have defined it, the model space appears to be much larger than this, but in fact it possesses a good deal of redundancy, in that a given partition $\mathcal{P}$ is likely to be shared by a number of C-RP's $\mathcal{P}^*$. This redundancy is due to our choosing to continue to cast our framework within the context of recursive partitioning, which in turn allows us to develop an extension of the tree-based algorithm underlying calculation of (2) for calculating (3). In effect, we are able to search a "forest" of trees corresponding to C-RP's $\mathcal{P}^* \in \mathcal{L}$ using an algorithm of only $O(n^3)$

complexity. And, not surprisingly in light of the properties of $\hat{\theta}_{RDP}$, the estimator $\hat{\theta}_{RP}$ enjoys similar near-optimality and adaptivity properties.

2.3 *Post-processing Using Moment-Interpolating Refinement.*

By construction our estimators will be discontinuous (unless a single polynomial is chosen to fit the full data). Although in the applications we present in this paper this characteristic lends critical information to the overall analysis, in other contexts it may be considered undesirable. One is led automatically then to the question of smoothing these estimators, in what is essentially a post-processing step. Of course, choice of method is important if one is to maintain the algorithmic and theoretical advantages inherent in the original estimators.

Consider the case of the estimator $\hat{\theta}_{RDP}$. Applying, say, a kernel smoother to this estimator takes an object produced by a multiscale method and forces it through a single-scale method, which is bound to defeat the advantages of the former. Instead, a natural alternative is to use the moment-interpolating refinement framework of Donoho *et al.* (2000). Specifically, an L_2-optimal piecewise-polynomial approximation on $\mathcal{P}^*_{Dy}$ of a function g can be linked in one-to-one correspondence with a truncated, hereditary expansion of g in a particular type of wavelet basis, the Alpert basis (e.g., Alpert 1993). Donoho *et al.* introduce a dual basis of so-called moment-interpolating wavelets, whose coefficients may be computed from those of the Alpert expansion and then used in turn to compute an alternative approximation of g that is in fact smooth and yet possesses the same order of approximation as the original. From this result it can be shown that the risk properties accompanying $\hat{\theta}_{RDP}$ are maintained. And from a computational perspective the entire process is equivalent to an $O(n)$ linear filtering of the original estimator.

3 Properties of MS-GLM Estimators.

In this section we provide a formal statement of the previously-mentioned algorithmic and risk properties of the estimators $\hat{\theta}_{RDP}$ and $\hat{\theta}_{RP}$.

3.1 *Algorithmic Properties.*

The estimators $\hat{\theta}_{RDP}$ and $\hat{\theta}_{RP}$ are simply piecewise polynomial functions of order D, with the component pieces of these functions supported on the intervals I in some optimally selected partition $\hat{\mathcal{P}}$, and the coefficients of each polynomial component chosen according to a standard maximum likelihood criterion. Given an interval I in some candidate partition $\mathcal{P}$, an optimal order-D polynomial may be fit to the subset of observations $\{y_{1_I}, \ldots, y_{n_I}\}$ in y corresponding to I, using a standard GLM fitting routine. Therefore, the primary algorithmic hurdle faced in calculating our estimators is that of comparing fits over all allowable partitions $\mathcal{P}$ without performing the maximum likelihood calculations on any given interval I more than once. The fitting and comparisons in the case of both estimators may be done in a number of steps that scales in a polynomial fashion with the sample size n, as summarized in the following theorem.

Theorem 1

i. The estimator $\hat{\theta}_{RDP}$ may be calculated using $O(n)$ calls to a GLM fitting routine and $O(n)$ comparisons of the resulting (penalized) likelihood values;

ii. the estimator $\hat{\theta}_{RP}$ may be calculated using $O(n^2)$ calls to a GLM fitting routine and $O(n^3)$ comparisons of the resulting (penalized) likelihood values.

Proof of these two statements involves tree-based arguments of a dynamic-programming flavor. In the case of $\hat{\theta}_{RDP}$, the argument is standard and parallels, for example, those found in Donoho (1997). Specifically, the framework of a bottom-up optimal tree-pruning algorithm, like that underlying CART (Breiman *et al.* 1983), may be used. In the case of the estimator $\hat{\theta}_{RP}$, the statement follows by an extension of this reasoning to what may be pictured as an inter-woven forest of trees. The justification follows an argument similar to that in Kolaczyk and Nowak (2002). Details may be found in the appendix. As an aside, it should be noted that the need for $O(n^3)$ comparisons in computing $\hat{\theta}_{RP}$ is not unexpected, given that similar algorithms of the same complexity have been offered for methods addressing, for example, the multiple changepoint problem (Barry and Hartigan 1993), approximation and compression of

digital signals (Prandoni and Vetterli 1999), and the segmentation of haplotype data in genetics (Zhang *et al.* 2002).

3.2 *Risk Properties.*

We now address the risk properties of our estimators. Define the loss in estimating θ by a value $\hat{\theta}$ through the squared Hellinger distance i.e.,

$$H_n^2\left(p_\theta, p_{\hat{\theta}}\right) = \int \left\{ \sqrt{p_\theta(y)} - \sqrt{p_{\hat{\theta}}(y)} \right\}^2 \nu_n(y) \ , \tag{4}$$

where $p_\theta(y) \equiv \prod_i p_\theta(y_i)$ and $\nu_n(y)$ is the dominating measure. Let $R_n \equiv (1/n)E\left[H_n^2\left(p_\theta, p_{\hat{\theta}}\right)\right]$ denote the corresponding risk.

To capture the notion of a one-dimensional, inhomogeneous function we will assume that the function f underlying the vector θ is a member of some Besov space $B_{p,q}^\alpha$, for some appropriately defined range of parameters α, p, and q. For f to be in $B_{p,q}^\alpha$ means essentially that f must have α derivatives in L_p; the parameter q is secondary in its role, allowing for additional fine tuning of the definition of the space. See Donoho *et al.* (1995) for an overview, for example, or DeVore (1998) for an accessible, detailed introduction to this and related topics.

The following may be said concerning the two estimators defined in Section 2.

Theorem 2 *Let $B_{p,q}^\alpha$ be a Besov space, for $0 < \alpha < D$ and $1 \leq p < \infty$ such that $1/p < \alpha + 1/2$, and $q > 0$. Suppose $f \in B_{p,q}^\alpha([0,1])$, where $|f(t)| \leq C$, for all $t \in [0,1]$, for $C > 0$, and assume that G^{-1} is Lipschitz on $[-C, C]$. Let $\theta = (\theta_1, \ldots, \theta_n)$ be derived from f through average sampling i.e., $\theta_i = n \int_{I_i} f(t)dt$.*

Then for $\lambda \equiv \lambda(n; D) = (1 + D/2)\log n$, with $n \geq 3$,

$$R_n \leq O\left((\log^c n/n)^{2\alpha/(2\alpha+1)} \right) \ ,$$

where $c = 2$ for $\hat{\theta}_{RDP}$ and $c = 1$ for $\hat{\theta}_{RP}$.

A few comments are in order. First, the risk rates $\{2\alpha/(2\alpha+1)\}$ are the standard optimal

rates for this type of problem. Since the risks R_n differ from these only by a power of $\log n$, our estimators have near-optimal risk rates. The bound for the RDP estimator is slightly larger than that for the RP estimator, but this likely could be improved through the use of more involved arguments than those we employ in the appendix, at the cost of simplicity. Second, neither of our estimators is provided any *a priori* knowledge of the particular values of the parameters (α, p, q) associated with the particular Besov space to which f is assumed to belong. Hence they are adaptive in achieving their rates without such knowledge. These two properties, namely near-optimality and adaptivity, are the same well-known strengths possessed by classical wavelet-shrinkage methods in the "function plus Gaussian noise" setting e.g., Donoho *et al.* (1995).

On the other hand, the method of proof in our case differs noticeably from those in the classical setting, following instead along the lines of that in Kolaczyk and Nowak (2002). The fact that the same type of risk rates appear in both settings, however, derives from the simple fact that orthonormal wavelet bases and free-knot, piecewise polynomial functions have the same ability to approximate functions in the stated range of Besov spaces in an optimal fashion (see DeVore 1998). For the reasons described in the introduction, we have chosen to use the latter class of functions, suitably constrained to the sampling resolution of the observations.

On a final note, we mention that the construction of our risk bounds allows one to prove that the same risk rates hold for certain other risk functions, as summarized in the following.

Corollary 1 *The risk rates of Theorem 2 hold in the Gaussian case for risk of the form* $(1/n)\sum_i E\left(\mu_i - \hat{\mu}_i\right)^2$, *in the Poisson case for risk of the form* $(1/n)\sum_i E\left(\mu_i^{1/2} - \hat{\mu}_i^{1/2}\right)^2$, *and more generally for risk of the form* $(1/n)\sum_i E\left[b(\theta_i) + b(\hat{\theta}_i) - 2b\{(\theta_i + \hat{\theta}_i)/2\}\right]$.

In particular, Corollary 1 indicates that our results hold under standard squared-error loss in the Gaussian case, and square-error loss on the square-root scale of intensity in the Poisson case.

4 Applications.

4.1 *Gamma-Ray Bursts*

Gamma-ray bursts are one of the most intriguing classes of objects in modern high-energy astrophysics, with a great deal of effort and resources being devoted to their study (e.g., Wijers 1998). Figure 1(a) shows a time series of data obtained during one such burst. It is typical to model such measurements, derived from photon arrival times, as binned counts from an inhomogeneous Poisson process. Although it is often said that such series can be as varied as snowflakes in their form, it has been found that large classes of them appear to be adequately modeled as super-positions of asymmetric exponential pulses. Information on the number, location, amplitudes, and relative width of the component pulses is of interest. A class of fast-rise, exponential-decay (FRED) models were fit by Norris *et al.* (1996), which specify that the underlying intensity of this process be a linear combination of pulse functions of the form

$$I(t) = \begin{cases} A \exp[-(|t - t_{max}|/\sigma_r)^\nu], & \text{if} \quad t < t_{max} \\ A \exp[-(|t - t_{max}|/\sigma_d)^\nu], & \text{if} \quad t > t_{max} \end{cases} , \tag{5}$$

where A is the pulse amplitude, t_{max} its location, and ν the peakedness, while σ_r and σ_d control the width of the rising and decaying portions, respectively. The fitting of such functions has been a fairly human-intensive task (J. Norris, Personal Communication).

In Figure 2(a), two estimates of the underlying intensity function are plotted. One is the FRED model fitted by Norris *et al.*, combining seven functions of the form (5). The other is the estimate obtained by the recursive partitioning (RP) method described in Section 2.2, calculated using a piecewise linear model for the natural parameter θ. Six of the seven peaks fit by Norris *et al.* are captured by the latter estimator. Note that due to our use of the canonical Poisson link function i.e., $G(\mu) = \log \mu$, we have effectively modeled the mean μ as piecewise exponential, which can be considered a rough approximation to the more sophisticated FRED model. Also plotted in Figure 2(a) are the boundaries of the intervals I in the optimally chosen partition $\hat{\mathcal{P}}$ underlying our RP estimator. It can be seen that these track the peaks and valleys

of each of the pulses in the data.

These observations suggest that from our estimator it is possible to extract starting values for the nonlinear least-squares routine that underlies the fitting of FRED models, regarding the number of pulses and their locations, amplitudes, and rates of decay. Hence our method can serve as both an effective analysis tool in its own right and a completely automated pre-processing routine in fitting FRED models. Only information on the peakedness parameter ν in (5) cannot be extracted from a single such fit, but additional insight on its values may be obtained through successive fitting of higher order models.

For example, fitting a piecewise quadratic model (in the natural parameterization) to the data yields a fit (Figure 2(b)) that matches the FRED model quite closely, including the capture of the slender fifth pulse at seven seconds missed by the piecewise linear model. In this case the nature of the partition intervals I in $\hat{\mathcal{P}}$ changes as well, running now from peak to peak, with a single curve used to fit each successive decay and rise between two adjacent peaks. Of some interest too is the fit of our method in the region of the seventh FRED pulse, near 11.5 seconds, where there are arguably two small, closely spaced pulses. Such phenomena were difficult to fit with two pulses in a numerically stable manner using the FRED model, and so the placement of the seventh pulse in this case reflects a user choice (J. Norris, Personal Communication). In the case of our own methodology, the location of a fitted pulse in between the two candidate pulses reflects an unguided attempt to fit the data in that region with a single 'pulse', which may be interpreted as a qualified measure of support for the decision in the fitting of the FRED model.

4.2 *Packet Loss Data.*

In Figure 1(b) are shown measurements resulting from an experiment in computer network traffic monitoring, conducted and analyzed by Yajnik *et al.* (1999). Network traffic consists of discrete packets of information. Packets were transmitted from a computer at the University of Massachusetts-Amherst to one located at the Swedish Institute of Computer Science (SICS), at a constant rate of one every 160 milliseconds (*ms*), for a period of two days. These packets

acted as 'probes' to measure the quality of the connection. Note was made as to whether each packet arrived or was 'lost'.

Packet loss is a phenomenon of fundamental importance to a variety of network-based applications, particularly those in multimedia. The packets are transmitted between the two computers by relaying each of them through a series of intermediate devices known as 'routers'. The loss of a packet is usually due to a decision by some router along the path between the sender and receiver *not* to pass the packet along towards its final destination, but instead to 'drop' it. This occurs when, for example, a buffer is full upon arrival of the packet at a router. Depending on the rate of loss along a path and the application at hand, packet loss has implications on such basic issues as transmission quality (e.g., audio/video) and effective use of bandwidth. An ability to effectively characterize loss rate functions in various networking contexts remains an open problem of great relevance.

As part of their analysis, Yajnik *et al.* determined that only packets with a separation of at least 1000 *ms* between their sending times could be expected to share a negligible degree of statistical dependency. Therefore, for the purpose of our own analysis we sub-sampled the original data at 1000 *ms* intervals, and then binned the resulting Bernoulli time series of loss events over disjoint five minute time intervals. The result is a time series that may be modeled as in (1), using a binomial model with a constant number $m = 300$ trials per five-second interval. These are the data displayed in Figure 1(b). Visual inspection reveals what appears to be very inhomogeneous behavior in the loss rate. This variation in the loss rate may be indicative of time-varying network conditions. The recursive partitioning estimator $\hat{\theta}_{RP}$ is depicted in Figure 3. The estimator automatically detects nine regions of distinct behavior, and provides accompanying fits that are piecewise linear on the logit scale. Evident in these results are a sharply defined constant, low-rate region between 10 and 17 hours, three brief and closely spaced spikes around 25 hours, and a succession of more smoothly varying rises and decays thereafter.

Interpretation of these results, particularly in conjunction with network operating conditions, is currently a highly challenging task. General knowledge may be lacking of even the

basic route(s) taken by sent packets along the underlying network. Moreover, the performance conditions associated with the route(s) typically are unknown. Therefore most interpretation involves a large degree of guess work, until the advent of techniques for measuring auxiliary information. For example, packet transmissions for this data began at 16:03 EST, Thursday, November 20, 1997. The spikes at hour 25 correspond to activity around 17:00 EST the following day, a Friday. Therefore, it is conceivable that this surge is due to some heavy flows deriving from processes associated with the end of the week, most likely at locations near either the sending or receiving locations, as opposed to the internet backbone itself. Furthermore, the 20+ hour periods before and after the spikes correspond to roughly the same periods of the day. And so the basic high-low-high character of each may be tied to patterns of typical 24-hour network usage, while the differences in each may possibly reflect a contrast of work-week (Thursday) versus weekend (Saturday) usage.

5 Discussion.

Motivated in part by seminal work of Donoho (1997), which establishes close connections between certain wavelet-based estimators and CART-like analyses, our approach in this paper uniquely combines nonparametric generalized linear models (NP-GLMs), tree-based modeling, and computational harmonic analysis. The result is a multiscale framework for extracting information on both the scale and trend of local components of an inhomogeneous function in an NP-GLM. Our estimators are founded on recursive partitions with piecewise polynomial fits supported on the partition intervals. This combination provides a very flexible class of estimators that admits a tractable theoretical analysis and that can be searched easily for an optimal estimator using certain tree-pruning algorithms and standard GLM fitting routines. Risk bounds for the proposed complexity penalized estimators show that the same rates one obtains with classical wavelet-shrinkage estimators in the Gaussian "function plus noise" model can be achieved more generally for a broad class of NP-GLM models.

The key device in establishing these risk bounds is our extension of the Li-Barron bound

(Li and Barron 2000, Kolaczyk and Nowak 2002). The generality of that bound allows us to handle the natural exponential families quite broadly within a single framework. The second crucial element is the use of piecewise polynomial approximations on recursive partitions, which parallels approximation by wavelet bases and allows us to take advantage of established approximation-theoretic results concerning Besov spaces.

An alternative approach to multiscale modeling in this context is through the explicit use of wavelets. For example, Sardy *et al.* (2002) introduce a wavelet-based, complexity-penalized likelihood estimator, with an ℓ_1-penalty on the size of the wavelet coefficients. However, optimization in this setting is non-trivial involving, for example, interior-point methods for finding solution of Kuhn-Tucker equations. On the other hand, optimization in our framework requires only calls to standard GLM fitting software routines, within the context of polynomial-time, tree-structured algorithms. The work of Antoniadis and Sapatinas (2001) and Antoniadis *et al.* (2001) adopts wavelet shrinkage type estimators, in analogy to the Gaussian setting, which require a computational overhead comparable to that of our approach. Their shrinkage estimators are shown to provide the usual rates of decay in classical Sobolev spaces, provided the variance functions are quadratic or cubic. The specialization to Sobolev spaces, rather than Besov, and the special requirements on the variance functions may be viewed as being more restrictive than our method, depending on the context of the problem at hand. For estimating comparatively smooth functions deriving from something like a Sobolev space, there are a variety of other methods as well, of course, including those based on splines (e.g., Green and Silverman 1994) and local polynomials (e.g., Fan and Gijbels 1996).

On a final note, we mention that our focus on the natural exponential families was only crucial for derivation of explicit bounds on the risk of our estimators. The modeling and algorithmic framework can be applied quite generally, to data types not in the exponential family, resulting in a broad class of what might be called *multiscale likelihood* models. Additionally, extensions are also possible to models using other than piecewise polynomials. Examples include the use of "wedgelet" and "platelet" models for image analysis in Willett and Nowak (2002).

Acknowledgement

This research was supported by the Army Research Office, the National Science Foundation, and the Office of Naval Research. The authors thank Rebecca Willet for her careful reading and helpful comments.

6 Appendix: Proofs.

Proof of Theorem 1: Consider the statement of part (i) of the theorem, concerning $\hat{\theta}_{RDP}$. The C-RDP $\mathcal{P}^*_{Dy}$ and its hierarchy of preceding partitions $\mathcal{P} \preceq \mathcal{P}^*_{Dy}$ may be associated with a full binary tree of depth $\log_2(n)$, with $n/2$ leaf nodes, $n/4$ nodes at the previous level, $n/8$ at the next, and so on. The algorithm starts at the lowest depth d allowing for the unique fitting of order D polynomials, as measured beginning at the root of the binary tree, i.e., containing $2^{log_2(n)-d}$ observations each. For each interval I at this depth, the polynomial maximizing the likelihood corresponding to $\{y_{1_I}, \ldots, y_{n_I}\}$ is found using a standard GLM fitting routine and saved. Then, at depth $d-1$, the process is repeated, but with the additional step that for each dyadic interval I at this depth the log-likelihood of the optimal polynomial fit is compared to the sum of the log-likelihoods of the analogous fits for the dyadic children intervals $I_{ch(I),l}$ and $I_{ch(I),r}$ at depth d, minus a penalty in the amount of 2λ for splitting I. Continuing in this fashion iteratively until the root of the underlying binary tree is reached, the estimator $\hat{\theta}_{RDP}$ is obtained at the finish. Counting the nodes of the tree, there are on the order of $(n/2 + n/4 + \cdots + 2 + 1) \sim n$ calls for GLM fits and a similar number of comparisons of the corresponding likelihoods.

Now consider part (ii) of the theorem, concerning $\hat{\theta}_{RP}$. While there are $(n-1)!$ C-RP's $\mathcal{P}^*$ in the library $\mathcal{L}$, and each may be associated with a binary tree describing the hierarchy of recursive partitions $\mathcal{P} \preceq \mathcal{P}^*$, all such $\mathcal{P}$ are in fact composed of subsets of only $n(n+1)/2 \sim n^2$ unique $I \in (0, 1]$. So the algorithm starts with the computing of a GLM fit on every such interval containing D or more observations. These fits then are used to select the solution to (3) as follows. Beginning with the intervals containing $2D$ observations, for each such interval

compare the log-likelihood value of the GLM fit with the sum of the log-likelihood values for the right and left half-intervals of length D minus an additional penalty of 2λ. Select the model (the single GLM fit or the two GLM fits) that has the larger total penalized log-likelihood value and record the selection along with this value. Then recursively consider intervals of successively longer lengths containing $m > 2D$ observations, for $m = 2D+1, \ldots, n$, and compare the log-likelihood value of the GLM fit on each interval with the optimal sum of maximum penalized log-likelihood values previously recorded for all possible pairs of subintervals at the previous step minus an additional penalty of 2λ. Note that these optimally inherited subintervals in turn may be accompanied by their own optimally inherited subintervals.

Any interval containing m observations may be partitioned into two subintervals in exactly $m-1$ ways, and only a fraction of these partitions involve pairs of subintervals both containing D or more observations. So, this last step requires fewer than $m-1$ comparisons for each m-length interval, and there are exactly $n-m+1$ unique such intervals. Therefore, fewer than $\sum_{m=1}^{n} m(n-m) = \frac{n^2(n+1)}{2} - \frac{n(n+1)(2n+1)}{6} \sim \frac{n^3}{6}$ comparisons are required to select the piecewise GLM fit that maximizes (3). In summary, the total computational cost is $O(n^2)$ calls to a GLM fitting routine and $O(n^3)$ comparisons.

Proof of Theorem 2: The overall method of proof follows along lines similar to those in Kolaczyk and Nowak (2002) which, in particular, rely on the following general result.

Theorem 3 (Kolaczyk and Nowak (2002)) *Let Γ_n be a finite collection of estimators θ' for θ, and $pen(\cdot)$ a function on Γ_n satisfying the condition*

$$\sum_{\theta' \in \Gamma_n} e^{-pen(\theta')} \leq 1 \quad . \tag{6}$$

Let $\hat{\theta}$ be a penalized maximum likelihood estimator of the form

$$\hat{\theta}(y) \equiv \arg\max_{\theta' \in \Gamma_n} \{ \ell(\theta') - 2 \ pen\,(\theta') \} \quad . \tag{7}$$

Then

$$E\left[H_n^2\left(p_\theta, p_{\hat{\theta}}\right)\right] \leq \min_{\theta' \in \Gamma_n} \left\{ K(p_\theta, p_{\theta'}) + 2\ pen(\theta') \right\} \quad , \tag{8}$$

where $K(p_\theta, p_{\theta'})$ *is the Kullback-Leibler divergence between* p_θ *and* $p_{\theta'}$.

Li and Barron (2000) first established a bound of this type, and Kolaczyk and Nowak (2002) extended it to the more general case above. To use this result, we proceed by constructing an estimator, within an appropriate space Γ_n of estimators satisfying (6), for which the quantity being minimized on the right-hand side of (8) is bounded by a term tending to zero at the required rate.

Let $\mathcal{C}_{PP} \equiv \mathcal{C}_{PP}(D, C, n)$ denote the collection of all piecewise-polynomial functions of order D on $(0, 1]$, bounded by C, such that the component pieces of each such function are restricted in number and support in one-to-one correspondence with the intervals $I \in \mathcal{P}$, for all $\mathcal{P} \in \mathcal{P}^*$ and all $\mathcal{P}^* \in \mathcal{L}$. We discretize this collection in the following manner to form our set Γ_n. First, note that any given polynomial piece on a sub-interval $I \in \mathcal{P}$, for some $\mathcal{P}$, may be expressed as a linear combination of the first D normalized Legendre polynomials, with their support scaled and shifted to I. Second, a simple argument shows that the the coefficients in such an expansion are bounded in magnitude by $B_{|I|} = \{(2D+1)|I|\}^{1/2} C'$, where $|I|$ is the width of the interval I and C' is a constant. Let $D_n[-B_{|I|}, B_{|I|}]$ denote a discretization of the range $[-B_{|I|}, B_{|I|}]$ into $n^{1/2}$ equi-spaced values. Then for $d = 1, \ldots, \lfloor n/D \rfloor$ and every d-piece member of $\mathcal{C}_{PP}$, we quantize the polynomial coefficients of each corresponding interval I to take values in $D_n[-B_{|I|}, B_{|I|}]$. Finally, for each d, we take the corresponding quantized piecewise polynomials and average-sample them on the intervals $I_i = ((i-1)/n, i/n]$, $i = 1, \ldots, n$, to obtain sequences $\theta^{(d)} = (\theta_1^{(d)}, \ldots, \theta_n^{(d)})$. Letting $\Gamma_n^{(d)}$ be the collection of such sequences, we set $\Gamma_n = \cup_{d=1}^{\lfloor n/D \rfloor} \Gamma_n^{(d)}$.

Two remarks are in order before proceeding further with the proof. First, suppose g is an element of $\mathcal{C}_{PP}$ and denote by g^q the result of quantizing the coefficients of g as described above. Then, because the magnitude of the normalized k-th Legendre polynomial supported on I is bounded by $\{(2k+1)/|I|\}^{1/2}$, it follows that the magnitude of the difference $g - g^q$ is bounded by $2C'(2D+1)n^{-1/2}$. From this bound it follows that the total quantization error

$||g - g^q||_{L_2}$ is bounded by $O(n^{-1/2})$, a fact we will make use of later in the proof. This approach to suitably quantizing the coefficients of piecewise polynomials parallels that of Willett and Nowak (2003), where a more detailed argument may be found in the context of an analogous framework of multiscale, piecewise polynomial modeling for density estimation. Second, note that technically the choice of average-sampling in defining Γ_n represents a deviation from the pointwise sampling used in a standard GLM fitting routine. However, although this choice is made for mathematical convenience, to match the average sampling that generated the unknown θ, it in fact introduces discrepancies that amount in total only to an additional error due to quadrature, and these are negligible in comparison to the rates ultimately governing our risk bounds. Hence they will be ignored.

Continuing with our proof, it follows using an argument similar to the proof of Lemma 1 in Kolaczyk and Nowak (2002) that (6) holds for the Γ_n we have defined. Specifically, note that each element $\theta^{(d)}$ in $\Gamma_n^{(d)}$, for fixed d, is an n-length sequence of polynomial-varying sub-sequences, with the sub-sequences induced by our sampling of a quantized piecewise polynomial of d pieces. There are at most $\binom{n-1}{d-1} n^{Dd/2}$ elements in $\Gamma_n^{(d)}$, since there are at most $n-1$ possible locations for the $d-1$ starting positions for the sub-sequences, not including the left-most sub-sequence. Our penalty function $\text{pen}(\cdot)$ will be of the form $\text{pen}(\theta') = \gamma \log(n) \cdot \#\{\mathcal{P}(\theta')\}$, where $\mathcal{P}(\theta')$ denotes the partition $\mathcal{P}$ corresponding to the piecewise polynomial underlying θ' and $\gamma \equiv (1 + D/2)$. Therefore,

$$\begin{aligned}
\sum_{\theta' \in \Gamma_n} e^{-\text{pen}(\theta')} &= \sum_{d=1}^{\lfloor N/D \rfloor} \sum_{\theta' \in \Gamma_n^{(d)}} e^{-\gamma d \log n} \\
&\leq \sum_{d=1}^{\lfloor N/D \rfloor} \binom{n-1}{d-1} n^{Dd/2} e^{-\gamma d \log n} \\
&\leq \sum_{d'=0}^{\lfloor N/D \rfloor} \binom{n-1}{d'} n^{Dd'/2} e^{-\gamma (d'+1) \log n} \\
&\leq \sum_{d'=0}^{\lfloor N/D \rfloor} \binom{n-1}{d'} n^{-(\gamma - D/2)(d'+1)}
\end{aligned}$$

$$
\begin{aligned}
&\leq n^{-1} \sum_{d'=0}^{\lfloor N/D \rfloor} \frac{n^{d'}}{d'!} n^{-d'(\gamma - D/2)} \\
&\leq n^{-1} \sum_{d'=0}^{\infty} \frac{1}{d'!} \\
&= n^{-1} e.
\end{aligned}
$$

Thus, (6) is satisfied for $n \geq 3$.

We now construct an estimator in Γ_n for which the quantity being minimized on the right-hand side of (8) will be bounded by a term tending to zero at the required rate. Let $\tilde{f}^{(d)}$ be the best approximation (in the L_2 sense) of the function f by a free-knot, piecewise polynomial function of d pieces. Given $f \in B^{\alpha}_{p,q}$ and the conditions on (α, p, q) stated in Theorem 2, the approximation error $||f - \tilde{f}^{(d)}||_{L_2}$ is $O(d^{-\alpha})$. See DeVore (1998), Section 6.3, for example. Noting that $\tilde{f}^{(d)}$ need not have knots at the endpoints of the I_i, we let $\tilde{f}^{(d),n}$ be that member of $\mathcal{C}_{PP}$ closest to $\tilde{f}^{(d)}$ (again, in the L_2 sense). Denote by $\tilde{f}^{(d),n,q}$ the result of quantizing the coefficients of $\tilde{f}^{(d),n}$ as described above. Finally, for the n-length sequence that results from average-sampling the function $\tilde{f}^{(d),n,q}$, we will write $\tilde{\theta}^{(d),n,q}$.

To finish the proof, consider the Kullback-Leibler distance in (8), for which we can write

$$
K(p_\theta, p_{\theta'}) = E_\theta \left\{ \log \frac{p_\theta(y)}{p_{\theta'}(y)} \right\} = E_\theta \left\{ \sum_{i=1}^{n} \log \frac{p_\theta(y_i)}{p_{\theta'}(y_i)} \right\} = \tau^{-1} \sum_{i=1}^{n} \left[\mu_i(\theta_i - \theta_i') - \{ b(\theta_i) - b(\theta_i') \} \right] \quad ,
$$

where $\mu_i \equiv E_\theta(y_i) = G^{-1}(\theta_i)$. From the expression $b(\theta_i) - b(\theta_i') = \log E_{\theta'} \left[\exp \{ (\theta_i - \theta_i') y_i \} \right]$ and Jensen's inequality it follows that $b(\theta_i) - b(\theta_i') > \mu_i'(\theta_i - \theta_i')$, and so by Cauchy-Schwartz we have

$$
K(p_\theta, p_{\theta'}) \leq \tau^{-1} ||\mu - \mu'||_{\ell_2} \cdot ||\theta - \theta'||_{\ell_2} \quad .
$$

But using the Lipschitz condition assumed for G^{-1}, we have $|\mu_i - \mu_i'| = |G^{-1}(\theta_i) - G^{-1}(\theta_i')| \leq A|\theta_i - \theta_i'|$, for some constant A, and therefore $K(p_\theta, p_{\theta'}) \leq A' ||\theta - \theta'||^2_{\ell_2}$.

Therefore, considering first the case $\hat{\theta} = \hat{\theta}_{RP}$, we have

$$
R_n(\theta, \hat{\theta}) \quad = \quad (1/n) E \left\{ H_n^2 \left(p_\theta, p_{\hat{\theta}} \right) \right\}
$$

$$
\begin{aligned}
&\leq (1/n) \min_{\theta' \in \Gamma_n} \left\{ K(p_\theta, p_{\theta'}) + 2 \text{ pen}(\theta') \right\} \\
&= (1/n) \min_{d} \min_{\theta' \in \Gamma_n^{(d)}} \left\{ K(p_\theta, p_{\theta'}) + (2d)(1 + D/2) \log n \right\} \\
&\leq (1/n) \min_{d} \min_{\theta' \in \Gamma_n^{(d)}} \left\{ A' ||\theta - \theta'||_{\ell_2}^2 + (2d)(1 + D/2) \log n \right\} \\
&\leq (1/n) \min_{d} \left\{ A' ||\theta - \tilde{\theta}^{(d),n,q}||_{\ell_2}^2 + (2d)(1 + D/2) \log n \right\} \quad . \qquad (9)
\end{aligned}
$$

However, for sequences θ and θ' produced by average sampling functions f and f', respectively, a simple argument relating coefficients of Haar functions on the discrete set $\{1, \ldots, n\}$ to those of Haar functions on the interval $[0, 1]$ is sufficient to show that $(1/n)||\theta - \theta'||_{\ell_2}^2 \leq ||f - f'||_{L_2}^2$. (See equation (27) of Kolaczyk and Nowak 2002.) So it follows that the norm within the brackets in (9) can be bounded by

$$
||f - \tilde{f}^{(d),n,q}||_{L_2}^2 = ||(f - \tilde{f}^{(d)}) + (\tilde{f}^{(d)} - \tilde{f}^{(d),n}) + (\tilde{f}^{(d),n} - \tilde{f}^{(d),n,q})||_{L_2}^2 \quad . \qquad (10)
$$

Finally, we apply the triangle inequality to the right-hand side of (10). The first resulting squared term $||f - \tilde{f}^{(d)}||_{L_2}^2$ will be of order $O(d^{-2\alpha})$. The second squared term will be of order $O(d/n)$, by construction, and the last will be of order $O(1/n)$ by virtue of our quantization. The magnitude of the cross-terms follow accordingly. Therefore, ignoring terms of no consequence, we obtain a bound that behaves like $d^{-2\alpha} + (d/n) \log n$, which minimized in d yields that

$$
R_n \leq O\left\{ (\log n / n)^{2\alpha/(2\alpha+1)} \right\} \quad . \qquad (11)
$$

In the case of $\hat{\theta} = \hat{\theta}_{RDP}$, the definition of Γ_n is adjusted accordingly, the inequality (6) follows similarly, and the quantity $\#(\mathcal{P}(\theta'))$ behaves like $d \log n$ instead of d on $\Gamma_n^{(d)}$, which accounts for the extra factor of $\log n$ in that case.

Proof of Corollary 1: This result is a simple consequence of the method of proof of Theorem 3 (see Kolaczyk and Nowak 2002). Specifically, in deriving equation (8) one actually begins with the inequality $H^2(p_\theta, p_{\hat{\theta}}) \leq -2 \log \mathcal{A}(p_\theta, p_{\hat{\theta}})$, where $\mathcal{A}(p_\theta, p_{\hat{\theta}}) \equiv \int (p_\theta(y) p_{\hat{\theta}}(y))^{1/2} \nu_n(y)$ is the 'affinity' between densities p_θ and $p_{\hat{\theta}}$. The rest of the proof (resulting in the right-hand side of

(8)) then continues by bounding minus twice the logarithm of the affinity. A simple calculation shows that if the y_i are independent observations from a NEF, as defined in (1), then (i) the total affinity is the sum of the marginal affinities (i.e., with respect to the densities of each marginal component y_i), and (ii) these marginal affinities have the form $b(\theta_i) + b(\hat{\theta}_i) - 2b[(\theta_i + \hat{\theta}_i)/2]$. In the Gaussian and Poisson cases this last expression reduces to $(\theta_i - \hat{\theta}_i)^2$ and $(\theta_i^{1/2} - \hat{\theta}_i^{1/2})^2$, respectively.

References.

Alpert, B.K. (1993). A class of bases in L^2 for the sparse representation of integral operators. *SIAM Journal of Mathematical Analysis,* **24**, 2466-262.

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika,* **88**, 805-820.

Antoniadis, A., Besbeas, P., and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhya, Series A,* **63**, 309-327.

Barry, D. and Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association,* **88**, 309-319.

Breiman, L., Friedman, J., Olshen, R., and Stone, C.J. (1983), *Classification and Regression Trees.* Belmont, CA: Wadsworth.

DeVore, R.A. (1998). Nonlinear approximation. *Acta Numerica,* **7** 51 - 150.

Donoho, D.L. (1997). CART and Best-Ortho-Basis: A Connection. *Annals of Statistics,* **25** 1870-1911.

Donoho, D.L., Dyn, N., Levin, D., and Yu, T.P.-Y. (2000). Smooth multiwavelet duals of Alpert bases by moment-interpolating refinement. *Applied and Computational Harmonic Analysis,* **9**, 166-203.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (Disc: p337-369), *Journal of the Royal Statistical Society B,* **57**,

301-337.

EFROMOVICH, S. (1999). *Nonparametric curve estimation: methods, theory, and applications.* Springer, New York.

FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modeling and its Applications.* Chapman and Hall, London.

GREEN, P.J. AND SILVERMAN, B.W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

LI, Q.J. AND BARRON, A.R. (2000). Mixture density estimation. In *Advances in Neural Information Processing Systems 12*, S.A. Solla, T.K. Leen, K.-R. Müller, eds., MIT Press.

KOLACZYK, E.D. AND NOWAK, R.D. (2002). Multiscale likelihood analysis and complexity penalized estimation. *Annals of Statistics,* (tentatively accepted for publication).

NORRIS, J.P. *et al.* (1996). Attributes of Pulses in Long Bright Gamma-Ray Bursts, *The Astrophysical Journal,* **459**, 393 - 412.

PRANDONI, P. AND VETTERLI, M. (1999). Approximation and compression of piecewise smooth functions. *Phil. Trans. R. Soc. Lond. A,* **357**, 2573-2591.

SARDY, S., ANTONIADIS, A., AND TSENG, P. (2002). Automatic smoothing with wavelets for a wide class of distributions. *Journal of Computational and Graphical Statistics,* (tentatively accepted).

WILLETT, R. M. AND NOWAK, R.D. (2003). Multiscale density estimation. Submitted to *IEEE Transactions on Information Theory.*

WILLETT, R. M., AND NOWAK, R. D. (2002). Platelets: A Multiscale Approach for Recovering Edges and Surfaces in Photon Limited Medical Imaging. To appear in *IEEE Transactions on Medical Imaging, Special Issue on Wavelets and Medical Imaging.*

WIJERS, W. (1998). The Burst, the Burster and its Lair," *Nature,* **393**, 13.

YAJNIK, M., MOON, S., KUROSE, J., AND TOWSLEY, D. (1999). Measurement and modeling of the temporal dependence in packet loss. *Proceedings of the 18th Annual Conference of the IEEE Computer and Communications Societies (INFOCOM), 345-353.*

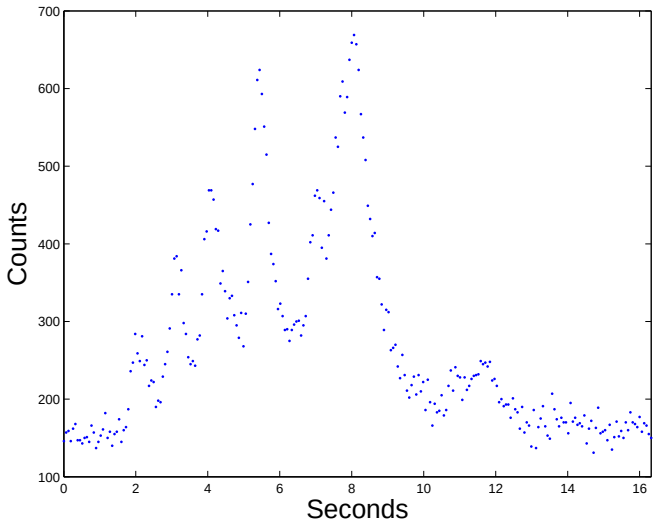
ZHANG, K., DENG, M., CHEN, T., WATERMAN, M., AND FENGZHY, S. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences,* **11**, 7225-7339.

CAPTIONS FOR FIGURES.

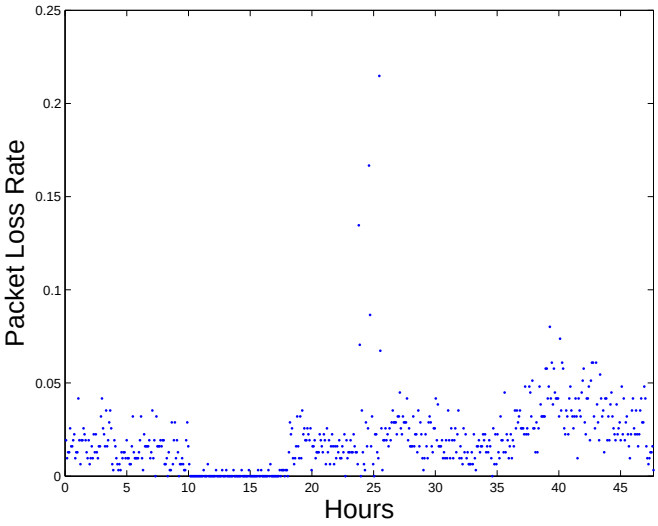
Figure 1. – Examples of two inhomogeneous time series: (a) Gamma-ray burst data obtained by the BATSE instruments on board NASA's former Compton Gamma Ray Observatory. Time series were created by aggregating original photon arrival times (photon energies 25-55 keV) from burst #1425 into a total of $n = 256$ equi-spaced bins; (b) Observed packet loss rates among packets sent from the University of Massachusetts-Amherst to the Swedish Institute of Computer Science over a two-day period by Yajnik *et al.* (1999). Packets, originally sent at 160 *ms* time intervals, were sub-sampled at 1000 *ms* time intervals. Rates were calculated as observed fraction of packets lost, for each successive group of 300.

Figure 2. – Comparison of the MS-GLM estimate (solid) and the estimate of Norris *et al.* 1996 (dot-dashed) for the intensity underlying BATSE burst # 1425. The library-based recursive partitioning estimator (i.e., $\hat{\theta}_{RP}$) is shown, based on a Poisson model and choice of (a) piecewise log-linear and (b) piecewise log-quadratic intensity. The vertical, dotted lines denote the boundary points of the optimal recursive partition.

Figure 3. – MS-GLM estimate of packet loss rate for the data of Yajnik *et al.* (1999), using library-based recursive partitioning (i.e. $\hat{\theta}_{RP}$), under a binomial model and piecewise linear specification for the logit. The vertical, dotted lines denote the boundary points of the optimal recursive partition.

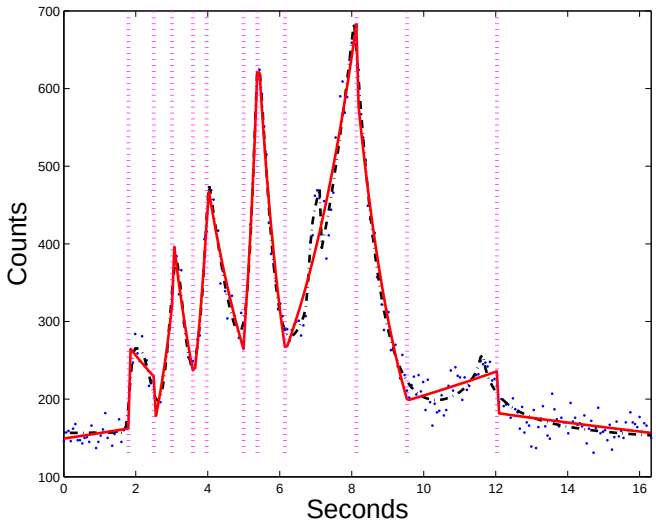

Counts
Seconds
700
600
500
400
300
200
100
0
2
4
6
8
10
12
14
16

(a)

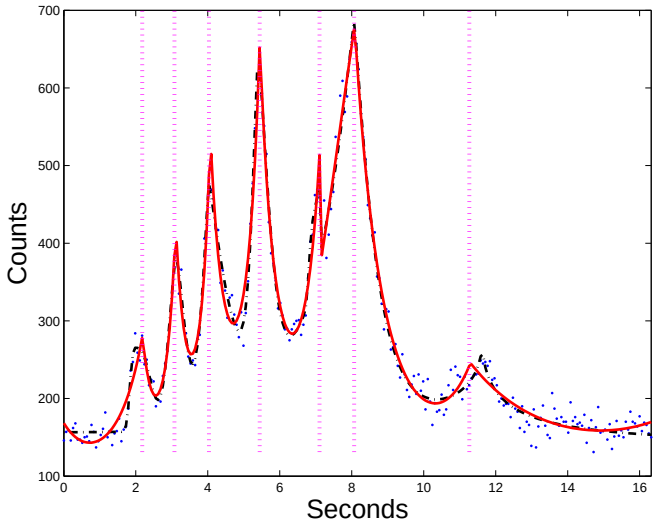

Packet Loss Rate
Hours
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
35
40
45

(b)

Figure 1:


Counts
Seconds
700
600
500
400
300
200
100
0
2
4
6
8
10
12
14
16

(a)


Counts
Seconds
700
600
500
400
300
200
100
0
2
4
6
8
10
12
14
16

(b)

Figure 2:

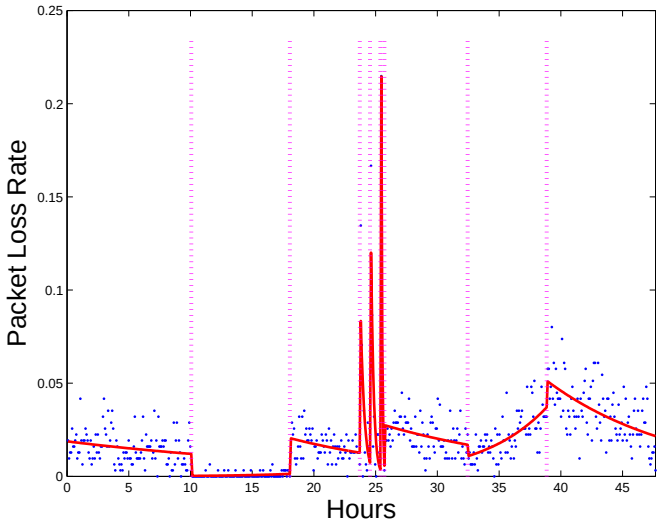

Packet Loss Rate
Hours
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
35
40
45

Figure 3: