Multiscale Generalised Linear Models for Nonparametric Function Estimation

by

Eric D. Kolaczyk

Department of Mathematics and Statistics

Boston University, Boston, MA, 02215

kolaczyk@math.bu.edu

and

Robert D. Nowak

Department of Electrical and Computer Engineering

University of Wisconsin, Madison, WI, 53706

nowak@engr.wisc.edu

Summary

We present a method for extracting information about both the scale and trend of local components of an inhomogeneous function in a nonparametric generalised linear model. 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 summarise the smooth trend within each interval. Our estimators are formulated as solutions of complexity-penalised likelihood optimisations, 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 software routines for generalised linear models, within the context of efficient, tree-based, polynomial-time algorithms. A risk analysis shows that these estimators achieve the same asymptotic rates in the nonparametric generalised linear model as the classical wavelet-based estimators in the Gaussian 'function plus noise' model, for suitably defined ranges of Besov spaces. Numerical simulations show that the method tends to perform at least as well, and often better, than alternative wavelet-based methodologies in the context of finite samples, while applications to gamma-ray burst data in astronomy and packet loss data in computer network traffic analysis confirm its practical relevance.

Some key words: Astronomy; Computer network traffic; Minimax; Piecewise polynomial; Recursive partitioning; Wavelet.

1 Introduction

Multiscale methods, particularly those based on wavelets and related paradigms, have become a leading choice for nonparametric curve estimation when the curve to be estimated is inhomogeneous in nature. Many 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 additive noise, often independent, identically distributed and Gaussian. The goal then is to produce an accurate estimate of $\theta = (\theta_1, \ldots, \theta_n)$, particularly so as not to obscure the inhomogeneous structure of f in the process, such as jumps, cusps and so on.

In this paper we are interested in 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. Secondly, the function f may itself be a composite of simpler, localised homogeneous functions, and accurate indication of both the scale, or extent, and overall trend of each component is felt to be useful in addressing the scientific questions accompanying the data.

In § 4, two such examples are examined. The first example concerns photon arrival-time data from an astronomical gamma-ray burst, modelled as a realisation of a Poisson counting process, such as that shown in Fig. 2. Accurate characterisation of the number, location and extent of component pulses underlying such bursts is a fundamental part of analysing such data in the field of high-energy astrophysics (Norris et al. 1996). The second example concerns the monitoring of computer network traffic and, more specifically, the loss, or 'dropping', of packets in transmitting information between two locations, such as shown in Fig. 3. Packet loss data takes the form of a Bernoulli process which, with appropriate subsampling and aggregation, may be modelled 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 (Yajnik et al.

1999).

We present a method for extracting information about both the scale and trend of local components of an inhomogeneous function in a nonparametric generalised linear model. 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 summarise the smooth trend within each interval. Our estimators are formulated as solutions of complexity-penalised likelihood optimisations, 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 software routines for generalised linear models, 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 nonparametric generalised linear model as the classical wavelet-based estimators in the Gaussian 'function plus noise' model, for suitably defined ranges of Besov spaces.

We are currently aware only of a small amount of related work. For example, Sardy et al. (2002) introduce a class of wavelet-based complexity-penalised likelihood estimators for nonparametric generalised linear models, where the penalty is based on the ℓ_1 norm of the wavelet coefficients. However, the optimisation involved in calculating these estimators is expensive, requiring the use of interior-point methods for the solution of the corresponding Kuhn-Tucker equations. Antoniadis & Sapatinas (2001) and Antoniadis et al. (2001), adopting a mean-squared error criterion, extend the wavelet shrinkage paradigm to natural exponential families. In this case the calculations are computationally efficient and optimal risk rates are proved, 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, such as Gaussian, Poisson and binomial models.

More generally, Kohler (1999) has made use of piecewise polynomials in a manner similar to that herein, but in the context of complexity-penalised least-squares estimation. For his estimators, near-optimal risk rates are derived for recovering members of the class of p-smooth functions, and the existence of a computational algorithm of complexity similar to one of the two presented herein is established. Lastly, we point out that the tenor of the work presented here is inspired by that of the seminal paper by Donoho (1997), which establishes close connections between certain wavelet-based estimators and CART-like analyses. By extension, therefore, our work may be associated with the large body of work relating to CART (Breiman et al. 1983) and similar methods.

Matlab code implementing the methodology proposed in this paper, as illustrated in § 4, is available at http://math.bu.edu/people/kolaczyk/software.html .

2 Multiscale Generalised Linear Models

2.1 Background

Let $y = (y_1, \ldots, y_n)$ be independent observations from a natural exponential family, with marginal probability or density functions of the form

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

which are common up to the canonical parameters θ_i. These parameters are assumed in turn to relate to an underlying mean parameterisation μ_i through the expression $\theta_i = G(\mu_i)$, where $G(\cdot)$ is the canonical link function; that is G is defined implicitly through $b(\cdot)$ in the usual manner.

As our interest is in one-dimensional nonparametric regression, we suppose that the vectors $\theta = (\theta_1, \ldots, \theta_n)$ and $\mu = (\mu_1, \ldots, \mu_n)$ are associated with unknown functions,

f and h say, respectively, through some sort of discretisation. Motivated by the types of data presented in § 4, we choose to specify the discretisation in the mean parameterisation through average sampling, so that $\mu_i \equiv n \int_{I_i} h(t)dt$, where $I_i \equiv ((i-1)/n, i/n]$. We then define h through f via the link function, i.e. $h(t) = G^{-1}\{f(t)\}$, and suppose that f is a member of some appropriately defined class of functions $\mathcal{F}([0,1])$. The dispersion parameter τ will be taken to be fixed and known.

Our focus in this paper is on the estimation of θ, from which estimates of μ can be obtained. 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: a nested hierarchy of recursive, dyadic partitions; and 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 as structure in the function beyond a certain degree of polynomial smoothness. The second element then states that this remaining smoothness is represented also concisely, through a handful of so-called scale coefficients.

We therefore propose an approach to modelling 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, and corresponding estimators $\hat{\mu}$ of μ, that are simple to define, easily interpreted and produced by efficient computational algorithms. Yet they also attain the type of near-optimal asymptotic 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.

2.2 *The basic method: recursive dyadic partitioning*

Consider the class of recursive dyadic partitions of the unit interval $(0, 1]$. Assume that $n = 2^J$ is a power of two; this assumption will be relaxed in § 2.3. Beginning with the unit interval, we partition 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 $\mathcal{P}^*_{Dy} \equiv \{I_i\}_{i=1}^n$ is achieved. Note that implicit in our definition of $\mathcal{P}^*_{Dy}$ is its association with all of the dyadic partitions $\mathcal{P}$ encountered intermediate to $(0, 1]$ and $\{I_i\}_{i=1}^n$. The partial ordering induced by the operation of successive refinement will be expressed using the symbols '$\prec$' and '$\preceq$'; for example $\mathcal{P} \prec \mathcal{P}^*_{Dy}$.

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 not greater than 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 functions on $(0, 1]$ involving constant and/or linear components, with the first component restricted to the subinterval $(0, 0.5]$ and the second to the subinterval $(0.5, 1]$. Note that a polynomial of order D has degree $D - 1$.

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

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

where $\lambda \equiv \lambda(n; D) = (D/2)\log(n)$ is a smoothing parameter. This is a maximum complexity-penalised likelihood estimator of θ, where the penalty attempts to control the number of polynomial components that are used. As such, it is similar in spirit

to model-selection procedures based on penalised contrasts (Birgé & Massart, 1998; Baraud 2000). The optimisation in (2) is over all piecewise-polynomial fits on a given partition $\mathcal{P}$, over all partitions $\mathcal{P} \preceq \mathcal{P}^*_{Dy}$. In the absence of a penalty, the optimal choice of partition would be $\hat{\mathcal{P}} = \mathcal{P}^*_{Dy}$, with n components, which simply returns the nonparametric maximum likelihood estimate. In the special case of $D = 1$, where piecewise-constant functions are used, the estimator $\hat{\theta}_{RDP}$ reduces to a one-dimensional version of the method of Donoho (1997) under the Gaussian noise model, and is equivalent to one of the estimators introduced in Kolaczyk & Nowak (2004) under the Gaussian and Poisson models, although the latter derives from a rather different conceptual foundation.

Note that contained in $\hat{\theta}_{RDP}$ is information about both the scale, or extent, of localized components in θ, because of 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 to possess properties of asymptotic near-optimality and adaptivity in terms of risk; see § 3.

2.3 *Extension of the basic method: library of partitions*

Despite its advantages, the choice of a single fixed hierarchy of dyadic partitions underlying $\hat{\theta}_{RDP}$ may be seen as a limitation in some contexts, if, for example, it is hypothesised that separation of localised 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 § 4.2.

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 subintervals, $I_{ch(I),l}$ and $I_{ch(I),r}$ say, the left and right 'children' intervals. The final partition $\mathcal{P}^*$ resulting from this process will be called a complete recursive partition. We note again that, by analogy with the case of $\mathcal{P}^*_{Dy}$ above, implicit in our definition of any given $\mathcal{P}^*$ is its association with a particular sequence of recursively defined partitions $\mathcal{P}$ encountered in moving from $(0,1]$ to $\{I_i\}_{i=1}^n$.

While the basic form of estimator defined in (2) for the complete recursive dyadic partition $\mathcal{P}^*_{Dy}$ generalises immediately to arbitrary fixed complete recursive partitions $\mathcal{P}^*$, such extensions may suffer from an analogous lack of flexibility in the placement of changepoints. Instead, a more promising alternative involves the use of a library of complete recursive partitions. Let $\mathcal{L}$ be the library of all $(n-1)!$ possible complete recursive partitions $\mathcal{P}^*$, and define $PP(\mathcal{P}; D)$ as before, for all $\mathcal{P} \preceq \mathcal{P}^*$ and $\mathcal{P}^* \in \mathcal{L}$. Then, generalising (2), we define

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

where here $\lambda \equiv \lambda(n; D) = (1 + D/2)\log(n)$.

This too is a maximum complexity-penalised likelihood estimator, but now defined over a much richer model space. In particular, it includes all possible partitions of $(0,1]$ into m subintervals, for $m = 0, 1, \ldots, n$, where the subintervals 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 complete recursive partitions $\mathcal{P}^*$. This redundancy is caused by 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 complete recursive partitions $\mathcal{P}^* \in \mathcal{L}$ using an algorithm of only $O(n^3)$ complexity. Furthermore, not surprisingly in light of the properties of $\hat{\theta}_{RDP}$, the estimator $\hat{\theta}_{RP}$ enjoys similar near-optimality and adaptivity properties.

3 Properties of Estimators

3.1 *Algorithmic properties*

The estimators $\hat{\theta}_{RDP}$ and $\hat{\theta}_{RP}$ are 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 fitted to the subset of observations $\{y_{1_I}, \ldots, y_{n_I}\}$ in y corresponding to I, using a standard fitting routine for generalised linear models. Therefore, the primary algorithmic hurdle 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 summarised in the following theorem.

Theorem 1 . (i.) *The estimator $\hat{\theta}_{RDP}$ may be calculated using $O(n)$ calls to a generalised linear model fitting routine and $O(n)$ comparisons of the resulting penalised likelihood values.* (ii.) *The estimator $\hat{\theta}_{RP}$ may be calculated using $O(n^2)$ calls to a generalised linear model fitting routine and $O(n^3)$ comparisons of the resulting penalised*

likelihood values.

Note that the stated complexities are at the level of fits of generalised linear models and model comparisons, as opposed to being at the level of elementary operations; it is difficult to pin down the latter in a precise fashion because of the iterative nature of the iteratively reweighted least squares algorithm typically used in fitting a generalised linear model. However, we mention that it took just over four minutes on a personal computer with a 3.2GHz processor to fit the recursive partitioning estimator $\hat{\theta}_{RP}$ to the gamma-ray burst data in § 4.2; to fit the recursive dyadic partitioning estimator $\hat{\theta}_{RDP}$ to the same data took just under two seconds on the same machine.

Proof of Theorem 1 involves tree-based arguments of a dynamic-programming flavour. In the case of $\hat{\theta}_{RDP}$, the argument is standard and parallels, for example, those found in Donoho (1997), based on the framework of a bottom-up optimal tree-pruning algorithm, like that underlying CART (Breiman et al., 1983). In the case of $\hat{\theta}_{RP}$, 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 & Nowak (2004); details may be found in the Appendix. Note 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 & Hartigan, 1993), approximation and compression of digital signals (Prandoni & Vetterli, 1999) and the segmentation of haplotype data in genetics (Zhang et al., 2002).

3.2 *Risk properties*

Define the loss in estimating θ by a value $\hat{\theta}$ through the squared Hellinger distance,

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

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 has a secondary role, allowing for additional fine tuning of the definition of the space; see Donoho et al. (1995) for an overview or DeVore (1998) for an accessible, detailed introduction to this and related topics.

The following may be said concerning the two estimators defined in § 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 and G^{-1} are Lipschitz on their respective domains. Let $\theta = (\theta_1, \ldots, \theta_n)$ be derived from f as described at the start of § 2.*

Then for the estimators $\hat{\theta}_{RDP}$ and $\hat{\theta}_{RP}$, defined in (2) and (3) respectively,

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

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

A few comments are in order. First, from these results it is apparent that both estimators have near-optimal asymptotic risk rates, as they differ from the standard asymptotically optimal rates only by a factor involving a power of $\log n$. The factor for the recursive dyadic partition estimator is slightly larger than that for the recursive partition estimator, but this could well be improved through the use of more involved arguments than those we employ in the Appendix, at the cost of simplicity. Secondly, neither of our estimators is provided with 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, other than that $\alpha < D$. 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 noise' model with Gaussian noise (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 & Nowak (2004). The fact that the same type of risk rates appear both here and in the classical setting, 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.

Finally, 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 summarised in the following corollary, proof of which parallels that of Corollary 1 of Kolaczyk & Nowak (2004).

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

4 Applications

4.1 *Simulation results*

A small simulation study was conducted to evaluate the finite-sample properties of our proposed estimators. The design of our study parallels that of Antoniadis & Sapatinas (2001). We simulated from their Poisson and binomial noise models, with

$n = 256$, using either their smooth function or their burst-like function as $\mu(t)$, calibrated to achieve their 'medium' signal-to-noise ratio. A total of $M = 100$ trials were run for each of the four resulting cases, i.e. Poisson or binomial, and smooth or burst. For the Poisson cases, the underlying intensity function was estimated using our recursive dyadic partition method and our recursive partition method, as well as three wavelet-based methods from the literature, namely the wavelet shrinkage estimator of Antoniadis & Sapatinas (2001), the ℓ_1-penalised likelihood estimator of Sardy et al. (2002) and the method of Donoho (1993), based on the use of Anscombe's transformation to normalise the data. For the binomial case, the underlying probability function was estimated using our recursive dyadic partition and recursive partition methods again, the method of Antoniadis & Sapatinas (2001), and an extension of Donoho's (1993) method, based on the arcsine transformation. The method of Sardy et al. (2002) was omitted because an implementation of their approach for the binomial case is currently not available. In both cases, piecewise linear models were used for the recursive dyadic partition and recursive partition estimators, and the least-asymmetric wavelets of order 8 were used for all of the wavelet-based estimators.

Results of the simulation study are shown in Fig. 1. From the four sets of boxplots it can be seen that the recursive dyadic partition method of this paper does at least as well as any of the others, and, in all but one case, noticeably better than the wavelet-based methods. The recursive partition method, in contrast, tends to do about as well as the other wavelet-based methods for the smooth signal, but better for the burst signal. The difference in performance between the recursive dyadic partition method and the recursive partition method is probably due primarily to the relative size of their penalties, λ, with the implication being that a less conservative choice could be useful. Among the wavelet-based methods, the relative performance varies.

4.2 *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 (Wijers, 1998). Figure 2 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 modelled as superpositions of asymmetric exponential pulses. Information about the number, location, amplitudes and relative width of the component pulses is of interest. A class of asymmetric exponential pulse models were fitted by Norris et al. (1996), and called fast-rise, exponential decay models by these authors. These models specify that the underlying intensity of this process be a linear combination of pulse functions of the form

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

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, as mentioned by J. Norris in a personal communication.

In Fig. 2(a), two estimates of the underlying intensity function are plotted. One is the asymmetric exponential pulse model fitted by Norris et al. (1996), which combines seven functions of the form (5). The other is the estimate obtained by the recursive partitioning method described in § 2.3, calculated using a piecewise linear model for the natural parameter θ. Six of the seven peaks fitted by Norris et al. are captured by the latter estimator. Note that because we used the canonical Poisson link function, i.e. $G(\mu) = \log \mu$, we have effectively modelled the mean μ as piecewise exponential,

which can be considered a rough approximation to the more sophisticated asymmetric exponential pulse model. Also plotted in Fig. 2(a) are the boundaries of the intervals I in the optimally chosen partition $\hat{\mathcal{P}}$ underlying our recursive partitioning 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 asymmetric exponential pulse 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 this other type of model. 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 parameterisation, yields a result, shown in Fig. 2(b), that matches the fit of Norris et al. 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 pulse of Norris et al., at approximately 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 asymmetric exponential pulse model, and so the placement of the seventh pulse in this case reflects a user choice, according to a personal communication from J. Norris. 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 of Norris et al.

4.3 *Packet loss data*

In Fig. 3 are shown measurements resulting from an experiment in computer network traffic monitoring, conducted and analysed by Yajnik et al. (1999). Network traffic consists of discrete packets of information. Packets were transmitted from a computer at the University of Massachusetts at Amherst to one located at the Swedish Institute of Computer Science, at a constant rate of one every 160 milliseconds, 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 for such basic issues as transmission quality and effective use of bandwidth. An ability to characterise loss rate functions effectively in various networking contexts remains an open problem of great relevance.

As part of their analysis, Yajnik et al. (1999) determined that only packets with a separation of at least 1000 milliseconds between their sending times could be expected to share a negligible degree of statistical dependency. Therefore, for the purpose of our own analysis we subsampled the original data at 1000-millisecond 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 modelled as in (1), using a binomial model with a constant number $m = 300$ of trials per five-second interval. These are the data displayed in Fig. 3. Visual inspection reveals what appears to be

very inhomogeneous behaviour 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 Fig. 3. The estimator automatically detects nine regions of distinct behaviour, 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. Such results can be of major benefit in the inference of network operating conditions, as even the most basic information, such as the basic route(s) taken by packets sent along the underlying network, is generally lacking.

5 Discussion

Various extensions on our basic framework are possible. First, one might allow for the use of variable-degree polynomials. It is straightforward to define and implement such estimators in practice, although the corresponding risk theory appears at present to be rather limited, because of limitations in the associated approximation theory; see Willett and Nowak (2003b) for details in the context of density estimation. Secondly, it is natural to consider the case in which the dispersion parameter τ in (1) is unknown. One approach is to produce an estimate $\hat{\tau}$ from the deviance residuals of an initial fit $\hat{\theta}_{RDP}$ or $\hat{\theta}_{RP}$, after which this fit is re-calculated using $\tau = \hat{\tau}$. Thirdly, although the discontinuous nature of our estimators lends important information to the applications in this paper, in other contexts it may be considered undesirable. In that case one may wish to adapt the moment-interpolating wavelets of Donoho et al. (2000) to post-process our estimates, which should provide a computationally efficient way of producing smoothed versions of the estimators with the same risk properties. Lastly, we note that our focus on the natural exponential families in this paper was only crucial

for the derivation of explicit bounds on the risk of our estimators. The modeling and algorithmic framework can be applied quite generally, using densities $p(y)$ of fairly arbitrary nature, yielding a broad class of what might be called 'multiscale likelihood' models. Additionally, extensions are also possible to models using other than piecewise polynomials, as in the use of 'wedgelet' and 'platelet' models for image analysis in Willett & Nowak (2003a).

Acknowledgment

This research was supported by the U.S. Army Research Office, the U.S. National Science Foundation, and the U.S. Office of Naval Research. The authors thank Rebecca Willett for her careful reading and helpful comments, and Fanis Sapatinas and Sylvain Sardy for supplying their Matlab code for the wavelet-based methods involved in the comparisons of § 4.1. The comments and suggestions of the editor and three anonymous referees are also gratefully acknowledged.

Appendix

Proof of Theorem 1. Consider the statement of part (i) of the theorem, concerning $\hat{\theta}_{RDP}$. The complete recursive dyadic partition $\mathcal{P}^*_{Dy}$ 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. For any given interval $I \in \mathcal{P} \preceq \mathcal{P}^*_{Dy}$, let ℓ_I be the loglikelihood value of an interval I, and let CPL_I be the complexity penalised loglikelihood value of the optimal submodel on I. 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 fitting routine for generalised linear models and the value $CPL_I = \ell_I - \lambda$ is 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 value $\ell_I - \lambda$ is compared to $CPL_{I_{ch(I),l}} + CPL_{I_{ch(I),r}}$, the cost of simply joining the optimal submodels for the dyadic children intervals $I_{ch(I),l}$ and $I_{ch(I),r}$ at depth d. If the former is larger, then we select the generalised linear model for the full interval I; otherwise, the optimal submodels for the two children subintervals are chosen. The quantity CPL_I is then set accordingly. Continuing in this fashion iteratively until the root of the underlying binary tree is reached, we finally obtain the estimator $\hat{\theta}_{RDP}$. If we count the nodes of the tree, there are of the order of $(n/2 + n/4 + \cdots + 2 + 1) \sim n$ calls for generalised linear model 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 $(n-1)!$ complete recursive partitions $\mathcal{P}^*$ in the library $\mathcal{L}$, all such partitions are, in fact, composed of subsets of only $\binom{n+1}{2}$ unique $I \in (0, 1]$, since each interval is defined by the selection of two endpoints from the set $\{0, 1, \ldots, n\}$. Therefore, the algorithm starts with the computation of a generalised linear model fit and corresponding loglikelihood value ℓ_I on every interval I containing D or more observations. This requires $O(n^2)$ calls to a fitting routine for generalised linear models, one for each candidate interval. These fits then are used to select the solution to (3) as follows. Beginning with the intervals containing $2D$ observations, not necessarily dyadic, for each such interval I, compare $\ell_I - \lambda$ with the sum of the complexity-penalised likelihoods of the optimal submodels in the two subintervals containing D observations. Record CPL_I accordingly. Then recursively consider intervals of successively longer lengths containing $m > 2D$ observations, for $m = 2D+1, \ldots, n$. If I is of length m, it can be partitioned into two subintervals in $m-1$ ways. Let $\{I_l^i, I_r^i\}_{i=1}^{m-1}$ denote all possible pairs of subintervals of I such that $I_l^i \bigcup I_r^i = I$. We compare the quantity $\ell_I - \lambda$ with $\max_i\{CPL_{I_l^i} + CPL_{I_r^i}\}$. If $\ell_I - \lambda$ is larger, then we select the generalised linear model on the full interval I; otherwise, we select the models inherited from the optimal subintervals. The quantity CPL_I is then set accordingly. This selection process operates recursively, with inter-

vals of increasing lengths $m = 2D+1, \ldots, n$, and terminates with the final model for the entire interval $(0, 1]$.

The computational complexity of this recursion is $O(n^3)$. This result is established as follows. At each stage of the recursion, no more than $m-1$ comparisons are required for each interval containing m observations. Furthermore, there are only $n-m+1$ intervals containing m observations. Therefore, the total number of comparisons required to compute the estimator $\hat{\theta}_{RP}$ is fewer than $\sum_{m=D}^{n}(m-1)(n-m+1) = O(n^3)$.

Proof of Theorem 2. The overall method of proof follows along lines similar to those in Kolaczyk & Nowak (2004). In particular, we begin with the fact that, for Γ_n a finite collection of estimators θ' for θ and $\mathrm{pen}(\cdot)$ a function on Γ_n satisfying the condition

$$\sum_{\theta' \in \Gamma_n} e^{-\mathrm{pen}(\theta')} \leq 1 \ , \tag{A1}$$

it can be shown that penalised likelihood estimators defined to maximise $\{\, \ell(\theta') - 2 \ \mathrm{pen}\,(\theta')\}$ over $\theta' \in \Gamma_n$ satisfy the inequality

$$E\left\{H_n^2\left(p_\theta, p_{\hat{\theta}}\right)\right\} \leq \min_{\theta' \in \Gamma_n} \left\{\, K(p_\theta, p_{\theta'}) + 2 \ \mathrm{pen}(\theta') \,\right\} \ , \tag{A2}$$

where $K(p_\theta, p_{\theta'})$ is the Kullback-Leibler divergence between p_θ and $p_{\theta'}$. Li & Barron (2000) first established a bound of this type, and Kolaczyk & Nowak (2004) modified it to apply to more general contexts like the case above.

To use this result, we begin by constructing an appropriate space Γ_n associated with the estimator $\hat{\theta}_{RP}$. The case of $\hat{\theta}_{RDP}$ follows similarly. 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 $\mathcal{P} \preceq \mathcal{P}^*$ and

all $\mathcal{P}^* \in \mathcal{L}$. Next, associating each member of this collection with a set of normalised Legendre polynomials, we discretise it by quantising the coefficients in the normalised representation on a grid of $n^{1/2}$ equispaced values; see Willett & Nowak (2003b), for example. Average-sampling each of these quantised piecewise polynomials, and letting $\Gamma_n^{(d)}$ be the set of such resulting sequences, $\theta^{(d)} = (\theta_1^{(d)}, \ldots, \theta_n^{(d)})$ say, we set $\Gamma_n = \cup_{d=1}^{\lfloor n/D \rfloor} \Gamma_n^{(d)}$. It then follows using an argument similar to the proof of Lemma 1 in Kolaczyk & Nowak (2004) that (A1) holds for the Γ_n we have defined, with $\mathrm{pen}(\theta') = (1 + D/2)\log(n)\,\#\{\mathcal{P}(\theta')\}$, where $\mathcal{P}(\theta')$ denotes the partition $\mathcal{P}$ corresponding to the piecewise polynomial underlying θ'. The penalty for $\hat{\theta}_{RDP}$ can be chosen to be smaller, i.e. $\mathrm{pen}(\theta') = (D/2)\log(n) \cdot \#\{\mathcal{P}(\theta')\}$, simply because there are fewer partitions in that case; see Willett and Nowak (2003b) for more details.

We now construct an estimator in Γ_n for which the quantity being minimised on the right-hand side of (A2) 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 $f \in B_{p,q}^{\alpha}$ 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, §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. Lastly, denote by $\tilde{f}^{(d),n,q}$ the result of quantising the coefficients of $\tilde{f}^{(d),n}$ as described above. In addition, recalling the relationship between canonical and mean parameterisations outlined at the start of § 2, let $\tilde{h}^{(d),n,q} = G^{-1}\left\{\tilde{f}^{(d),n,q}\right\}$, $\tilde{\mu}_i^{(d),n,q} = n\int_{I_i} \tilde{h}^{(d),n,q}(t)dt$, and $\tilde{\theta}_i^{(d),n,q} = G\left\{\tilde{\mu}_i^{(d),n,q}\right\}$.

To finish the proof, consider the Kullback-Leibler distance in (A2), 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 $b(\theta_i) - b(\theta_i') = \log E_{\theta'} \left[\exp\left\{(\theta_i - \theta_i')y_i\right\}\right]$ and Jensen's inequality it follows that $b(\theta_i) - b(\theta_i') > \mu_i'(\theta_i - \theta_i')$, and so by Cauchy-Schwarz we have that

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

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

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

$$\begin{aligned}
R_n(\theta, \hat{\theta}) &= (1/n) E\left\{ H_n^2\left(p_\theta, p_{\hat{\theta}}\right)\right\} \\
&\leq (1/n) \min_{\theta' \in \Gamma_n} \left\{ K(p_\theta, p_{\theta'}) + 2 \ \mathrm{pen}(\theta') \right\} \\
&= (1/n) \min_d \min_{\theta' \in \Gamma_n^{(d)}} \left\{ K(p_\theta, p_{\theta'}) + (2d)(1 + D/2) \log n \right\} \\
&\leq (1/n) \min_d \left\{ A' ||\mu - \tilde{\mu}^{(d),n,q}||_{\ell_2}^2 + (2d)(1 + D/2) \log n \right\} \ ,
\end{aligned}$$

where the last line follows from our bound on the Kullback-Leibler distance and the fact that the $\tilde{\theta}^{(d),n,q}$ corresponding to $\tilde{\mu}^{(d),n,q}$ can be no worse than the θ' that achieves the inner minimum in the second last line.

However, for sequences μ and μ' produced by average sampling functions h and h', 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 $(1/n)||\mu - \mu'||_{\ell_2}^2 \leq ||h - h'||_{L_2}^2$; see equation (27) of Kolaczyk & Nowak (2004). Therefore, using this fact and the assumption that G^{-1} is Lipschitz, we have that

$$(1/n)||\mu - \tilde{\mu}^{(d),n,q}||_{\ell_2}^2 \leq ||h - \tilde{h}^{(d),n,q}||_{L_2}^2 \leq A'||f - \tilde{f}^{(d),n,q}||_{L_2}^2 \ ,$$

for some constant A'.

Now write

$$||f - \tilde{f}^{(d),n,q}||_{L_2}^2 = ||(f - \tilde{f}^{(d)}) + (\tilde{f}^{(d)} - \tilde{f}^{(d),n}) + (\tilde{f}^{(d),n} - \tilde{f}^{(d),n,q})||_{L_2}^2 \ , \qquad \text{(A3)}$$

and apply the triangle inequality to the right-hand side of (A3). 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 quantisation. The magnitude of the cross-terms follow accordingly. Therefore, ignoring constants and terms of no consequence, we obtain a bound on our risk that behaves like $d^{-2\alpha} + (d/n)\log n$, which minimised with respect to d yields that

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

In the case of $\hat{\theta} = \hat{\theta}_{RDP}$, the definition of Γ_n is adjusted accordingly, the inequality (A1) 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.

References.

Antoniadis, A. & Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika* **88**, 805-20.

Antoniadis, A., Besbeas, P. & Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhya* A **63**, 309-27.

Barry, D. & Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *J. Am. Statist. Assoc.* **88**, 309-19.

Baraud, Y. (2000). Model selection for regression on a fixed design. *Prob. Theory Rel. Fields* **117**, 467-93.

Birgé, L. & Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329-75.

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

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

Donoho, D.L. (1993). Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data. In *Proc. Sympos. Appl. Math.* Vol. **47**, Ed. I. Daubechies, pp. 173-205. San Antonio, TX: American Mathematical Society.

Donoho, D.L. (1997). CART and best-ortho-basis: A connection. *Ann. Statist.* **25**, 1870-911.

Donoho, D.L., Dyn, N., Levin, D. & Yu, T.P.-Y. (2000). Smooth multiwavelet duals of Alpert bases by moment-interpolating refinement. *Appl. Comp. Harmonic Anal.* **9**, 166-203.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. & Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with Discussion). *J. R. Statist. Soc.* B **57**, 301-69.

Li, Q.J. & Barron, A.R. (2000). Mixture density estimation. In *Advances in Neural Information Processing Systems 12*, Ed. S.A. Solla, T.K. Leen and K.-R. Müller, pp. 279-285. Cambridge, MA: MIT Press.

Kohler, M. (1999). Nonparametric estimation of piecewise smooth regression functions. *Statist. Prob. Lett.* **43**, 49-55.

Kolaczyk, E.D. & Nowak, R.D. (2004). Multiscale likelihood analysis and complexity penalised estimation. *Ann. Statist.* **32**, xxx-yyy.

Norris, J.P., Nemiroff, R.J., Bonnell, J.T., Scargle, J.D., Kouveliotou, C., Paciesas, W.S., Meegan, C.A. & Fishman, G.J. (1996). Attributes of pulses in long bright gamma-ray bursts. *Astrophys. J.* **459**, 393 - 412.

PRANDONI, P. & VETTERLI, M. (1999). Approximation and compression of piecewise smooth functions. *Phil. Trans. R. Soc. Lond.* A **357**, 2573-91.

SARDY, S., ANTONIADIS, A. & TSENG, P. (2004). Automatic smoothing with wavelets for a wide class of distributions. *J. Comp. Graph. Statist.* **13**, xxx - yyy.

WILLETT, R. M., & NOWAK, R. D. (2003a). Platelets: A multiscale approach for recovering edges and surfaces in photon limited medical imaging. *IEEE Trans. Med. Im.* **22**, 332-350.

WILLETT, R. M. & NOWAK, R.D. (2003b). Multiscale density estimation. Rice University Technical Report TREE0303 (at http://cmc.rice.edu/docs/).

WIJERS, W. (1998). The burst, the burster and its lair. *Nature* **393**, 13.

YAJNIK, M., MOON, S., KUROSE, J. & TOWSLEY, D. (1999). Measurement and modeling of the temporal dependence in packet loss. In *Proc. 18th Annual Conf. IEEE Computer and Communications Societies (INFOCOM), New York, NY, pp. 345-53. IEEE.*

ZHANG, K., DENG, M., CHEN, T., WATERMAN, M. & FENGZHY, S. (2002). A dynamic programming algorithm for haplotype block partitioning. *Proc. Nat. Acad. Sci.* **11**, 7335-39.

CAPTIONS FOR FIGURES.

Figure 1. – Simulation results, showing performance of estimators based on the methods of recursive dyadic partitioning, RDP, and recursive partitioning, RP, described in this paper, and the wavelet-based methods of Antoniadis & Sapatinas (2002), AS, and Donoho (1993), D. For the Poisson sampling model, results from the estimator of Sardy et al. (2002), SAT, are also shown. Results for the Poisson sampling model are shown in (a) and (b), while those for the binomial sampling model are shown in (c) and (d); the true unknown function was the 'smooth' function of Antoniadis & Sapatinas in (a) and (c), and the 'burst' function in (b) and (d).

Figure 2. – Estimates of intensity for gamma-ray burst data from the BATSE instruments on board NASA's former Compton Gamma Ray Observatory. Time series were created by aggregating original photon arrival times, for photons with energies 25-55 keV energies, from burst #1425 into a total of $n = 256$ equispaced bins. Each plot compares a multiscale generalised linear model estimate, solid, and the estimate of Norris et al. (1996), dot-dashed. The multiscale generalised linear model estimator shown, $\hat{\theta}_{RP}$, is 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. – Estimate of packet loss rates for packet trace experiment of Yajnik et al. (1999). Packets, originally sent at 160 millisecond time intervals, were subsampled at 1000 millisecond time intervals. A multiscale generalised linear model was fitted to the observed fraction of packets lost, for each successive group of 300 subsampled packets. Shown is the estimate $\hat{\theta}_{RP}$, based on a binomial model and piecewise linear specification for the logit. The vertical, dotted lines denote the boundary points of the optimal recursive partition.

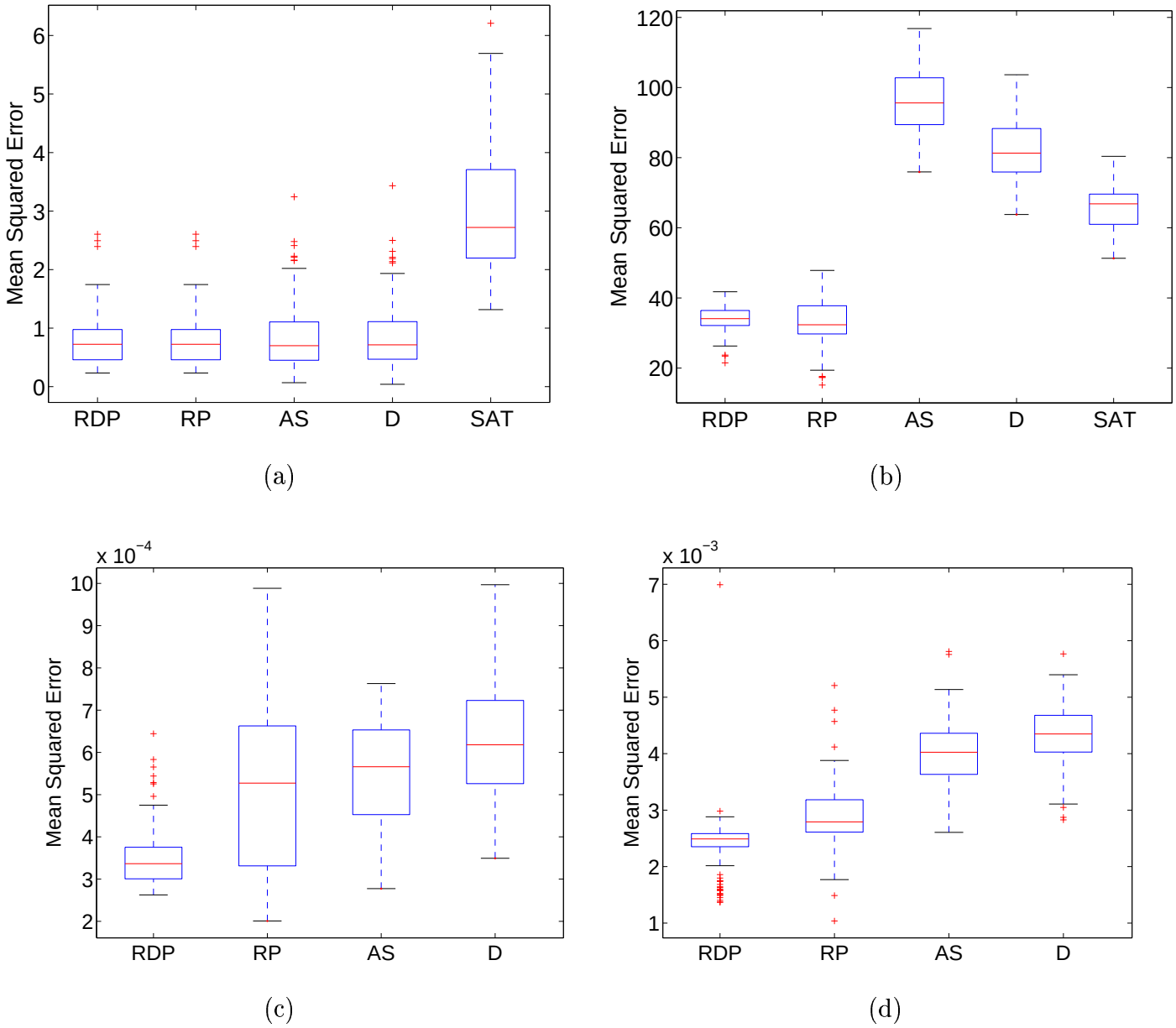
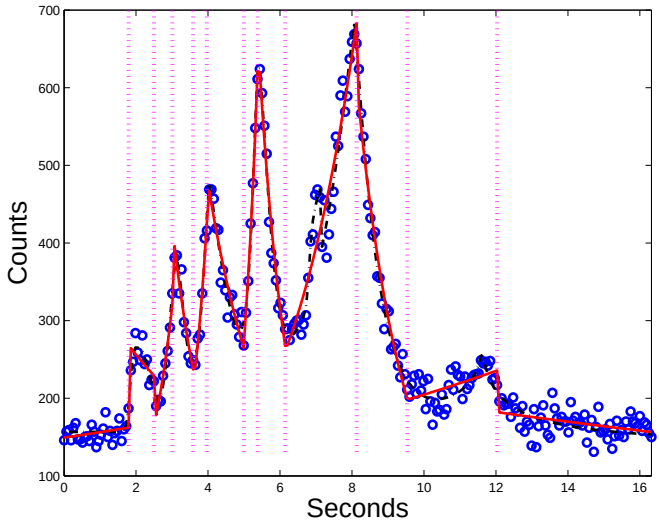
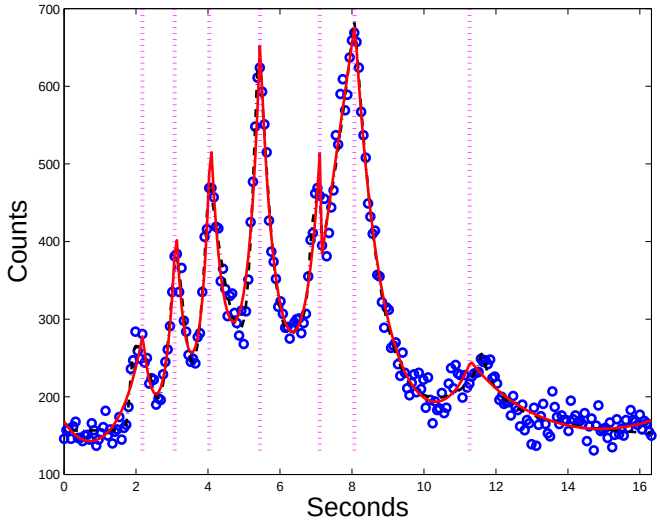


Figure 1:


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:

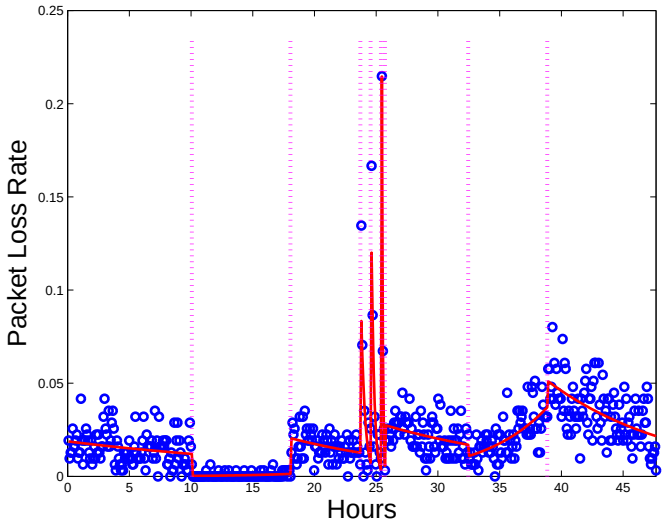

Packet Loss Rate
Hours
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
35
40
45

Figure 3: