MULTIRESOLUTION NONPARAMETRIC INTENSITY AND DENSITY ESTIMATION

Rebecca M. Willett and Robert D. Nowak

Department of Electrical and Computer Engineering
Rice University, 6100 S. Main, Houston, TX, 77005-1892 USA

ABSTRACT

This paper introduces a new multiscale method for nonparametric piecewise polynomial intensity and density estimation of point processes. Fast, piecewise polynomial, maximum penalized likelihood methods for intensity and density estimation are developed. The recursive partitioning scheme underlying these methods is based on multiscale likelihood factorizations which, unlike conventional wavelet decompositions, are very well suited to applications with point process data. Experimental results demonstrate that multiscale methods can outperform wavelet and kernel based density estimation methods.

1. POINT PROCESS ESTIMATION

In a variety of applications, data is acquired by observing the times or locations at which events occur. Frequently, either by choice or due to limitations of the measuring device, events are only observed on discrete intervals, resulting in a discrete signal representing the number of events occurring within each interval. Examples include the arrival of photons at a detector or observations of a random variable; we can estimate the intensity of photon emission or the probability density function by modeling these processes as Poisson and multinomial processes, respectively.

Wavelet-based methods are powerful tools for nonparametric signal denoising and have been used in applications of this nature; however, in the context of point process estimation, most wavelet-based approaches are based on Gaussian or other simplifying approximations to the Poisson or multinomial likelihood. The Haar wavelet system is the only exception; it does provide a tractable multiscale analysis framework for Poisson data, and this paper builds on the Haar-based multiscale likelihood factorizations we developed in [1]. Gaussian approximations are undesirable primarily because the approximations are usually only reasonable if the numbers of events occurring in each discrete interval is sufficiently large (so that the data, possibly after a suitable transformation, is roughly Gaussian distributed). To insure that the count levels are large, the detections must be binned or aggregated over sufficiently large intervals. Thus, one must immediately sacrifice resolution in order to accommodate the approximations. This runs counter to the entire philosophy of wavelet and multiscale methods which attempt to achieve some degree of resolution adaptivity in order to recover as much of the signal detail and structural nuances as possible from the data.

This paper introduces new multiscale methods for intensity and density estimation with piecewise polynomial approximation capabilities while maintaining the computational simplicity of wavelet methods. A key feature of our new approach is that it is nonparametric, meaning that no *a priori* limit is placed on the degrees of freedom used to describe the observed data. These methods constitute a nontrivial extension of the work done in [1], in which minimax optimal Haar wavelet-based methods restricted the intensity estimate to a piecewise-constant signal.

The paper is laid out as follows. Section 2 describes maximum likelihood estimation of polynomial fits to intensities and densities. Section 3 discusses multiscale analysis methods for Poisson and multinomial data and the notion of multiscale likelihood factorizations. We show that Poisson and multinomial likelihoods, parameterized by piecewise polynomial intensity/density functions, admit multiscale likelihood factorizations, which are probabilistic analogs of classical multiresolution analysis. Section 4 proposes a new penalized likelihood criterion for intensity and density estimation based on the multiscale factorization. The multiscale likelihood factorization enables a fast, globally optimal algorithm that computes the maximum penalized likelihood estimate (MPLE). Applications of the MPLE to network traffic analysis, gamma-ray burst data, and density estimation are demonstrated in Section 5. Section 6 summarizes the new methodology and discusses ongoing and future research directions.

2. MLE OF LINEAR SIGNAL PARAMETERS

Before introducing our multiscale intensity estimation algorithm, we describe here our method of calculating the maximum likelihood estimate of a polynomial (one interval) fit

R. Nowak was partially supported by the National Science Foundation, grant no. MIP–9701692, the Army Research Office, grant no. DAAD19-99-1-0349, the Office of Naval Research, grant no. N00014-00-1-0390. R. Willett was partially supported by the National Science Foundation Graduate Student Fellowship.

to Poisson and multinomial data. This technique extends the work in [2] and will be applied to every dyadic interval in the multiresolution algorithm.

The model for the intensity of the process is constrained such that $\boldsymbol{\mu} = T \cdot \boldsymbol{\theta}$, where T is a known Vandermonde matrix transforming the vector of polynomial coefficients to be estimated, $\boldsymbol{\theta}$, to the polynomial signal, $\boldsymbol{\mu}$. In the Poisson case, $\boldsymbol{\mu}$ would represent the Poisson intensity vector, while in the multinomial case, $\boldsymbol{\mu}$ would represent the probability vector and be constrained such that $\sum \boldsymbol{\mu} = 1$. Their exists no closed form solution for the MLE of the parameter vector $\boldsymbol{\theta}$; therefore a numerical solution must be calculated using a gradient or steepest descent algorithm. We have shown that this is a convex minimization problem [3], which demonstrates that an optimization algorithm will correctly identify a global maxima. The convexity of this optimization problem plays a key role in computational efficiency of the algorithm described below.

3. MULTISCALE ANALYSIS OF POINT PROCESSES

Now suppose that $x(u)$ is a realization of a Poisson or multinomial process. Underlying this process is an continuous intensity or density function $\mu(u)$, $(u) \in [0,1]$. Assume that either by choice or perhaps the limitations of measuring instruments, $x(u)$ is observed only discretely on the measurement intervals I_n, $n = 0, \ldots, N-1$. It is assumed that the effect of the discretization is to yield a vector of count measurements $\boldsymbol{x} \equiv \{x_n\}_{n=0}^{N-1}$, associated with an array of intensity parameters or multinomial probabilities $\boldsymbol{\mu} \equiv \{\mu_n\}_{n=0}^{N-1}$. Each x_n is simply the number of events in the interval I_n and $\mu_n \equiv \int_{I_n} \mu(u)$. The likelihood of $\boldsymbol{x}$, given the intensities or probabilities $\boldsymbol{\mu}$, is denoted by $p(\boldsymbol{x}|\boldsymbol{\mu})$.

A Haar multiscale analysis of the count data is obtained by associating a count statistic $x_{I_{n,j}} \equiv \sum_{k:(\frac{k}{N}) \in I_{n,j}} x_k$ with each dyadic interval $I_{n,j} \equiv [n/2^j, (n+1)/2^j)$, $j = 0, \ldots, J-1$, $n = 0, \ldots, 2^j - 1$, and $J = \log_2(N)$. The set of all dyadic intervals $\{I_{n,j}\}$ corresponds to a *complete* recursive dyadic partition (RDP) of $[0,1]$. There also exists a Haar multiscale analysis of the intensity/density function which is defined analogously on dyadic intervals. This RDP is called complete because all terminal nodes in the partition are intervals of width $1/N$ at the finest scale. An incomplete RDP would contain larger terminal intervals which could correspond to intervals of homogeneous or smoothly varying intensities. In earlier work, we introduced in this context a class of *multiscale likelihood factorizations* that provide an alternative probabilistic representation (i.e., in addition to that of the original likelihood) of the information in $\boldsymbol{x}$, in a manner indexed by the various time/scale combinations offered by a given RDP [1]. For a given RDP $\mathcal{P}$ the likelihood $p(\boldsymbol{x}|\boldsymbol{\mu})$ may be factorized as

$$
\begin{aligned}
& p(\boldsymbol{x}|\boldsymbol{\mu}(\mathcal{P},\boldsymbol{\theta})) = \\
& \quad p(x_{I_0}|\mu_{I_0}) \times \prod_{I \in NT(\mathcal{P})} p(\{x_{ch(I)}\}|x_I, \theta_I) \\
& \quad \times \prod_{ch(I) \in T(\mathcal{P})} p(\{x_n\}_{n/N \in ch(I)}|x_{ch(I)}, \theta_{ch(I)}), \quad (1)
\end{aligned}
$$

where $I_0 \equiv [0,1]$, $NT(\mathcal{P})$ is the set of all non-terminal intervals in $\mathcal{P}$, and $T(\mathcal{P})$ is the set of all terminal intervals in $\mathcal{P}$. The terminal node likelihood factors $p(\{x_n\}_{n/N \in ch(I)}|x_{ch(I)}, \theta_{ch(I)})$ are the multinomial likelihoods of the data in $ch(I)$ given a polynomial model $\theta_{ch(I)}$ of the intensity on $ch(I)$. Note that in this factorization the intensity is constrained to be piecewise polynomial on each interval in the partition, as indicated by the notation $\boldsymbol{\mu}(\mathcal{P},\boldsymbol{\theta})$. In Poisson processes, μ_{I_0} is a parameter to be estimated, while in multinomial processes, μ_{I_0} is known to be one.

The expression in (1) serves as a probabilistic analogue of an orthonormal wavelet decomposition of a function. The parameters of the conditional likelihoods play the same role as wavelet coefficients in a conventional wavelet-based multiscale analysis. The factorization in (1) can be shown to follow from a set of sufficient conditions whose form and function are remarkably similar to those of a Haar wavelet analysis – effectively a multiresolution analysis of the likelihood function. Details may be found in [1].

4. DENSITY AND INTENSITY ESTIMATION

4.1. Maximum Penalized Likelihood Estimation

The multiscale likelihood factorizations above provide for a very simple framework for maximum penalized likelihood estimation, wherein the penalization is based on the complexity of the underlying partition. The complexity of a given partition is proportional to the total number of intervals. Our goal here is to maximize the penalized likelihood function

$$L_\gamma(\boldsymbol{\mu}) \equiv \log p(\boldsymbol{x} \mid \boldsymbol{\mu}) - \gamma\{\#\boldsymbol{\theta}\}, \quad (2)$$

where $p(\boldsymbol{x} \mid \boldsymbol{\mu})$ denotes a likelihood (factorization) of the form (1) and $\{\#\boldsymbol{\theta}\}$ is the number of parameters in the vector $\boldsymbol{\theta}$ (one for each constant interval, M for each polynomial interval of degree M). The constant $\gamma > 0$ is a weight that balances between fidelity to the data (likelihood) and complexity regularization (penalty), which effectively controls the bias-variance trade-off.

The solution of

$$
\begin{aligned}
(\widehat{\mathcal{P}}, \widehat{\boldsymbol{\theta}}) &\equiv \arg\max_{\mathcal{P},\boldsymbol{\theta}} L_\gamma(\boldsymbol{\mu}(\mathcal{P},\boldsymbol{\theta})) \\
\widehat{\boldsymbol{\mu}} &\equiv \boldsymbol{\mu}(\widehat{\mathcal{P}}, \widehat{\boldsymbol{\theta}}) \quad (3)
\end{aligned}
$$

is called a maximum penalized likelihood estimator (MPLE). Larger values of γ produce smoother, less complex estimators; smaller values of γ produce more complicated estimators. The best overall performance (as measured by MSE) depends on the choice of γ. We have studied the performance of the MPLE in simulated intensity and density estimation experiments and investigated the effect of γ; our experiments reveal that $\gamma = \frac{1}{5}\log(\#\text{counts})$ consistently results in a low MSE over a broad range of intensity levels. There are also theoretical reasons for this form of penalization [1]. This setting for γ is used in all experiments described in this paper.

Maximizing (2) involves adaptively pruning the complete RDP based on the data. This pruning can be performed optimally and very efficiently as described in the next section. The pruning process is akin to a "keep or kill" wavelet thresholding rule. An MPLE provides higher resolution and detail in areas of the signal where there are dominant edges or singularities with higher count levels (higher SNR). The partition underlying the MPLE is pruned to a coarser scale (lower resolution) in areas with lower count levels (low SNR) and where the data suggest that the intensity is fairly smooth.

4.2. Optimal Pruning Algorithm

Observe that the structure of the penalized likelihood criterion stated in (2) and the likelihood factorization described in Section 3 allow an optimal intensity estimate to be computed quickly. The likelihood factorization allows the likelihood of the entire signal to be represented in a tree structure in which both likelihoods and parameter penalties of children are inherited by parents. Using this, it is possible to optimally prune an RDP of the data using a fast algorithm reminiscent of dynamic programming and the CART algorithm [1].

The goal is to estimate the intensity $\boldsymbol{\mu}$ according to (3). In order to perform the estimation, the algorithm considers each dyadic interval in the partition of the observation interval and performs an M-ary hypothesis test. The hypotheses for each dyadic interval are as follows:

- $\mathbf{H}_m$: Degree m ($m = 0, 1, 2, \ldots$,M) polynomially varying intensity segment (terminal node)
- $\mathbf{H}_{M+1}$: Inherit from children (non-terminal node)

When the maximum polynomial degree $M = 0$, the algorithm coincides with Haar analysis. It is also possible to consider a subset of $\{\mathbf{H}_i\}$; *e.g.* if $M = 2$, one might use only $\mathbf{H}_2$ and $\mathbf{H}_3$ to restrict the set of estimates to piecewise quadratic intensities. The algorithm begins one scale above the leaf nodes in the binary tree and traverses upwards, performing a tree-pruning operation at each stage. For each node (i.e., dyadic interval) at a particular scale, the maximum likelihood parameter vector is determined for each hypothesis as described in 2 and the penalized log likelihoods for each hypothesis are calculated. In particular, the penalized log likelihood for the split is computed using the optimal penalized log likelihoods computed at the previous, finer scale for both of the two children. The algorithm pseudocode is in Table 1. In the table, $L_\gamma(\boldsymbol{\theta}_{H_i}; I_{j,n})$ denotes the penalized log likelihood term for segment $I_{j,n}$ under hypothesis H_i.

Initialize:	$j = J - 1$
Loop:	for each node $I_{j,n}$ at level j
Calculate:	$L_\gamma(\boldsymbol{\theta}_{H_i}; I_{j,n})$ for $0 \le i \le M$ $L_\gamma(\boldsymbol{\theta}_{H_{M+1}}; I_{j,n}) = \sum_{I' \in ch(I_{j,n})} L_{min}(I')$
Save:	$L_{min}(I_{j,n}) = \min_{0 \le i \le M+1} L_\gamma(\boldsymbol{\theta}_{H_i}; I_{j,n})$ $\boldsymbol{\theta}_{min}(I_{j,n}) = \arg\min_{0 \le i \le M+1} L_\gamma(\boldsymbol{\theta}_{H_i}; I_{j,n})$
Coarsen:	Scale $j = j - 1$
Goto Loop:	if $j \ge 0$
Prune:	Perform a depth first search for terminal nodes. When a terminal node is found, record the MPLE for each terminal interval descending from the current node.

Table 1. Algorithm Pseudocode

5. APPLICATIONS

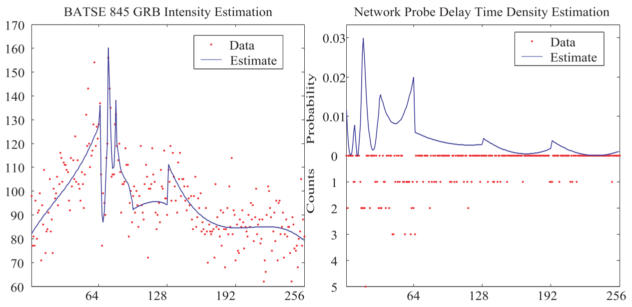
Two applications of multiscale MPLE are now explored, and density estimation capabilities are compared with those of wavelet-based methods and kernel methods.

5.1. Poisson Processes

One of the most fascinating problems in astrophysics today is the nature and origin of gamma ray bursts – quick, extremely intense bursts of gamma rays commonly associated with star formation and supernovae. Kolaczyk has done preliminary work in this field using Haar wavelets, an approach similar to the one described in this paper with the restriction that the intensity be piecewise constant [4]. Figure 1(a) demonstrates our algorithm's ability to produce an intensity estimate balanced between fidelity to the data and the underlying complexity. This estimate is much more faithful to the data than the piecewise constant Haar estimate in [4].

5.2. Density Estimation

Queuing delays are one of the most critical performance metrics in data networks. Accurate estimates of the queuing delay distribution can aid in optimizing communication network routing and service strategies. Here we apply our density estimation method to this problem. Queuing delay measurements were generated using the *ns*-2 network simulator. In the simulation, the queue buffer size was 35 packets and the traffic was a mixture of TCP and UDP flows. The estimate in Figure 1(b) below is plotted above a histogram of the measurements. Observe that the density estimate is smoothly varying and yet detects sharp peaks and discontinuities in the density.



(a) (b)

Fig. 1. Gamma Ray Burst and Network Queue Applications

5.3. Comparison with Wavelets

Our final experiment consisted of comparing the polynomial density estimation technique presented here with the commonly used normal kernel and wavelet methods. We consider three densities, generated from the well known test functions 'HeaviSine,' 'Bumps,' and 'Blocks' [5]. Probability mass functions (pmfs) of length 1024 were constructed from these test signals by shifting them to be strictly positive and normalizing them to one. This mimics a density estimation problem in which the measurement system has an accuracy of 10 bits. To simulate a set of observations from each density, approximately 1024 iid samples from each pmf were generated by a random number generator. Notice that the total number of samples is approximately the same as the dimension of the pmfs, simulating the ideal situation in which the data are not binned.

The densities were estimated with the MPLE algorithm described in this paper, the normal kernel method described in [6] (p. 56) and the wavelet hard-thresholding method described in [7]. Ten estimations were performed using the kernel method (with adaptively chosen optimal bandwidths for each observation), the wavelet thresholding method (using D6 wavelets with threshold levels chosen clairvoyantly in each simulation to obtain the best MSE), and our MPLE piecewise quadratic polynomial fits. In the wavelet thresholding case, the clairvoyant thresholds could not be obtained in practice, but here they provide a lower bound on the achievable MSE performance for any practical hard-thresholding scheme. The MSE of these estimates normalized to the MSE of the MPLE piecewise polynomial estimates are displayed in Table 2. Clearly, even without the benefit of setting the penalization factor clairvoyantly or data adaptively, the multiscale MPLE yields much smaller errors than the kernel and wavelet techniques for both smooth and spiky densities. Notably, unlike wavelet-based techniques, the polynomial technique is guaranteed to result in a non-negative density estimate.

	HeaviSine	Bumps	Blocks
Normal Kernel	1.24	1.98	1.23
Clairvoyant D6 Wavelet	2.06	1.71	2.02
Multiscale MPLE	1	1	1

Table 2. Density Estimation MSE, Normalized to MSE of Multiscale MPLE Estimate

6. CONCLUSIONS AND ONGOING WORK

This paper introduced a multiresolution nonparametric algorithm for intensity and density estimation. Our technique outperforms conventional kernel and wavelet-based methods because of its ability to efficiently represent detailed structure and its accurate (non-Gaussian) model of the data.

It should be possible to quantify the theoretical error performance of the multiresolution polynomial MPLE very precisely. In [1] it is shown that the Haar-based MPLE is near minimax optimal when the underlying Poisson intensity belongs to Bounded Variation or Besov function spaces. Our interest in this work is piecewise polynomial estimates in contrast to the piecewise constant estimates obtained by Haar analysis, and we expect that improved minimax bounds can be obtained for our likelihood-based MPLE in such cases. We are currently pursuing this work. In addition to this ongoing theoretical analysis, more experimental comparisons between this method and conventional methods are necessary.

7. REFERENCES

[1] E. Kolaczyk and R. Nowak, "Risk analysis for multiscale penalized maximum likelihood estimators," submitted to Annals of Stat. Available at http://www.ece.rice.edu/ nowak/publications.html.

[2] M. Unser and M. Eden, "Maximum likelihood estimation of linear signal parameters for poisson processes," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 6, pp. 942–5, 1988.

[3] R. Willett and R. Nowak, "Platelets: A multiscale intensity estimation of piecewise linear poisson processes," Tech. Rep. TREE0105, Rice University, 2001.

[4] Eric D. Kolaczyk, "Nonparametric estimation of gamma-ray burst intensities using haar wavelets," *The Astrophysical Journal*, vol. 483, pp. 340–349, 1997.

[5] David L. Donoho and Iain M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[6] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London, 1986.

[7] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Density estimation by wavelet thresholding," *Ann. Statist.*, vol. 24, pp. 508–539, 1996.