Multiscale Poisson Intensity and Density Estimation

R. M. Willett, *Member, IEEE,* and R. D. Nowak, *Senior Member, IEEE*

***Abstract*— The nonparametric Poisson intensity and density estimation methods studied in this paper offer near minimax convergence rates for broad classes of densities and intensities with arbitrary levels of smoothness. The methods and theory presented here share many of the desirable features associated with wavelet-based estimators: computational speed, spatial adaptivity, and the capability of detecting discontinuities and singularities with high resolution. Unlike traditional wavelet-based approaches, which impose an upper bound on the degree of smoothness to which they can adapt, the estimators studied here guarantee non-negativity and do not require any *a priori* knowledge of the underlying signal's smoothness to guarantee near-optimal performance. At the heart of these methods lie multiscale decompositions based on free-knot, free-degree piecewise-polynomial functions and penalized likelihood estimation. The degrees as well as the locations of the polynomial pieces can be adapted to the observed data, resulting in near minimax optimal convergence rates. For piecewise analytic signals, in particular, the error of this estimator converges at nearly the parametric rate. These methods can be further refined in two dimensions, and it is demonstrated that platelet-based estimators in two dimensions exhibit similar near-optimal error convergence rates for images consisting of smooth surfaces separated by smooth boundaries.**

Keywords: CART, complexity regularization, nonparametric estimation, piecewise polynomial approximation, platelets, wavelets

I. Density and Poisson Intensity Estimation

Poisson intensity estimation is a vital task in a variety of critical applications, including medical imaging, astrophysics, and network traffic analysis. Several multiresolution methods for estimating the time- or spatially-varying intensity of a Poisson process in these and other applications have been presented in the literature [1]–[3], generating wide interest [4]–[6]. Experimental results suggest that these methods can produce state-of-the-art results, but until now there has not been a thorough analysis of the theoretical underpinnings of these methods. This paper addresses this gap by casting the Poisson intensity estimation problem in a density estimation framework. Not only does this allow us to theoretically characterize multiscale methods for photon-limited imaging applications, but it also leads to a general framework for univariate and multivariate density estimation which both performs well in practice and exhibits several important theoretical properties. Accurate and efficient density estimation is often a fundamental first step in many applications, including source coding, data compression, statistical learning, and signal processing.

*Corresponding author: R. Willett. R. Nowak (nowak@engr.wisc.edu) was supported by the National Science Foundation, grants CCR-0310889 and ANI-0099148, and the Office of Naval Research grant N00014-00-1-0390. R. Nowak is with the Department of Electrical and Computer Engineering at the University of Wisconsin-Madison and R. Willett (willett@duke.edu) is with the Department of Electrical and Computer Engineering at Duke University.

The primary contributions of this paper are two-fold: (1) a theoretical characterization of photon-limited (Poisson) image processing tools, and (2) a data-adaptive multiscale density estimation method with several advantages over traditional wavelet-based approaches. These theoretical results will be supported with a number of experiments which demonstrate that our techniques can frequently outperform the best known wavelet-based techniques. The performance improvement is due to two key factors: (1) the ability of our method to adapt not only to singularities or discontinuities in the underlying intensity but also to arbitrary degrees of smoothness, and (2) the ability of our method to adapt to boundaries and edge structures in image data.

The approach studied in this paper involves using penalized likelihood estimation on recursive dyadic partitions in order to produce near-optimal, piecewise polynomial estimates, analogous to the methodologies in [7]–[9]. This results in a multiscale method that provides spatial adaptivity similar to wavelet-based techniques [10], [11], with a notable advantage. Wavelet-based estimators can only adapt to a function's smoothness up to the wavelet's number of vanishing moments; thus, some *a priori* notion of the smoothness of the true density or intensity is required in order to choose a suitable wavelet basis and guarantee optimal rates. The partition-based method, in contrast, automatically adapts to arbitrary degrees of the function's smoothness without any user input or *a priori* information. (Although the Meyer wavelet basis has infinitely many vanishing moments, its applications to density and intensity estimation on compact sets is unclear because the wavelets are defined in the frequency domain and have infinite time domain support.) Like wavelet-based estimators, the partition-based method admits fast estimation algorithms and exhibits near minimax optimal rates of convergence in many function spaces. The partition-based method has several additional advantages: estimates are guaranteed to be positive and the method exhibits rates of convergence within a logarithmic factor of the parametric rate for certain classes of densities and intensities. (While some methods (e.g. [12]) produce guaranteed positive density estimates by estimating the log-density, these methods are akin to fitting piecewise exponential functions to the density and hence are optimal for different classes of densities.) We elaborate on these points below.

While we focus on a particular class of problems in this paper, the ideas presented here are very general and simple to extend to other frameworks. For example, the partition-based technique could easily be used to find a piecewise polynomial estimate of the log of the density or intensity to form piecewise exponential estimates. The work in [13] extended the results

presented here and described in a technical report [14] to show that nonparametric estimation using generalized linear models in conjunction with the techniques described in this paper also results in nearly optimal rates of convergence for certain classes of functions.

A. Problem Formulation

The basic set-up considered in this paper is as follows. Assume a series of n independent and identically distributed observations, x_i, $i = 1, \ldots, n$ are made of a random variable, X, with density f^*. Let $\boldsymbol{x} \equiv \{x_i\}_{i=1}^n$. In this paper we consider penalized likelihood estimation, in which the density estimate is

$$\widehat{f} = \arg\min_{f \in \Gamma_n} L(f)$$

where Γ_n is a finite collection of candidate estimates,

$$L(f) \equiv -\log_e p_f(\boldsymbol{x}) + \mathrm{pen}(f), \tag{1}$$

and

$$p_f(\boldsymbol{x}) = \prod_{i=1}^n f(x_i)$$

denotes the likelihood of observing $\boldsymbol{x}$ if X had density f and where $\mathrm{pen}(f)$ is the penalty associated with a density f.

The methods presented in this paper are also applicable to estimating the temporally- or spatially-varying intensity of a Poisson process: both problems are concerned with estimating the distribution of events over some domain. The critical distinction between the two problems is that in density estimation, the density f^* is known to integrate to one, while in the Poisson case, there is no such constraint on the integral of the intensity. The number of observed events is random, with a mean equal to the integral of the intensity, and the mean must be estimated along with the distribution of events. In general, intensity estimation can be broken into two distinct subproblems: (1) estimation of the distribution of events, and (2) estimation of the integral of the intensity. The first subproblem is exactly the density estimation problem, and so everything said about density estimation above extends to Poisson intensity estimation. In the context of univariate Poisson intensity estimation, we let $\boldsymbol{x} = \{x_i\}_{i=1}^n$ be a series of n events, and let $x_i \in [0, 1]$ be the time or location of the i^{th} event. The underlying intensity is denoted by f^*, and the total intensity is denoted $I_{f^*} \equiv \int f^*(x)dx$.

Because of the close ties between Poisson intensity and density estimation and for simplicity of exposition, we focus on density estimation for most of this paper, and then explain the connections to and differences from Poisson intensity estimation in Section III-B.

B. Relation to Classical and Wavelet Density and Intensity Estimators

Classical nonparametric estimation techniques, e.g. kernel or histogram methods, have been thoroughly explored in the density estimation literature [15]–[21]. Most of the theoretical analysis associated with these methods pertains to linear estimators, which are known to be sub-optimal (in the sense of rates of convergence) for many classes of densities, e.g., Besov spaces [22]–[25]. In fact, is has been demonstrated that the L_1 error of non-negative, fixed-bandwidth kernel density estimators cannot exceed the rate of $n^{-2/5}$ (where n is the number of observations) for any density [16], [26]. Because linear estimators do not adapt to spatial changes in the structure of the data, their density estimates are in practice frequently oversmoothed where the density is changing rapidly or undersmoothed where the density is changing more slowly. Such estimators do not preserve singularities or sharp changes in the underlying density. Similar issues arise when using a single (not piecewise) polynomial for density estimation. Barron and Sheu [27] use Legendre polynomials to approximate the log of a density, resulting in a near minimax optimal exponential estimate when the log of the density is in a Sobolev space. The much larger class of densities in Besov spaces cannot be optimally estimated with their method due to its lack of spatial adaptivity. Spatially adaptive kernel methods [28]–[30], and wavelet-based density estimation techniques [22], [23] have been proposed to overcome such limitations; however, these methods generally require wavelets or kernels with more vanishing moments than degrees of density smoothness (e.g. the Besov smoothness parameter α in (7); this is explained in detail below); this limits the ability of these estimators to adapt to arbitrary degrees of smoothness. Histograms on data-dependent partitions also produce tractable, spatially adaptive density estimators, but while such estimators exhibit strong L_1 and L_2 consistency [31], [32], they can only achieve minimax rates of convergence for limited degrees of smoothness [33].

Wavelet-based techniques overcome this lack of spatial adaptivity because wavelets are well localized in both time and frequency and hence can provide good local estimates of the density. The estimation scheme presented by Donoho, Johnstone, Kerkyacharian, and Picard [23], is representative of many wavelet-based density estimators and summarized here in order to highlight its similarities to and differences from the partition-based in this paper. Any piecewise smooth density, $f(\cdot)$, such as one in a Besov space [24], [25], can be represented in terms of scaling and wavelet coefficients:

$$f(t) = \sum_k c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_k d_{j,k}\psi_{j,k}(t), \tag{2}$$

where $\phi_{j,k}$ is a scaling function and $\psi_{j,k}$ is a wavelet function, dilated to scale j and shifted by k units, and j_0 is the coarsest scale considered. In an orthogonal system, each wavelet coefficient is the inner product of the density and the wavelet function at a particular scale and shift, so if X is a random variable with density f, then we can express each coefficient as:

$$d_{j,k} = \int f(x)\psi_{j,k}(x)dx = \mathbb{E}\left[\psi_{j,k}(X)\right].$$

Thus a Monte Carlo estimate of each wavelet coefficient can be computed as

$$\widehat{d}_{j,k} = \frac{1}{n}\sum_{i=1}^n \psi_{j,k}(x_i),$$

where x_i is the i^{th} realization of X. Assuming that there are enough observations falling in the support of $\psi_{j,k}$, the central limit theorem can be invoked and $\widehat{d}_{j,k}$ can be assumed to be approximately Gaussian distributed with mean $d_{j,k}$ and some variance. In wavelet-based density estimation, the means of these empirical coefficients are improved using a hard or soft thresholding scheme based on the Gaussianity of the coefficients, and then the thresholded coefficients are used to synthesize the final density estimate. To guarantee that (on average) a sufficient number of samples fall within the support of each wavelet basis function to justify the Gaussian approximation, wavelet-based density estimates are restricted to scales no finer than $j = \log_2(n/\log_2 n)$.

Similar problems arise with classical and wavelet-based estimators in the context of Poisson intensity estimation. Statistical methods which account for the unique properties of the Poisson distribution can be effective [34]–[39], but are not well-suited for the detection of discontinuities or singularities. Wavelet-based techniques [40]–[47], designed for effective approximation of singularities are difficult to analyze in the presence of Poisson noise. Gaussian approximations are usually only appropriate when the number of events per interval or pixel is suitably large. This constraint is typically satisfied by binning observations until each interval or pixel contains a fairly large number of events; this process immediately limits the ultimate resolution the system can attain and any method's ability to reconstruct some fine scale structures.

C. Multiscale Partition-Based Estimators

Wavelet-based techniques are advantageous for both their near minimax convergence rates and the computational simplicity of filter-bank implementations. Near optimal convergence rates are possible as long as *a priori* knowledge of the density or intensity smoothness can be used to select a wavelet function ψ which is smooth enough (i.e., with a sufficient number of vanishing moments). The method introduced in this paper also admits a computationally efficient analysis and spatial adaptivity, but it exhibits the same convergence rates as wavelet-based techniques without any *a priori* upper bounds on smoothness. The partition-based method has two key additional benefits. First, the estimator always results in *bona fide* estimates (i.e. non-negative estimates which integrate to one). Second, we demonstrate that for piecewise analytic densities and intensities, the proposed free-knot, free-degree estimator results in near-parametric rates of convergence.

In our partition-based method, polynomials are fitted to a recursive dyadic partition (RDP) of the support of the density or the Poisson intensity. Our approach, based on complexity-regularization, is similar in spirit to the seminal work of Barron and Cover [48]. This work expands upon previous results (see, e.g., [49], [50], and [51]) by introducing an adaptivity to spatially varying degrees of smoothness. Barron *et al* [49] consider estimation of log densities and show that maximum penalized likelihood estimation using piecewise polynomials on regular partitions can result in a near minimax optimal estimator when the log density is in a Hölder smoothness class (a much more restrictive assumption than the Besov space considered in this paper [24]). Furthermore, the authors assume that the estimator uses polynomials with degree no less than the smoothness of the density. Castellan [50] and Reynaud-Bouret [51] independently addresses a problem similar to the one studied in this paper, but, like [49], only consider uniform partitions of the domain of the density; such partitions are not spatially adaptive and so cannot achieve optimal convergence rates for densities or log densities in Besov spaces. Nonuniform partitions are mentioned as a viable alternative in [50], but Castellan does not prove bounds associated with these partitions and does not propose a computationally tractable method for choosing the optimal nonuniform partition. This paper addresses these theoretical and practical challenges.

The RDP framework studied here leads to a model selection problem that can be solved by a tree pruning process. Appropriate pruning of this tree results in a penalized likelihood estimate of the signal as described in Section II. The main convergence results are summarized in Section III. Upper bounds on the estimation error (expected squared Hellinger distance) are established using several recent information-theoretic results, most notably the Li-Barron bound [52], [53] and a generalization of this bound [8]. We focus on multivariate density and Poisson intensity estimation in Section IV. A computationally efficient algorithm for computing piecewise polynomial estimates is presented and computational complexity is analyzed in Section V, and experimental results demonstrate the advantages of the partition-based approach compared to traditional wavelet-based estimators in Section VI. Section VII discusses some of the implications of our results and directions for future work.

II. Multiscale Density Estimation in One Dimension

The multiscale method presented here finds the optimal free-knot, free-degree piecewise polynomial density estimate using penalized likelihood estimation. The partition-based method determines the optimal partition of the interval $[0, 1]$ and optimal polynomial degree for each interval in the partition based on the observations; maximum likelihood polynomials of the optimal degree are then fit to the data on each interval. The optimal partition and polynomial degrees are selected using a simple framework of penalized likelihood estimation, wherein the penalization is based on the complexity of the underlying partition and the number of degrees of freedom in each polynomial.

The minimization is performed over a nested hierarchy of partitions defined through a recursive dyadic partition (RDP) of the unit interval, and the optimal partition is selected by optimally pruning a tree representation of the initial RDP of the data range. The effect of polynomial estimation on dyadic intervals is essentially an estimator with the same approximation capabilities as a wavelet-based estimator (for a wavelet with sufficiently many vanishing moments); this is established using approximation theoretic bounds in [25]. Thus there no disadvantage (in an approximation-theoretic sense) in using a piecewise polynomial basis instead of a wavelet basis.

As mentioned above, the piecewise polynomial multiscale analysis presented here is performed on recursive dyadic

partitions (RDPs) of the unit interval. The set of all intervals formed by recursively splitting the unit interval into equally sized regions until there are $2^{\lceil \log_2(n/\log_2 n)\rceil}$ regions with width no greater than $\log_2 n/n$ is referred to as the *complete* RDP (C-RDP). Any RDP can be represented with a binary tree structure. In general, the RDP framework can be used to perform model selection via a tree pruning process. Each of the terminal intervals in the pruned RDP corresponds to a region of homogeneous or smoothly varying density. Such a partition can be obtained by merging neighboring intervals of (i.e. pruning) a C-RDP to form a data-adaptive RDP $\mathcal{P}$ and fitting polynomials to the density on the terminal intervals of $\mathcal{P}$. Let $\boldsymbol{\theta}$ be a vector of polynomial coefficients for all of the intervals in $\mathcal{P}$. Note that some intervals of $\mathcal{P}$ may contain higher degree polynomials than others, so that the length of $\boldsymbol{\theta}$ may not be an integer multiple of the number of intervals in $\mathcal{P}$. Then any candidate density estimate is completely described by $\mathcal{P}$ and $\boldsymbol{\theta}$; i.e. $f = f(\mathcal{P}, \boldsymbol{\theta})$.

We penalize the piecewise polynomial estimates according to a codelength required to uniquely describe each such model (i.e., codes which satisfy the Kraft inequality). These codelengths will lead to near-minimax optimal estimators, as discussed in the next section. Because the proposed codelengths are proportional to the partition size and the number of polynomial coefficients associated with each model, penalization leads to estimates that favor fewer degrees of freedom. In particular, the penalty assigned to $f(\mathcal{P}, \boldsymbol{\theta})$ is

$$\mathrm{pen}(f(\mathcal{P}, \boldsymbol{\theta})) \equiv (2|\mathcal{P}| + |\boldsymbol{\theta}| - 1)\log_e 2 + \frac{|\boldsymbol{\theta}|}{2}\log_e n, \quad (3)$$

where $|\mathcal{P}|$ is the size of the RDP $\mathcal{P}$ (i.e. the number of terminal intervals) and $|\boldsymbol{\theta}| \equiv \|\boldsymbol{\theta}\|_{\ell_0}$ is the total number of polynomial coefficients in the vector $\boldsymbol{\theta}$. A detailed derivation of this penalty is in Appendix I. The penalty can be interpreted as a negative log-prior on the space of estimators. It is designed to give good guaranteed performance by balancing between fidelity to the data (likelihood) and the estimate's complexity (penalty), which effectively controls the bias-variance tradeoff. Since the penalty is proportional to $|\boldsymbol{\theta}|$, it facilitates estimation of the optimal polynomial degree on each interval of $\mathcal{P}$, leading to a "free-degree" piecewise polynomial estimate.

The solution of

$$(\widehat{\mathcal{P}}, \widehat{\boldsymbol{\theta}}) \equiv \underset{(\mathcal{P},\boldsymbol{\theta}):f(\mathcal{P},\boldsymbol{\theta})\in\Gamma_n}{\arg\min} L(f(\mathcal{P}, \boldsymbol{\theta})) \quad (4)$$

$$\widehat{f} \equiv f(\widehat{\mathcal{P}}, \widehat{\boldsymbol{\theta}}) \quad (5)$$

is called a penalized likelihood estimator (PLE). The collection of candidate estimates, Γ_n, is described in detail in Appendix I; it consists of all piecewise polynomial estimates, where the different polynomials are defined on the intervals of a RDP ($\mathcal{P}$), the polynomial coefficients ($\boldsymbol{\theta}$) have been quantized to one of $\sqrt{n}$ levels, and the resulting piecewise polynomial is non-negative and integrates to one. Section III demonstrates that this form of penalization results in near minimax optimal density estimates. Solving (4) involves adaptively pruning the C-RDP based on the data, which can be performed optimally and very efficiently. The pruning process is akin to a "keep or kill" wavelet thresholding rule. The PLE provides higher resolution and detail in areas of the density where there are dominant discontinuities or singularities with higher density. The partition underlying the PLE is pruned to a coarser scale (lower resolution) in areas with lower density and where the data suggest that the density is fairly smooth.

III. Error Analysis

In this section, we establish statistical risk bounds for free-degree piecewise polynomial estimation, as described above, and the resulting bound is used to establish the near-optimality of the partition-based estimation method. We then describe how these theoretical results can be applied to Poisson intensity estimation.

In this paper risk is defined to be proportional to the expected squared Hellinger distance between the true and estimated densities as in [48], [53]; that is,

$$\mathbb{E}\left[H^2(f^*, \widehat{f})\right] \equiv \mathbb{E}\left[\int\left(\sqrt{\widehat{f}} - \sqrt{f^*}\right)^2\right], \quad (6)$$

where the expectation is taken with respect to the observations. The squared Hellinger distance is an appropriate error metric here for several reasons. First, it is a general non-parametric measure appropriate for any density. In addition, the Hellinger distance provides an upper and lower bound on the L_1 error because of the relation $H^2(f_1, f_2) \le \int |f_1 - f_2| \le 2H(f_1, f_2)$ for all distributions f_1 and f_2 [16]. The L_1 metric is particularly useful for density estimation because of Scheffé's identity [16], which states that if $\mathcal{B}$ is the class of all Borel sets of $[0, 1]$, then

$$\sup_{B\in\mathcal{B}}\left|\int_B f_1 - \int_B f_2\right| = \frac{1}{2}\int |f_1 - f_2|.$$

Scheffé's identity shows that a bound on the L_1 error provides a bound on difference between the true probability measure and the density estimator's measure on every event of interest.

Lower bounds on the minimax risk decay rate have been established in [23]; specifically, consider densities in the Besov space

$$B_q^\alpha(L_p([0,1])) \asymp \Big\{f : \|c_{j_0,k}\|_{\ell_p} + \left(\sum_{j=j_0}^{\infty}\left(2^{\alpha j p}\sum_k |d_{j,k}|^p\right)^{q/p}\right)^{1/q} < \infty\Big\} \quad (7)$$

for $\alpha > 1/p \ge 1$, and $0 < q \le \infty$, where $\{c_{j_0,k}\}$ and $\{d_{j,k}\}$ are the scaling and wavelet coefficients in the wavelet expansion (2). Besov spaces are described in detail in [24], [25], and are useful for characterizing the performance of the proposed method because they include piecewise smooth densities which would be difficult to estimate optimally with classical, non-adaptive density estimation methods. The parameter α is the degree of smoothness (e.g. number of derivatives) of the functions in the space, p refers to the L_p space in which smoothness is measured, and q gives a more subtle measure

of smoothness for a given (α, p) pair. For these densities,

$$\begin{aligned}\inf_{\widehat{f}} \sup_{f^* \in B_q^\alpha(L_p([0,1]))} & \mathbb{E}\left[H^2(f^*, \widehat{f})\right] \\ \geq & \inf_{\widehat{f}} \sup_{f^* \in B_q^\alpha(L_p([0,1]))} \mathbb{E}\left[\frac{1}{4}\|\widehat{f} - f^*\|_{L_1}^2\right] \\ \geq & \; c n^{\frac{-2\alpha}{2\alpha+1}}\end{aligned}$$

for some $c > 0$ [23]. Likewise, the L_1 error is lower-bounded by $c' n^{\frac{-\alpha}{2\alpha+1}}$ for some $c' > 0$. We establish that the risk of the solution of (4) decays at a rate within a logarithmic factor of this lower bound on the rate.

A. Upper Bounds on Estimation Performance

Using the squared Hellinger distance allows us to take advantage of a key information-theoretic inequality derived by Li and Barron [52], [53] to prove the following main theorem:

Theorem 1 *Assume n samples are drawn from a density, f^*, which is a member of the Besov space $B_q^\alpha(L_p([0,1]))$ where $\alpha > 0$, $1/p = \alpha + 1/2$, and $0 \leq q \leq p$. Further assume that $0 < C_\ell \leq f \leq C_u < \infty$. Let $\widehat{f}$ be the free-degree penalized likelihood estimator satisfying (4) using the penalty in (3). Then*

$$\mathbb{E}\left[H^2(f^*, \widehat{f})\right] \leq C \left(\frac{\log_2^2 n}{n}\right)^{\frac{2\alpha}{2\alpha+1}} \tag{8}$$

for n sufficiently large and for some constant C that does not depend on n.

Theorem 1 is proved in Appendix I.

Remark 1 While the above theorem considers densities in a Besov space, it may be more appropriate in some contexts to assume that the density is in an exponential family and that the log of the density is in a Besov space (for examples, see [49], [50]). If desired, it is straightforward to adapt the method and analysis described in this paper to near optimal estimation of the log density.

Remark 2 The space of densities considered in the above theorem is quite general, and includes many densities for which optimal rates would not be achievable using nonadaptive kernel-based methods, such as a piecewise smooth (e.g. piecewise Hölder [24]) density with a finite number of discontinuities. Besov embedding theorems and other discussions on this class of densities can be found in [25] and [23].

Remark 3 The penalization structure employed here minimizes the upper bound on the risk. Furthermore, this upper bound is within a logarithmic factor of the lower bound on the minimax risk, demonstrating the near-optimality of the partition-based method, *even when α or an upper bound on α is unknown.*

Remark 4 The constant C in the above theorem and the proceeding theorems and corollaries is independent of n but still is a function of the "smoothness" of the class of densities under consideration. For example, in Theorem 1 it is related to the radius of the Besov ball in which f resides, in Example 1 below it is related to the number of pieces in a piecewise analytic function, and in Theorem 3 it is related to the Hölder exponents α and β. For ease of presentation, we state the bounds with constants, with the understanding that these constants depend on the function class under consideration, but we do not explicitly state this in each case.

The upper bound derived here is also within a logarithmic factor of the lower bound on the L_1 minimax error, as stated in the following corollary:

Corollary 1 *Let f^* and $\widehat{f}$ be defined as in Theorem 1. Then*

$$\mathbb{E}\left[\|\widehat{f} - f^*\|_{L_1}\right] \leq C \left(\frac{\log_2^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}}$$

for n sufficiently large and for some constant C that does not depend on n.

Corollary 1 is proved in Appendix II.

These results demonstrate the near-optimality of the penalization structure in (3) for free-degree piecewise polynomial estimation. In fact, as the smoothness of the density, α, approaches infinity, the asymptotic decay rate for this nonparametric method approaches the parametric rate of $1/n$. This can be made explicit for piecewise analytic densities, as in the following example:

Example 1 *Assume n samples are drawn from a piecewise analytic density with a finite number of pieces, f^*, such that $0 < C_\ell \leq f^*(\cdot) \leq C_u < \infty$. Let $\widehat{f}$ be the free-degree penalized likelihood estimator satisfying (4) using the penalty in (3). Then*

$$\mathbb{E}\left[H^2(f^*, \widehat{f})\right] \leq C \, \frac{\log_2^3 n}{n} \tag{9}$$

for n sufficiently large and some constant C.

For the piecewise analytic densities of the form in Example 1, the L_2 error of a free-knot, free-degree polynomial approximation with a total of m coefficients decays like 2^{-m}, and the variance of the estimator would decay like m/n because m coefficients must be estimated with n observations; balancing the approximation error with the estimation error leads to a total error decay of $(\log_2 n)/n$. The additional log terms are due to the recursive dyadic partition underlying the estimation method; a detailed derivation of the rate in Example 1 is provided in Appendix III.

B. Poisson Intensity Estimation

Recall that in Poisson intensity estimation, we let $\boldsymbol{x} = \{x_i\}_{i=1}^n$ be a series of n events, and let $x_i \in [0,1]$ be the time or location of the i^{th} event. The underlying intensity is denoted by f^*, and $I_{f^*} \equiv \int f^*(x)dx$. Using the above density estimation framework, it is possible to estimate the distribution

of events, $\widetilde{f}$, such that $\int \widetilde{f}(x)dx = 1$ and the maximum penalized likelihood intensity estimate is then $\widehat{f} \equiv n\widetilde{f}$; then

$$\mathbb{E}\left[H^2\left(\frac{\widehat{f}}{I_{\widehat{f}}}, \frac{f^*}{I_{f^*}}\right)\right] \leq C\ \left(\frac{\log_2^2 n}{n}\right)^{\frac{2\alpha}{2\alpha+1}}.$$

Since $\mathbb{E}[n] = I_{f^*}$, this renormalization generates an intensity estimate with overall intensity equal to the maximum likelihood estimate of I_{f^*}.

IV. Multidimensional Observations

In this section, we explore extensions of the above method to two-dimensional image estimation, particularly relevant in the context of Poisson intensity estimation, and multivariate estimation in higher dimensions.

A. Image Estimation

For the analysis in two dimensions, consider intensities which are smooth apart from a Hölder smooth boundary over $[0,1]^2$. Intensities of this form can be modeled by fusing two (everywhere) smooth intensities f_1 and f_2 into one single intensity according to

$$\begin{aligned} f(x,y) \quad = \quad & f_1(x,y)\cdot I_{\{y\geq H(x)\}} + \\ & f_2(x,y)\cdot\left(1 - I_{\{y\geq H(x)\}}\right), \end{aligned}$$

for all $(x,y) \in [0,1]^2$, where $I_{\{y\geq H(x)\}} = 1$ if $y \geq H(x)$ and 0 otherwise, and the function $H(x)$ describes a smooth boundary between a piece of f_1 and a piece of f_2. This is a generalization of the "Horizon" intensity model proposed in [54], which consisted of two constant regions separated by a smooth boundary. The boundary is described by $y = H(x)$, where

$$H \in \text{Hölder}_1^{\alpha}(C_\alpha), \quad \alpha \in (1,2], \quad C_\alpha > 0,$$

and $\text{Hölder}_1^{\alpha}(C_\alpha)$ for $\alpha \in (1,2]$ is the set of functions satisfying

$$\left|\frac{\partial}{\partial x}H(x_1) - \frac{\partial}{\partial x}H(x_0)\right| \leq C_\alpha |x_1 - x_0|^{\alpha-1},$$

for all $x_0, x_1 \in [0,1]$. For more information on Hölder spaces see [24].

The smoothness of the intensities f_1 and f_2 is characterized by a two-dimensional Hölder smoothness condition defined in [55]

$$f_i \in \text{Hölder}_2^{\beta}(C_\beta), \quad \beta \in (1,2], \quad C_\beta > 0, \quad i = 1,2,$$

where $\text{Hölder}_2^{\beta}(C_\beta)$ is the set of functions $f : [0,1]^2 \to \mathbb{R}^1$ with $k = \lfloor \beta \rfloor = 1$ continuous partial derivatives satisfying

$$|f(x_1) - p_{x_0}(x_1)| \leq C_\beta \left\|x_1 - x_0\right\|_2^{\beta},$$

for all $x_0, x_1 \in [0,1]^2$, where $p_{x_0}(x_1)$ is the Taylor polynomial of order k for $f(x_1)$ at point x_0.

The model describes a intensity composed of two smooth surfaces separated by a Hölder smooth boundary. This is similar to the "grey-scale boundary fragments" class of images defined in [55]. The boundary of the model is specified as a function of one coordinate direction (hence the name "Horizon"), but more complicated boundaries can be constructed with compositions of two or more Horizon-type boundaries, as in the following definition:

Definition 1 *Let $\mathcal{I}_{\alpha,\beta}$ denote the class of intensities $f : \Omega \to \mathbb{R}$ for $\Omega \subseteq [0,1]^2$ such that*

$$f(x,y) = f_1(x,y)\cdot I_{\{y\geq H(x)\}} + f_2(x,y)\cdot\left(1 - I_{\{y\geq H(x)\}}\right)$$

or

$$f(x,y) = f_1(x,y)\cdot I_{\{x\geq H(y)\}} + f_2(x,y)\cdot\left(1 - I_{\{x\geq H(y)\}}\right)$$

for all $(x,y) \in \Omega$ where $f_i \in \text{Hölder}_2^{\beta}(C_\beta)$, $i = 1,2$, and $H \in \text{Hölder}_1^{\alpha}(C_\alpha)$ with $\alpha, \beta \in (1,2]$. The class of piecewise (α,β)-smooth images *is the set of all images which can be written as a finite concatenation or superposition of $f \in \mathcal{I}_{\alpha,\beta}$.*

In [1], we introduced an atomic decomposition called "platelets", which were designed to provide sparse approximations for intensities in this class. Platelets are localized functions at various scales, locations, and orientations that produce piecewise linear two-dimensional intensity approximations. A wedgelet-decorated RDP, as introduced in [54], is used to efficiently approximate the boundaries. Instead of approximating the intensity on each cell of the partition by a constant, however, as is done in a wedgelet analysis, platelets approximate it with a planar surface. We define a platelet $f_S(x,y)$ to be a function of the form

$$f_S(x,y) = (A_S x + B_S y + C_S)\ I_S(x,y), \tag{10}$$

where $A_S, B_S, C_S \in \mathbb{R}$, S is a dyadic square or wedge associated with a terminal node of a wedgelet-decorated RDP, and I_S denotes the indicator function on S. Each platelet requires three coefficients, compared with the one coefficient for piecewise constant approximation. The dictionary is made discrete by quantizing both the platelet coefficients and the number of possible wedgelet orientations. A "resolution δ" approximation means that the spacing between possible wedgelet endpoints on each side of a dyadic square in $[0,1]^2$ is δ; see [54] for details.

The following theorem, which bounds the global squared L_2 approximation error of m-term platelet representations for intensities of this form, was proved in [1]:

Theorem 2 *Suppose that $2 \leq m \leq 2^J$, with $J > 1$. The squared L_2 error of an m-term, J-scale, resolution δ platelet approximation to a piecewise (α,β)-smooth image is less than or equal to $K_{\alpha,\beta} m^{-\min(\alpha,\beta)} + \delta$, where $K_{\alpha,\beta}$ depends on C_α and C_β.*

Theorem 2 shows that for intensities consisting of smooth regions ($\beta \in (1,2]$) separated by smooth boundaries ($\alpha \in (1,2]$), m-term platelet approximations may significantly outperform Fourier, wavelet, or wedgelet approximations. For example, if the derivatives in the smooth regions and along the boundary are Lipschitz ($\alpha, \beta = 2$, i.e., smooth derivatives), then the m-term platelet approximation error behaves like $O(m^{-2}) + \delta$, whereas the corresponding Fourier error behaves

like $O(m^{-1/2})$ and the wavelet and wedgelet errors behave like $O(m^{-1})$ at best. Wavelet and Fourier approximations do not perform well on this class of intensities due to the boundary. The reader is referred to [54], [56], [57] for the Fourier and wavelet error rates. Wedgelets can handle boundaries of this type, but produce piecewise constant approximations and perform poorly in the smoother (but non-constant) regions of intensities. Curvelets [56] offer another, in some ways more elegant, approach to the issue of efficient approximation of piecewise smooth images. However, while platelets and curvelets have the same approximation capabilities, platelets are much easier to apply in the context of Poisson imaging due to the fact that they're based on recursive dyadic partitions, just as tree-based methods offer several advantages over wavelets in the context of univariate intensity and density estimation.

As with the one-dimensional construction, we penalize the platelet estimates according to the codelength required to uniquely describe each model. The penalty assigned to $f(\mathcal{P}, \boldsymbol{\theta})$ is

$$\mathrm{pen}(f(\mathcal{P}, \boldsymbol{\theta})) = (7/3)|\mathcal{P}| \log_e 2 + (8/3)|\mathcal{P}| \log_e n. \quad (11)$$

The solution of (4), where $\mathcal{P}$ is a wedgelet-decorated RDP and $\boldsymbol{\theta}$ contains platelet coefficients is then the platelet penalized likelihood estimator. This construction can now be used to analyze platelet estimation error:

Theorem 3 *Assume n samples are drawn from a intensity, f^*, which is a piecewise (α, β)-smooth image. Further assume that $0 < C_\ell \leq f^* \leq C_u < \infty$. Let $\widehat{f}$ be the platelet estimator satisfying (4) using the penalty in (11). Then*

$$\mathbb{E}\left[H^2\left(\frac{\widehat{f}}{I_{\widehat{f}}}, \frac{f^*}{I_{f^*}}\right)\right] \leq C \left(\frac{\log_2^2 n}{n}\right)^{\frac{\min(\alpha,\beta)}{\min(\alpha,\beta)+1}} \quad (12)$$

for n sufficiently large and for some constant C that does not depend on n.

This is proved in Appendix IV. The denominators $I_{\widehat{f}}$ and I_{f^*} on the left hand side of the inequality normalize the intensities $\widehat{f}$ and f^*, respectively, so they both integrate to one. This rate is within a logarithmic factor of the minimax lower bound on the rate, $n^{-\min(\alpha,\beta)/(\min(\alpha,\beta)+1)}$; see [54], [55] for details.

B. Multivariate Estimation

The partition-based approach can easily be extended to multivariate estimation. We now assume that the true density is in a Hölder smoothness space because the relevance of singularities in multidimensional Besov spaces to practical problems is unclear. Specifically, information-bearing singularities in multiple dimensions, such as "ridges" or "sheets" have a much richer structure than one-dimensional singularities.

Assume that the true density $f : [0,1]^d \longrightarrow [C_\ell, C_u]$ is at least Hölder$_d^\alpha$ smooth everywhere. This condition means

$$|f(x_1) - p_{x_0}(x_1)| \leq C_\alpha \|x_1 - x_0\|_2^\alpha$$

for all $x_0, x_1 \in [0,1]^d$, where $p_{x_0}(x_1)$ is the k^{th}-order Taylor series polynomial expansion of $f(x)$ about x_0 evaluated at $x = x_1$, and where $k = \lfloor \alpha \rfloor$. For this class of densities, wavelet-based approaches can achieve an error decay rate of $O((\log_2 n/n)^{2\alpha/(2\alpha+d)})$ if a wavelet with more than α vanishing moments is selected [55]. Similarly, the same rate is achievable with a multivariate extension of the partition-based method studied in this paper without any *a priori* knowledge of the underlying smoothness.

From the Hölder condition, it is straightforward to verify that an order-k piecewise polynomial would accurately approximate a function in this class. Next note that multivariate tree pruning can be implemented in practice using 2^d-ary trees instead of binary trees to build a recursive dyadic partition. The appropriate penalty is

$$\mathrm{pen}(f(\mathcal{P}, \boldsymbol{\theta})) \equiv \left(\frac{2^d|\mathcal{P}| - 1}{2^d - 1} + |\boldsymbol{\theta}|\right) \log_e 2 + \frac{|\boldsymbol{\theta}|}{2} \log_e n;$$

to see this, follow the derivation of the one-dimensional penalty in Appendix I and note that a 2^d-ary tree with $|\mathcal{P}|$ leafs would have a total of $(2^d|\mathcal{P}| - 1)/(2^d - 1)$ nodes. It is straightforward to demonstrate, using arguments parallel to the ones presented in the univariate case, that this leads to an error decay rate of $(\log_2^2 n/n)^{2\alpha/(2\alpha+d)}$ without any prior knowledge of α. This is within a logarithmic factor of the minimax rate.

This is particularly significant when estimating very smooth densities in multiple dimensions. For example, consider a multivariate Gaussian, which is infinitely smooth. Any wavelet-based approach will be unable to exceed the rate $(\log_2 n/n)^{2r/(2r+d)}$, where r is the number of vanishing moments of the wavelet; kernel-based methods will also have a convergence rate limited by the bandwidth of the kernel. In contrast, the partition-based method will approach the parametric rate of $1/n$. We are unaware of any alternative nonparametric method with this property.

V. Algorithm and Computational Complexity

The previous sections established the near-optimality of the partition-based method using information theoretic arguments to bound the statistical risk. This section demonstrates that the partition-based estimator can be computed nearly as computationally efficiently as a traditional wavelet-based estimator in addition to having the theoretical advantages discussed in the previous sections.

A. Algorithm

Observe that the structure of the penalized likelihood criterion stated in (1) and the RDP framework allow an optimal density estimate to be computed quickly using a fast algorithm reminiscent of dynamic programming and the CART algorithm [7], [9]. This reduces the large optimization problem of computing the optimal free-degree, free-knot polynomial $\widehat{f}$ to a series of smaller optimization problems over disjoint intervals. The density f^* is estimated according to (4) with an algorithm which iterates from bottom to top through each level of the C-RDP of the observations. At each level, a multiple hypothesis test is conducted for each of the nodes at that level. The hypotheses for the node associated with interval I are as follows:

- $\boldsymbol{H}_{q_I}$ (terminal node): Order q_I ($q_I = 1, 2, \ldots, n_I$) polynomially varying segment which integrates to 1 on I, where $n_I \equiv \sum_{i=1}^{n} 1_{\{x_i \in I\}}$ is the number of observations falling in the interval I.
- $\boldsymbol{H}_{n_I+1}$ (non-terminal node): Concatenate optimal estimate of the left child, $\ell(I)$, scaled by $n_{\ell(I)}/n_I$ with the optimal estimate of the right child, $r(I)$, scaled by $n_{r(I)}/n_I$.

(Note that if we were to restrict our attention to polynomials of degree zero, the algorithm coincides with Haar analysis with a hereditary constraint [8].) 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 optimally determined for each hypothesis and the penalized log likelihoods for each hypothesis are calculated.

In particular, the penalized log likelihood for the split ($\boldsymbol{H}_{n_I+1}$) is computed using the optimal penalized log likelihoods computed at the previous, finer scale for both of the two children. To see the origin of the scaling factors $n_{\ell(I)}/n_I$ and $n_{r(I)}/n_I$, let $\widehat{f}_I$ be a density defined on I which minimizes $L(f(\mathcal{P}, \boldsymbol{\theta}))$ on the interval I, subject to the constraints $\int_I \widehat{f}_I = 1$ and $\widehat{f}_I > 0$. Note that $\widehat{f}_I$ can be computed independently of the observations which do not intersect I. Due to the additive nature of the penalized log likelihood function and the restriction of the estimator to a recursive dyadic partition, $\widehat{f}_I$ must either be a single polynomial defined on I or the concatenation of $\widehat{f}_{\ell(I)} a_{\ell(I)}$ and $\widehat{f}_{r(I)} a_{r(I)}$ for some positive numbers $a_{\ell(I)}$ and $a_{r(I)}$ which sum to one. A simple calculation reveals that $a_{\ell(I)} = n_{\ell(I)}/n_I$ and $a_{r(I)} = n_{r(I)}/n(I)$ minimize $L(f(\mathcal{P}, \boldsymbol{\theta}))$ over I subject to the given constraints.

The algorithm pseudocode is in Appendix V.

B. Computational Complexity

The partition-based method's overall computational complexity depends on the complexity of the polynomial fitting operation on each interval in the recursive dyadic partition. There is no closed-form solution to the MLE of the polynomial coefficients with respect to the likelihood; however, they can be computed numerically. The following lemma ensures that the polynomial coefficients can be computed quickly:

Lemma 1 *Assume a density, f, is a polynomial; that is, $f = T\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a vector containing the polynomial coefficients and T is a known linear operator relating the polynomial coefficients to the density. Denote the negative log likelihood of observing $\boldsymbol{x} \equiv \{x_i\}_{i=1}^n$ as $\ell_{\boldsymbol{x}}(\boldsymbol{\theta}) \equiv -\log_e p_{T\boldsymbol{\theta}}(\boldsymbol{x})$. Let Θ denote the set of all coefficient vectors $\boldsymbol{\theta}$ which result in a* bona fide *density. Then $\ell_{\boldsymbol{x}}(\boldsymbol{\theta})$ is a convex function on Θ, which is a convex set.*

Lemma 1 is proved in Appendix VI. Because $\ell_{\boldsymbol{x}}(\boldsymbol{\theta})$ is twice continuously differentiable and convex in the polynomial coefficients and the set of all admissible polynomial coefficients is convex, a numerical optimization technique such as Newton's method or gradient descent can find the optimal parameter values with quadratic or linear convergence rates, respectively. The speed can be further improved by computing Monte Carlo estimates of the polynomial coefficients to initialize the minimization routine. Specifically, if T_k is a k^{th}-order orthonormal polynomial basis function, then the optimal polynomial coefficient is

$$\int T_k(x) f(x) dx = \mathbb{E}[T_k],$$

which can be estimated as $(1/n)\sum_i T_k(x_i)$. In practice, we have found that computing such estimates with (appropriately weighted) Chebyshev polynomials is both very fast and highly accurate, so that calls to a convex optimization routine are often unnecessary in practice.

This lemma is a key component of the computational complexity analysis of the partition-based method. The theorem below is also proved in Appendix VI.

Theorem 4 *A free-degree piecewise polynomial PLE in one dimension can be computed in $O(n \log_2 n)$ calls to a convex minimization routine and $O(n \log_2 n)$ comparisons of the resulting (penalized) likelihood values. Only $O(n)$ log likelihood values and $O(n)$ polynomial coefficients need to be available in memory simultaneously. A platelet estimate of an image with n pixels can be calculated in $O(n^{4/3} \log_2 n)$ calls to a convex minimization routine.*

Note that the order of operations required to compute the estimate can vary with the choice of optimization method. Also, the computational complexity of the platelet estimator is based on the exhaustive search algorithm described in this paper, but recent work has demonstrated that more computationally efficient algorithms, which still achieve minimax rates of convergence, are possible [58].

VI. Simulation Results

The analysis of the previous sections demonstrates the strong theoretical arguments for using optimal tree pruning for multiscale density estimation. These findings are supported by numerical experiments which consist of comparing the density estimation techniques presented here with a wavelet-based method for both univariate density estimation and bivariate Poisson intensity estimation.

A. Univariate Estimation

Two test densities were used to help explore the efficacy of the proposed method. The first is a smooth Beta density: $f(x) = \beta(x; 2, 5)$, displayed in Figure 1(a). The second is a piecewise smooth mixture of beta and uniform densities designed to highlight the our method's ability to *adapt to varying levels of smoothness*:

$$\begin{aligned} f(x) \quad = \quad & \frac{3}{5}\left(\beta_{[0,\frac{3}{5}]}(x; 4, 4)\right) + \frac{1}{10}\left(\beta_{[\frac{2}{5},1]}(x; 4000, 4000)\right) \\ & + \frac{1}{40}\left(\mathrm{Unif}_{[0,1]}(x)\right) + \frac{11}{40}\left(\mathrm{Unif}_{[\frac{4}{5},1]}(x)\right), \end{aligned}$$

where $\beta_{[a,b]}$ refers to a Beta distribution shifted and scaled to have support on the interval $[a, b]$ and integrate to one. This

density is displayed in Figure 2(a). While the β distribution in particular could be very accurately estimated with a variety of methods designed for smooth densities, this experiment demonstrates that very accurate estimates of smooth densities are achievable by the proposed method *without* prior knowledge of the density's smoothness.

In each of one hundred experiments, an iid sample of one thousand observations was drawn from each density. The densities were estimated with the free-degree PLE method described in this paper (using only Monte Carlo coefficient estimates for speed), the wavelet hard- and soft-thresholding methods described in [23], and the wavelet block thresholding method described in [59]; Daubechies 8 wavelets were used for the second two methods. Like the method described in this paper, both of the wavelet-based approaches have strong theoretical characteristics and admit computationally fast implementations, although as described above, they have some limitations. The hard and soft wavelet threshold levels were chosen to minimize the average L_1 estimation error over the two distributions. (L_1 errors were approximated using discretized versions of the densities and estimates, where the length of the discrete vector, 2^{15}, was much greater than the number of observations, $1,000$.) A data-adaptive thresholding rule was proposed in [11], but the computational complexity of determining the threshold is combinatorial in the number of observations, which is impractical for large sets of observations. Furthermore, it entails either keeping or killing all wavelet coefficients on a single scale. This lack of spatial adaptivity could easily lead to poorer numerical results than the "clairvoyant" threshold weights used for this experiment. The clairvoyant thresholds used in this simulation could not be obtained in practice; in fact, the optimal threshold weights vary significantly with the number of observations. However, here they provide an empirical lower bound on the achievable MSE performance for any practical thresholding scheme. The MSE of these estimates are displayed in Table I. Clearly, even without the benefit of setting the penalization factor clairvoyantly or data adaptively, the multiscale PLE yields significantly lower errors than wavelet-based techniques for both smooth and piecewise smooth densities. Notably, unlike wavelet-based techniques, the polynomial technique is guaranteed to result in a non-negative density estimate. Density estimates can be viewed in Figures 1 and 2. Note that both the partition-based method and the wavelet-based methods result in artifacts for small numbers of observations. Piecewise polynomial estimates may have breakpoints or discontinuities at locations closely aligned with the underlying RDP. Wavelet-based estimates have negative segments and either undersmooth or oversmooth some key features; artifacts in all situations can be significantly reduced by cycle-spinning. This method can also be used effectively for univariate Poisson intensity estimations in applications such as network traffic analysis or Gamma Ray Burst intensity estimation, as demonstrated in [60].

Method	**Beta Density, Average L_1 Error**	**Mixture Density, Average L_1 Error**
Donoho et al, Hard Threshold, Clairvoyant Threshold [23]	0.1171	0.2115
Donoho et al, Soft Threshold, Clairvoyant Threshold [23]	0.1129	0.1968
Chicken and Cai, Hard Threshold, Clairvoyant Threshold [59]	0.1803	0.2855
Chicken and Cai, Soft Threshold, Clairvoyant Threshold [59]	0.1620	0.2638
Free Degree PLE, Theoretical Penalty	0.0494	0.1255

TABLE I
DENSITY ESTIMATION L_1 ERRORS.

B. Platelet estimation

In this section, we compare platelet-based Poisson intensity estimation with wavelet denoising of the raw observations and wavelet denoising of the Anscombe transform [61] of the observations. For this simulation, we assumed that observations could only be resolved to their locations on a 1024×1024 grid, as when measurements are collected by counting photons hitting an array of photo-multiplier tubes. An average of 0.06 counts were observed per pixel. The true underlying intensity is displayed in Figure 3(a), and the Poisson observations are displayed in Figure 3(b).

For each of the intensity estimation techniques shown here, we averaged over four shifts (no shift, $256/3$ in the vertical direction only, $256/3$ in the horizontal direction only, and $256/3$ in both the horizontal and vertical directions) to reduce the appearance of gridding artifacts typically associated with multiscale methods. The wavelet denoised image in Figure 3(c) was computed using a Daubechies 6 wavelet and a threshold was chosen to minimize the L1 error. The artifacts in this image are evident; their prevalence is intensity dependent because the variance of Poisson observations is equal to the intensity. The Anscombe transformed data ($y = 2(x+3/8)^{1/2}$, where x is a Poisson count statistic) was also denoised with Daubechies 6 wavelets (Figure 3(d)), again with a threshold chosen to minimize the L1 error. Here artifacts are no longer intensity dependent, because the Anscombe transform is designed to stabilize the variance of Poisson random variables. However, there are still distinct ringing artifacts near the high-contrast edges in the image. Furthermore, the overall intensity of the image is not automatically preserved when using the Anscombe transform ($\int \widehat{f}_{anscombe} \neq \sum_i x_i$), and important feature shared by the platelet- and wavelet-based methods.

We compared the above wavelet-based approaches with two RDP-based estimators: one composed of linear fits on the optimal rectangular partition (called the piecewise linear estimator), and one composed of linear fits on the optimal wedgelet partition (called the platelet estimator). Like the wavelet estimators, the piecewise linear estimator is unable to optimally adapt to image edges, as seen in Figure 3(e). However, comparing the images, we see that the piecewise linear estimator significantly outperforms the wavelet estimators. The wedgelet partition underlying the platelet estimator (Figure 3(f)), in contrast, is much better at recovering edges in the image and provides a marked improvement over the piecewise linear estimator. It is important to note that both the piecewise linear and platelet estimates were computed using

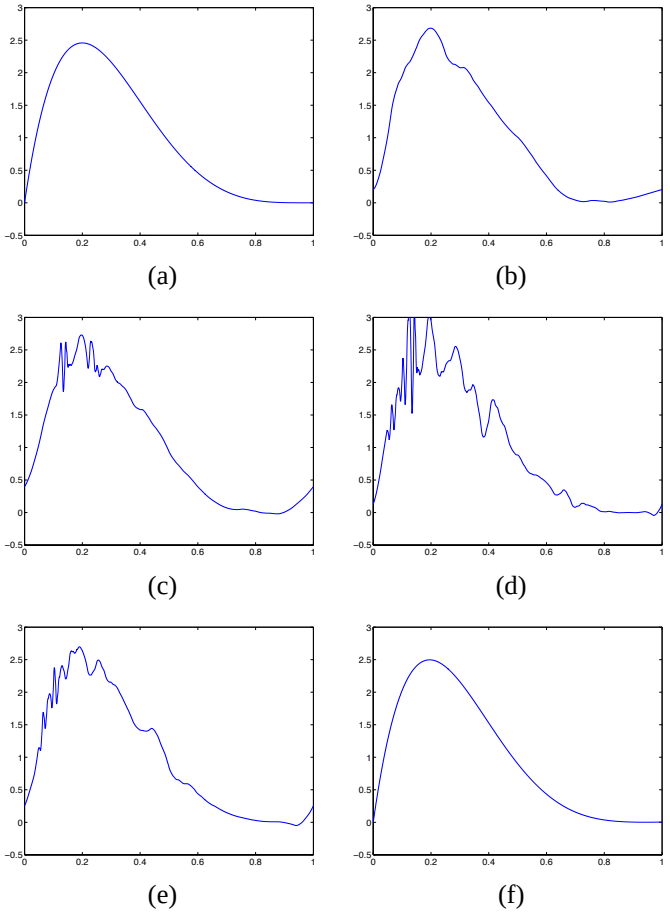

Fig. 1. Density estimation results for the Beta density. (a) True Beta density. (b) Wavelet estimate [23] with clairvoyant hard threshold; L_1 error $= 0.0755$. (c) Wavelet estimate [23] with clairvoyant soft threshold; L_1 error $= 0.0870$. (d) Wavelet estimate [59] with clairvoyant hard *block* threshold; L_1 error $= 0.1131$. (e) Wavelet estimate [59] with clairvoyant soft *block* threshold; L_1 error $= 0.0701$. (f) Free-degree estimate (with theoretical penalty); L_1 error = 0.0224

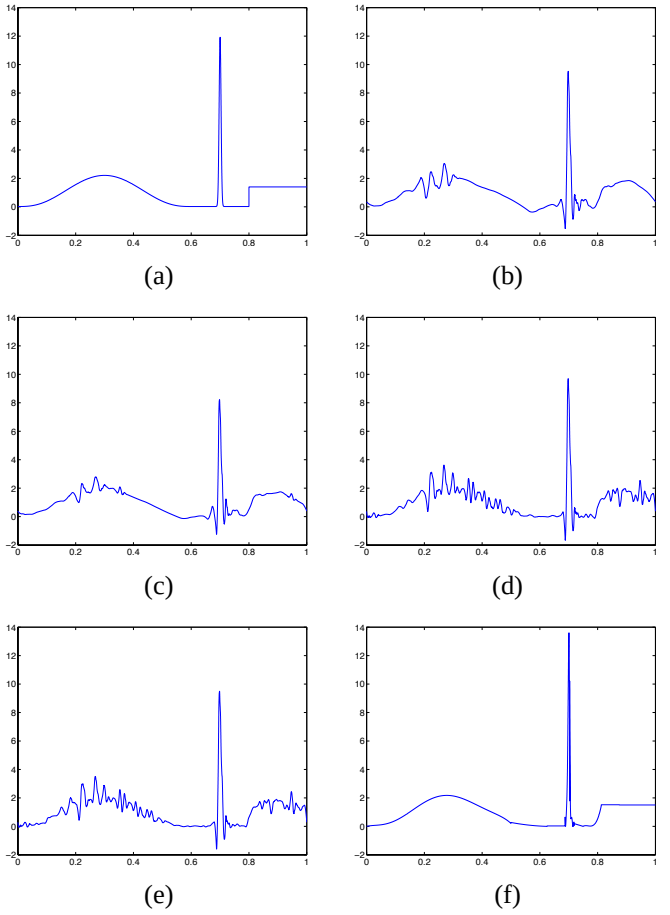

Fig. 2. Density estimation results for the mixture density. (a) True mixture density. (b) Wavelet estimate [23] with clairvoyant hard threshold; L_1 error $= 0.2806$. (c) Wavelet estimate [23] with clairvoyant soft threshold; L_1 error $= 0.2320$. (d) Wavelet estimate [59] with clairvoyant hard *block* threshold; L_1 error $= 0.2702$. (e) Wavelet estimate [59] with clairvoyant soft *block* threshold; L_1 error $= 0.2495$. (f) Free-degree estimate (with theoretical penalty); L_1 error = 0.1048

the theoretical penalties without the benefit of clairvoyant penalty weightings given to the wavelet-based estimates. Of course curvelets, mentioned in Section IV-A, also have the ability to adapt to edges in images; however, we anticipate that the platelet estimator would outperform the curvelet estimator for intensity estimation just as the piecewise linear estimator outperforms the wavelet-based estimates. Because of use of curvelets for intensity and density estimation is beyond the scope of this paper, we do not provide experimental curvelet results here.

VII. Conclusions and Ongoing Work

This paper studies methods for density estimation and Poisson intensity estimation based on free-degree piecewise polynomial approximations of functions at multiple scales. Like wavelet-based estimators, the partition-based method can efficiently approximate piecewise smooth functions and can outperform linear estimators because of its ability to isolate discontinuities or singularities. In addition to these features, the partition-based method results in non-negative density estimates and does not require any *a priori* knowledge of the density's smoothness to guarantee near optimal performance rates. Experimental results support this claim, and risk analysis demonstrates the minimax near-optimality of the partition-based method. In fact, the partition-based method exhibits near optimal rates for any piecewise analytic density regardless of the degree of smoothness; we are not aware of any other density estimation technique with this property.

The methods analyzed in this paper demonstrates the power of multiscale analysis in a more general framework than that of traditional wavelet-based methods. Conventional wavelets are effective primarily because of two key features: (1) adaptive recursive partitioning of the data space to allow analysis at multiple resolutions, and (2) wavelet basis functions that are blind to polynomials according to their numbers of vanishing moments. The alternative method presented here is designed to exhibit these same properties without retaining other wavelet properties which are significantly more difficult to analyze in the case of non-Gaussian data. Furthermore, in contrast to wavelet-based estimators, this method allows the data to adaptively determine the smoothness of the underlying density instead of forcing the user to select a polynomial order or

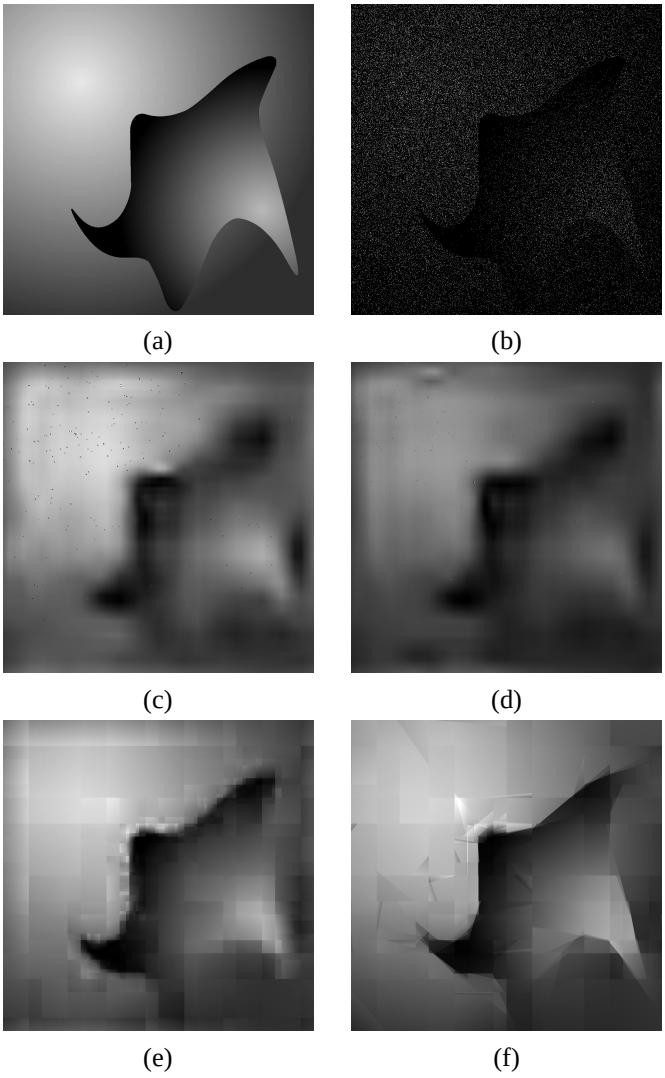

Fig. 3. Poisson intensity estimation. (a) True intensity. (b) Observed counts (mean = 0.06). (c) Wavelet denoised image, using $D6$ wavelets and a clairvoyant penalty to minimize the L1 error. Mean (per pixel) absolute error = $8.28e-3$. (d) Wavelet denoised image after applying the Anscombe transform, using $D6$ wavelets and a clairvoyant penalty to minimize the L_1 error. Mean absolute error = $2.00e-3$. (e) Piecewise linear estimate, using theoretical penalty. Mean absolute error = $4.47e-3$. (f) Platelet estimate, using theoretical penalty described in analysis above. Mean absolute error = $3.09e-3$.

wavelet smoothness. Because of their ability to adapt to smooth edges in images, platelet-based estimators also offer a notable advantage over traditional wavelet-based techniques; this is a critical feature for photon-limited imaging applications. These estimators have errors that converge nearly as quickly as the parametric rate for piecewise analytic densities and intensities.

As with wavelet-based and most other forms of multiscale analysis, the estimates produced by the partition-based PLE method commonly exhibit change-points on the boundaries of the underlying recursive dyadic partition. Because we only consider piecewise polynomials with first-order knots, and not splines, density estimates produced by the partition-based method often exhibit such discontinuities. Smoother estimates with the same theoretical advantages can be obtained through the use of Alpert bases [62] for moment interpolation as described by Donoho [63]. Fast, translation-invariant tree-pruning methods for first-order polynomials have been developed in [64]. Future work in multiscale density and intensity estimation includes the investigation of translation invariant methods for higher order polynomials.

Finally, note that in many practical applications, observations have been quantized by the measurement device, sometimes to such an extent that one can only observe binned counts of events. The effect of this binning or quantization is to limit the accuracy achievable by this or any other method. Nevertheless, the partition-based method studied in this paper can easily handle binned data to produce accurate estimates with near-optimal rates of convergence.

Appendix I
Proof of the Risk Bound Theorem

Proof of Theorem 1 The proof of this theorem consists of four steps. First, we will apply the Li-Barron theorem [53] to show that, if we consider all density estimates in a class Γ_n and if the penalties for each density in Γ_n satisfy the Kraft inequality, then

$$\mathbb{E}\left[H^2\left(\widehat{f},f^*\right)\right] \leq \min_{g\in\Gamma_n}\left\{K(f^*,g)+\frac{2}{n}\mathrm{pen}(g)\right\},$$

where

$$K(f^*,g)\equiv\int f^*\log_e\left(\frac{f^*}{g}\right)$$

denotes the Kullback-Leibler (KL) divergence between f^* and g. Second, we will verify that the proposed penalties satisfy the Kraft inequality. Third, we will upper bound the KL term, and finally, we will apply approximation-theoretic results to bound the risk.

The first step closely follows Kolaczyk and Nowak's generalization of the Li-Barron theorem [8], [53], but exhibits some technical differences because we consider continuous time (not discrete) densities.

Theorem 5 *Let Γ_n be a finite collection of estimators g for f^*, and $\mathrm{pen}(\cdot)$ a function on Γ_n satisfying the condition*

$$\sum_{g\in\Gamma_n} e^{-\mathrm{pen}(g)} \leq 1 \ . \tag{13}$$

Let $\widehat{f}$ be a penalized likelihood estimator given by

$$\widehat{f}(\boldsymbol{x}) \equiv \arg\min_{g\in\Gamma_n}\left\{-\log_e p_g(\boldsymbol{x})+2\mathrm{pen}\,(g)\right\}. \tag{14}$$

Then

$$\mathbb{E}\left[H^2\left(\widehat{f},f^*\right)\right] \leq \min_{g\in\Gamma_n}\left\{K(f^*,g)+\frac{2}{n}\mathrm{pen}(g)\right\} \ . \tag{15}$$

Remark 5 Minimizing over a finite collection of estimators, Γ_n, in (14) is equivalent to minimization over the finite collection of recursive partitions, $\mathcal{P}$, and coefficients, $\boldsymbol{\theta}$, described in (4) in Section II.

Remark 6 The first term in (15) represents the approximation error, or squared bias; that is, it is an upper bound on how

well the true density can be approximated by a density in the class Γ_n. The second term represents the estimation error, or variance associated with choosing an estimate from Γ_n given n observations. Both of these terms contribute to the overall performance of the estimator, and it is only by careful selection of Γ_n and the penalty function that we can ensure that the estimator achieves the target, near minimax optimal error decay rate.

Proof of Theorem 5 Following Li [52], define the affinity between two densities as

$$\mathcal{A}(f,g) \equiv \int (fg)^{1/2}.$$

Also, given a random variable X with density $f : [0,1] \longrightarrow [C_\ell, C_u]$, let $p_f : [0,1]^n \longrightarrow [C_\ell^n, C_u^n]$ denote the probability density function associated with drawing the n observations $\boldsymbol{x}$ from X. Then

$$\begin{aligned}\mathbb{E}\left[H^2(f^*,\widehat{f})\right] &= \mathbb{E}\left[2\left(1-\mathcal{A}(f^*,\widehat{f})\right)\right] \\ &\leq \mathbb{E}\left[-2\log_e \mathcal{A}(f^*,\widehat{f})\right] \\ &= \mathbb{E}\left[-\frac{2}{n}\log_e \mathcal{A}(p_{\widehat{f}}, p_{f^*})\right].\end{aligned}$$

From here it is straightforward to follow the proof of Theorem 7 in [8] to show

$$\begin{aligned}\mathbb{E}\left[H^2(f^*,\widehat{f})\right] &\leq \min_{g\in\Gamma_n}\left\{\frac{1}{n}K(p_{f^*},p_g)+\frac{1}{n}\mathrm{pen}(g)\right\} \\ &= \min_{g\in\Gamma_n}\left\{K(f^*,g)+\frac{1}{n}\mathrm{pen}(g)\right\}.\end{aligned}$$

■

We now define Γ_n as follows. First consider the collection of all free-knot, free-degree piecewise-polynomial functions which map $[0,1]$ to $[C_\ell, C_u]$ and which integrate to one. (Note that the knots in these densities will not normally lie on endpoints of intervals in the C-RDP, but rather within one of these intervals.) For each of these densities, shift each knot to the nearest dyadic interval endpoint, quantize the polynomial coefficients, clip the resulting function to be positive, and normalize it to integrate to one. This collection of densities constitutes Γ_n. We quantize the coefficients of an orthogonal polynomial basis expansion of each polynomial segment to one of $\sqrt{n}$ levels; this will be discussed in detail later in the proof. This definition of Γ_n allows us to prove the Kraft inequality when the penalty is defined as in (3):

Lemma 2 *Let $g \in \Gamma_n$, and let $\mathcal{P}$ denote the partition on which g is defined, and $\boldsymbol{\theta}$ be the vector of quantized polynomial coefficients defining g (prior to clipping and renormalization). If $\mathrm{pen}(g(\mathcal{P},\boldsymbol{\theta})) \equiv (2|\mathcal{P}|+|\boldsymbol{\theta}|-1)\log_e 2 + \frac{|\boldsymbol{\theta}|}{2}\log_e n$, then*

$$\sum_{g\in\Gamma_n} e^{-\mathrm{pen}(g)} \leq 1. \tag{16}$$

Proof of Lemma 2 Note that any $g \in \Gamma_n$ can be described by the associated quantized density (denoted g_q) prior to the deterministic processes of clipping and renormalization. Consider constructing a unique code for every g_q. If g_q consists of free-degree polynomials on each of $|\mathcal{P}|$ dyadic intervals, then both the locations of the $|\mathcal{P}|$ intervals and all the $|\boldsymbol{\theta}| < n$ coefficients need to be encoded. The $|\mathcal{P}|$ intervals can be encoded using $2|\mathcal{P}|-1$ bits. To see this, note that dyadic intervals can be represented as leaf nodes of a binary tree, and a binary tree with $|\mathcal{P}|$ leaf nodes has a total of $2|\mathcal{P}|-1$ nodes. Thus each node could be represented by one bit–a zero for an internal node and a one for a leaf node. This can easily be verified with an inductive argument.

The i^{th} of these $|\mathcal{P}|$ intervals, I_i, contains n_{I_i} observations, and the density on this interval is a polynomial of order r_i, $i = 1,\ldots,|\mathcal{P}|$, where $r_i \in \{1,\ldots,n_{I_i}\}$ and $\sum_i r_i = |\boldsymbol{\theta}|$. For the i^{th} interval, $\frac{r_i}{2}\log_2 n$ bits are needed to encode each quantized coefficient. These coefficients can be prefix encoded by following each encoded quantized coefficient with a single bit indicating whether all r_i coefficients have been encoded yet. A total of $|\boldsymbol{\theta}|$ of these indicator bits will be required. Thus the total number of bits needed to uniquely represent each $g \in \Gamma_n$ is $2|\mathcal{P}|-1+\sum_{i=1}^{|\mathcal{P}|}\left(\frac{r_i}{2}\log_2 n + r_i\right) = 2|\mathcal{P}| + |\boldsymbol{\theta}| - 1 + \frac{|\boldsymbol{\theta}|}{2}\log_2 n$.

We know that the existence of this uniquely decodable scheme guarantees that

$$\sum_{g\in\Gamma_n} 2^{-\left(2|\mathcal{P}|+|\boldsymbol{\theta}|-1+\frac{|\boldsymbol{\theta}|}{2}\log_2 n\right)} \leq 1.$$

Therefore, if $\mathrm{pen}(g) = (2|\mathcal{P}|+|\boldsymbol{\theta}|-1)\log_e 2 + \frac{|\boldsymbol{\theta}|}{2}\log_e n$, then

$$\begin{aligned}\sum_{g\in\Gamma_n} e^{-\mathrm{pen}(g)} &= \sum_{g\in\Gamma_n} 2^{-\log_2(e)\left((2|\mathcal{P}|+|\boldsymbol{\theta}|-1)\log_e 2+\frac{|\boldsymbol{\theta}|}{2}\log_e n\right)} \\ &= \sum_{g\in\Gamma_n} 2^{-\left(2|\mathcal{P}|+|\boldsymbol{\theta}|-1+\frac{|\boldsymbol{\theta}|}{2}\log_2 n\right)} \\ &\leq 1,\end{aligned}$$

as desired.

■

The next step in bounding the risk is to bound the KL divergence in (15).

Lemma 3 *For all densities $f : [0,1] \longrightarrow [C_\ell, C_u]$ and all $g \in \Gamma_n$,*

$$K(f,g) \leq \frac{1}{C_\ell}\|f-g\|_{L_2}^2.$$

Proof of Lemma 3

$$\begin{aligned}K(f,g) &= \int_0^1 f\log_e\left(\frac{f}{g}\right) \\ &\leq \int_0^1 f\left(\frac{f}{g}-1\right)+g-f \\ &= \int_0^1 \left(\frac{1}{g}\right)\left(g^2-2gf+f^2\right) \\ &\leq \frac{1}{C_\ell}\|f-g\|_{L_2}^2 \end{aligned} \tag{17}$$

where first inequality follows from $\log_e(z) \leq z-1$ and the second inequality follows from $g \geq 1/C_\ell$. ■

The above construction of Γ_n can be used to bound the approximation error $\|f-g\|_{L_2}^2$:

Lemma 4 *Let $f \in B_q^\alpha(L_p([0,1]))$, where $\alpha > 0$, $1/p = \alpha + 1/2$, and $0 < q \leq p$, be a density, let $g \in \Gamma_n$ be the best m-piece approximation to f, and let d denote the number of polynomial coefficients in this approximation. Then*

$$\|f - g\|_{L_2}^2 \leq C \left(m^{-\alpha} + \frac{m^{1/2}}{n^{1/2}} + \frac{d}{n^{1/2}} \right)^2 \qquad (18)$$

for n sufficiently large and for some constant C that does not depend on n.

Proof of Lemma 4 Using the construction of g outlined above and the triangle inequality, we have

$$\|f - g\|_{L_2} \leq \|f - g_p\|_{L_2} + \|g_p - g_s\|_{L_2} + \|g_s - g\|_{L_2}, \quad (19)$$

where g_p is the best free-knot, free-degree piecewise polynomial approximation of f, g_s is g_p after its knots have been shifted to the nearest dyadic interval endpoint, and g is g_s after the polynomial coefficients have been quantized, and the resulting function has been clipped and renormalized to produce a *bona fide* density.

These three terms can each be bounded as follows:

- $\|f - g_p\|_{L_2}$: The L_2 approximation error for either a m-piece free-degree piecewise polynomial approximation decays faster than $C_a m^{-\alpha}$ for some constant C_a which does not depend on m when $f \in B_q^\alpha(L_p([0,1]))$ [25].
- $\|g_p - g_s\|_{L_2}$: Because $f \leq C_u$ and f has compact support, we know $g_p < \infty$ and $g_s < \infty$. By construction, g_p has $m-1$ breakpoints, so for all but $m-1$ of the n intervals in the C-RDP, $g_p = g_s$. For the remaining $m-1$ intervals, each of length $1/n$, the L_∞ error is bounded by constant independent of m, leading to the bound

$$\|g_p - g_s\|_{L_2} \leq C_b \left(\frac{m-1}{n} \right)^{1/2} \qquad (20)$$

  where C_b is a constant independent of m and n.
- $\|g_s - g\|_{L_2}$: Quantization of each of the d polynomial coefficients produces the final error term. The polynomials can be expressed in terms of an orthogonal polynomial basis (e.g. the shifted Legendre polynomials), which allows the magnitudes of the coefficients to be bounded and hence quantized. Let T_I^k denote the k^{th}-order polynomial basis function on the interval I, so that $\langle T_I^\ell, T_I^k \rangle = 1_{k=\ell}$. Let $\theta_{i,k} = \langle g_s, T_{I_i}^k \rangle$. By the Cauchy-Schwarz inequality,

$$|\theta_{i,k}| \leq \|g_s\|_{L_2(I_i)} \|T_{I_i}^k\|_{L_2(I_i)}.$$

  Let $C_s = \sup_x g_s(x) < \infty$; then it is possible to quantize $\theta_{i,k}$ to one of $n^{1/2}$ levels in $\left[-C_s \|T_{I_i}^k\|_{L_2(I_i)}, C_s \|T_{I_i}^k\|_{L_2(I_i)}\right]$. Let the quantized version of coefficient $\theta_{i,k}$ be denoted $[\theta_{i,k}]$. This quantization results in the function g_q and induces the following error:

$$\begin{aligned}
\|g_s - g_q\|_{L_2} &= \sum_{i=1}^m \|g_s - g_q\|_{L_2(I_i)} \\
&= \sum_{i=1}^m \left[\sum_{k=1}^{r_i} (\theta_{i,k} - [\theta_{i,k}])^2 \|T_{I_i}^k\| \right]^{1/2} \\
&\leq \sum_{i=1}^m \left(\sum_{k=1}^{r_i} \frac{C_q}{n} \right)^{1/2} \\
&\leq \sum_{i=1}^m r_i \left(\frac{C_q}{n} \right)^{1/2} \\
&= C_c d / n^{1/2}
\end{aligned}$$

  for some constants C_q and C_c independent of g_s and g_q. Next, let g denote g_q after imposing the constraints that $\int g = 1$ and $g \geq 0$ by clipping and normalizing g_q. These operations do not increase the approximation error decay rate. For any density f and any function g, $\int |f - g| \geq \int |f - \max(g, 0)|$. In addition, for any density f and any non-negative function $g_q \geq 0$ such that $\int |f - g_q| < \epsilon$ for some $\epsilon < 1/2$, $\int |f - \frac{g_q}{\int g_q}| \leq 8\epsilon/3$ [31]. Set $\epsilon = C_a m^{-\alpha}$; then $\epsilon < 1/2$ for m sufficiently large. Thus $\|g_s - g\|_{L_2} \leq \|g_s - g_q\|_{L_2}$.

■

Finally, note that estimating densities on recursive dyadic partitions typically requires a larger number of polynomial pieces than free-knot approximation would require. The term $\|f - g\|_{L_2}^2$ was bounded assuming polynomial approximation was conducted on m (not necessarily dyadic) intervals. In practice, however, the binary tree pruning nature of the estimator would necessitate that any of the polynomial segments represented by g that do not lie on a dyadic partition be repartitioned a maximum of $\log_2 n$ times. This means that the best approximation to the density with m pieces and d coefficients must be penalized like a density with $|\mathcal{P}| = m \log_2 n$ pieces and $|\boldsymbol{\theta}| = d \log_2 n$ coefficients.

This, combined with the bounds in (15), (17), and (18), yield the bound

$$\begin{aligned}
&\mathbb{E}\left[H^2(f^*, \widehat{f}) \right] \\
&\leq \min_{g \in \Gamma_n} \left\{ \frac{1}{C_\ell} \|f^* - g\|_{L_2}^2 + \frac{2}{n} \mathrm{pen}(g) \right\} \\
&\leq \min_{m,d} \left\{ \begin{array}{l} \frac{1}{C_\ell} \left(C_a m^{-\alpha} + C_b \frac{m^{1/2}}{n^{1/2}} + C_c \frac{d}{n^{1/2}} \right)^2 + \\ \frac{2}{n} \left[(2m \log_2 n + d \log n - 1) \log_e 2 + \right. \\ \left. \quad \frac{d \log_2 n}{2} \log_e n \right] \end{array} \right\}.
\end{aligned}$$

Recalling that $m \leq d$, this expression is minimized for $d \sim \left(\frac{\log_2^2 n}{n} \right)^{\frac{-1}{2\alpha+1}}$. Substitution then yields that $\mathbb{E}\left[H^2(f^*, \widehat{f}) \right]$ is bounded above by $C \left(\frac{\log_2^2 n}{n} \right)^{\frac{2\alpha}{2\alpha+1}}$ for some constant C.

■

APPENDIX II
PROOF OF THE L_1 ERROR BOUND

Proof of Corollary 1 The risk bound of Theorem 1 can be

translated into an upper bound on the L_1 error between f^* and $\widehat{f}$ as follows. First note that $H^2(f^*,\widehat{f}) \leq \int |f^* - \widehat{f}| \leq 2H(f^*,\widehat{f})$ [16]. By Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}\left[\|f^* - \widehat{f}\|_{L_1}\right] &\leq 2\mathbb{E}\left[H(f^*,\widehat{f})\right] \\ &\leq 2\left(\mathbb{E}\left[H^2(f^*,\widehat{f})\right]\right)^{1/2} \\ &\leq C\left(\frac{\log_2^2 n}{n}\right)^{\frac{\alpha}{2\alpha+1}}.\end{aligned}$$

■

APPENDIX III
PROOF OF THE NEAR-PARAMETRIC RATES

Discussion of Example 1

The derivation of this rate closely follows the analysis of Theorem 1. Assume that f^* is composed of $1 \leq m < \infty$ analytic pieces, and the best free-knot, free-degree polynomial has a total of d coefficients. Then

$$\|f^* - g_p\|_{L_2} \leq C_a 2^{-d/m}.$$

This is a result of Jackson's Theorem V(iii) in [65]:

Theorem 6 *Let $f^* \in C[-1,1]$ and*

$$E_d(f^*) \equiv \inf_{a_0,\ldots,a_{d-1}} \sup_{-1\leq x\leq 1} \left| f^*(x) - \sum_{i=0}^{d-1} a_i x^i \right|.$$

If $f^{*(k)} \in C[-1,1]$ and $d \geq k$, then

$$E_d(f^*) \leq (\pi/2)^k \|f^{*(k)}\| \frac{(d-k+1)!}{(d+1)!}.$$

Applying Stirling's inequality and assuming $d = k \geq 8$, we have

$$E_d(f^*) \leq C_a' \|f^{*(k)}\| 2^{-d}.$$

For f^* with m analytic pieces, the minimax error in approximating f with piecewise polynomials with a total of d coefficients must decay *at least* as fast as $C_a'' 2^{-d/m}$. (Faster rates may be possible via a non-uniform distribution of the d coefficients over the m analytic pieces.) This results in the risk bound

$$\begin{aligned}&\mathbb{E}\left[H^2(f^*,\widehat{f})\right] \\ &\leq \min_{g\in\Gamma_n}\left\{\log_2(n)\|f^* - g\|_{L_2}^2 + \frac{2}{n}\mathrm{pen}(g)\right\} \\ &\leq \min_{m,d}\left\{\begin{array}{l}\frac{1}{C_\ell}\left(C_a 2^{-d/m} + C_b\frac{m^{1/2}}{n^{1/2}} + C_c\frac{d}{n^{1/2}}\right)^2 + \\ \frac{2}{n}\left[(2m\log_2 n - 1)\log_e 2 + \frac{d\log_2 n}{2}\log_e n\right]\end{array}\right\}.\end{aligned}$$

Recalling that $m \leq d$, this expression is minimized for $d = \log_2 n$. Substitution then yields that $\mathbb{E}\left[H^2(f^*,\widehat{f})\right]$ is bounded above by $C\frac{\log_2^3 n}{n}$ for some constant C.

■

APPENDIX IV
PROOF OF PLATELET ESTIMATION RISK BOUNDS

Proof of Theorem 3 This proof is highly analogous to the proof of Theorem 1 above, and so we simply highlight some of the most significant differences here.

First, a platelet estimate may be uniquely encoded with a prefix code (satisfying the Kraft inequality) as follows: for each (square- or wedgelet-decorated) leaf in the RDP, $7/3$ bits are needed to uniquely encode its location. To see this, let s denote the number of square-shaped leafs, and note that $s = 3k+1$ for some $k \geq 0$, where k is the number of interior nodes in the quad-tree representation of the RDP. This structure has a total of $4k+1$ nodes, and can be encoded using $4k+1$ bits. Next, each of the s square-shaped leafs may or may not be split into two wedgelet-shaped cells; these decisions can be encoded with a single bit, for a total of s additional bits. Thus, ignoring wedgelet orientations, the entire tree structure can be encoded using a total of $7k+2 < 7s/3$ bits. Let m denote the total number of square- or wedgelet-decorated leafs in the RDP; $s < m$, and so at most $7m/3$ bits can be used to encode the structure.

For each of the m cells in the partition, $8/3\log_2 n$ bits must be used to encode its intensity: $2/3\log_2 n$ bits for each of the three platelet coefficients, and $2/3\log_2 n$ bits to encode part of the wedgelet orientation. These numbers can be derived by noting that the best *quantized* m-term squared L_2 platelet approximation error behaves like $O(m^{-\min(\alpha,\beta)} + \delta + m^2/n^{2q})$, where n^q is the number of possible levels to which a platelet coefficient may be quantized and δ is the spacing between possible wedgelet endpoints. In order to guarantee that the risk converges at nearly the minimax rate of $n^{-2/3}$, δ must be set to $n^{-2/3}$ and q must be $2/3$. Then for any dyadic square contained in $[0,1]^2$, the total number of possible wedgelet orientations is no greater than $(1/\delta)^2 = n^{4/3}$. A single orientation can then be described using $4/3\log_2 n$ bits; each of the two wedgelets in a square-shaped region of the RDP is allotted half of these bits.

With this encoding scheme in mind, we set

$$\mathrm{pen}(f(\mathcal{P},\boldsymbol{\theta})) = (8/3)|\mathcal{P}|\log_e n + (7/3)|\mathcal{P}|\log_e 2.$$

This, combined with the bounds in (15), (17), and Theorem 2, yield the bound

$$\begin{aligned}&\mathbb{E}\left[H^2(f^*,\widehat{f})\right] \\ &\leq \min_{g\in\Gamma_n}\left\{\frac{1}{C_\ell}\|f^* - g\|_{L_2}^2 + \frac{2}{n}\mathrm{pen}(g)\right\} \\ &\leq \min_{\mathcal{P},\boldsymbol{\theta}}\left\{\begin{array}{l}\frac{1}{C_\ell}\left(C_a m^{-\min(\alpha,\beta)} + \right. \\ \quad \left. C_b n^{-2/3} + C_c\frac{m^2}{n^{4/3}}\right) + \\ \frac{2}{n}\left[(8/3)(m\log_2 n)\log_e n + \right. \\ \quad \left. (7/3)(m\log_2 n)\log_e 2\right]\end{array}\right\}.\end{aligned}$$

This expression is minimized for $m \sim \left(\frac{\log_2^2 n}{n}\right)^{\frac{-1}{\min(\alpha,\beta)+1}}$. Substitution then yields that $\mathbb{E}\left[H^2(f^*,\widehat{f})\right]$ is bounded above by $C\left(\frac{\log_2^2 n}{n}\right)^{\frac{\min(\alpha,\beta)}{\min(\alpha,\beta)+1}}$ for some constant C.

■

Loop: for $j = J$ downto 1, where J is the maximum depth of the C-RDP binary tree
  Loop: for each dyadic interval I at level j
    If $n_I == 0$:
      $\boldsymbol{\theta}_{\min}(I) = \mathbf{0}$
    Else:
      Loop: for $q = 1$ to n_I
        $\boldsymbol{\theta}_{\boldsymbol{H}_q} = \arg\min_{\boldsymbol{\theta}:|\boldsymbol{\theta}|=q} L(f(I, \boldsymbol{\theta}))$
      Goto loop: next q
      $q^* = \arg\min_{1 \le q \le n_I} L(f(I, \boldsymbol{\theta}_{\boldsymbol{H}_q}))$
      If $I \in \mathcal{T}(\mathrm{C-RDP})$:
        $\mathcal{P}_{\min}(I) = I$
        $\boldsymbol{\theta}_{\min}(I) = \boldsymbol{\theta}_{\boldsymbol{H}_{q^*}}$
      Else:
        $\mathcal{P}\text{no prune}(I) = \text{concat}(\mathcal{P}_{\min}(\ell(I)), \mathcal{P}_{\min}(r(I)))$
        $\boldsymbol{\theta}\text{no prune}(I) = \text{concat}(\boldsymbol{\theta}_{\min}(\ell(I)), \boldsymbol{\theta}_{\min}(r(I)))$
        If $L(f(\mathcal{P}\text{no prune}(I), \boldsymbol{\theta}\text{no prune}(I))) < L(f(I, \boldsymbol{\theta}_{\boldsymbol{H}_{q^*}}))$:
          $\mathcal{P}_{\min}(I) = \mathcal{P}\text{no prune}(I)$
          $\boldsymbol{\theta}_{\min}(I) = \boldsymbol{\theta}\text{no prune}(I)$
        Else:
          $\mathcal{P}_{\min}(I) = I$
          $\boldsymbol{\theta}_{\min}(I) = \boldsymbol{\theta}_{\boldsymbol{H}_{q^*}}$
        End if
      End if
    End if
  Goto loop: next node I at level j
Goto loop: next depth j
Estimate: $\widehat{f} = f(\mathcal{P}_{\min}([0,1]), \boldsymbol{\theta}_{\min}([0,1]))$

TABLE II
FREE-DEGREE PIECEWISE POLYNOMIAL ESTIMATION ALGORITHM PSEUDOCODE

APPENDIX V
ALGORITHM

Table II contains the algorithm pseudocode. In the pseudocode, $L(f(I, \boldsymbol{\theta}_{\boldsymbol{H}_r}))$ denotes the penalized log likelihood term for segment I under hypothesis $\boldsymbol{H}_q$, $\boldsymbol{\theta}(I)$ denotes the polynomial coefficients associated with interval I, and $\mathcal{T}(\mathrm{C-RDP})$ is the set of all intervals in the C-RDP corresponding to a terminal node (leaf) in the binary tree representation.

APPENDIX VI
PROOF OF COMPUTATIONAL COMPLEXITY LEMMA AND THEOREM

Proof of Lemma 1 If $\boldsymbol{\theta}$ is a vector of polynomial coefficients and $\boldsymbol{x}$ consists of n observations, then

$$\ell_{\boldsymbol{x}}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log_e \left(\sum_{k=0}^{|\boldsymbol{\theta}|-1} \theta_k x_i^k \right).$$

Let $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$ be two $|\boldsymbol{\theta}|$-dimensional vectors in Θ, and let $\theta_{a,k}$ and $\theta_{b,k}$ denote the k^{th} elements of $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$, respectively. Using the convexity of the negative log function, we have for all $0 \le \lambda \le 1$,

$$-\sum_{i=1}^{n} \log_e \left(\sum_{k=0}^{|\boldsymbol{\theta}|-1} \lambda\theta_{a,k} x_i^k + (1-\lambda)\theta_{b,k} x_i^k \right) \le$$
$$-\lambda \sum_{i=1}^{n} \log_e \left(\sum_{k=0}^{|\boldsymbol{\theta}|-1} \theta_{a,k} x_i^k \right)$$
$$-(1-\lambda) \sum_{i=1}^{n} \log_e \left(\sum_{k=0}^{|\boldsymbol{\theta}|-1} \theta_{b,k} x_i^k \right)$$

and hence $\ell_{\boldsymbol{x}}(\boldsymbol{\theta})$ is a convex function of $\boldsymbol{\theta}$.

To see that Θ is a convex set, consider two admissible coefficient vectors $\boldsymbol{\theta}_a$ and $\boldsymbol{\theta}_b$ defining two *bona fide* densities f_a and f_b, respectively. Then for any $\lambda < 1$ the density $f_c = \lambda f_a + (1-\lambda) f_b$ is also a *bona fide* density, and can be described by the coefficient vector $\boldsymbol{\theta}_c = \lambda\boldsymbol{\theta}_a + (1-\lambda)\boldsymbol{\theta}_b$ is also admissible. As a result, the set is convex. ■

Proof of Theorem 4 Recall that we start with $2^{\lceil \log_2(n/\log_2 n) \rceil} = O(n)$ terminal intervals in the C-RDP. Let n_I denote the number of observations in interval I. The tree-pruning algorithm begins at the leafs of the tree and progresses upwards. At the deepest level, the algorithm examines n pairs of intervals; for each interval I at this level, all of the k^{th}-order polynomial fits for $k = 1, \ldots, n_I$ are computed. This means that, at this level, a total of n polynomial fits must be calculated and compared. At the next coarser level, the algorithm examines $n/2$ intervals, and for each interval I at this level, all of the k^{th}-order polynomial fits for $k = 1, \ldots, n_I$ are computed, for a total of n polynomial fits which must be computed and compared. This continues for all levels of the tree, which means a total of $O(n \log_2 n)$ polynomial fits must be computed and compared. Further note that, at each level, only the optimal polynomial fit must be stored for each interval. Since there is a total of n intervals considered in the algorithm, only $O(n)$ likelihood values and polynomial coefficients must be stored in memory. ■

REFERENCES

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

[2] K. Timmermann and R. Nowak, "Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 846–862, April, 1999.

[3] R. Nowak and E. Kolaczyk, "A multiscale statistical framework for Poisson inverse problems," *IEEE Trans. Info. Theory*, vol. 46, pp. 1811–1825, 2000.

[4] J. Liu and P. Moulin, "Complexity-regularized denoising of poisson-corrupted data," in *International Conference on Image Processing*, 2000, vol. 3, pp. 254–257.

[5] T. Frese, C. Bouman, and K. Sauer, "Adaptive wavelet graph model for bayesian tomographic reconstruction," *IEEE Transactions on Image Processing*, vol. 11, no. 7, 2002.

[6] A. Willsky, "Multiresolution markov models for signal and image processing," *Proceedings of the IEEE*, vol. 90, no. 8, pp. 1396–1458, 2002.

[7] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1983.

[8] E. Kolaczyk and R. Nowak, "Multiscale likelihood analysis and complexity penalized estimation," to appear in Annals of Stat.. Available at http://www.ece.wisc.edu/∼nowak/pubs.html.

[9] D. Donoho, "Cart and best-ortho-basis selection: A connection," *Annals of Stat.*, vol. 25, pp. 1870–1911, 1997.

[10] D. Donoho, I. Johnstone, G. Kerkyacharian, and D. Picard, "Wavelet shrinkage: Asymptopia?," *Journal of the Royal Statistical Society*, pp. 301–337, 1995.

[11] G. Kerkyacharian, D. Picard, and K. Tribouley, "l_p adaptive density estimation," *Bernoulli*, vol. 2, pp. 229–247, 1996.

[12] J. Koo and W. Kim, "Wavelet density estimation by approximation of log-densities," *Statistics and Probability Letters*, vol. 26, pp. 271–278, 1996.

[13] E. Kolaczyk and R. Nowak, "Multiscale generalized linear models for nonparametric function estimation," submitted to Biometrica 2003. Available at http://www.ece.wisc.edu/∼nowak/pubs.html.

[14] R. Willett and R. Nowak, "Multiscale density estimation," Tech. Rep., Rice University, 2003, Available at http://www.ece.rice.edu/∼willett/papers/WillettIT2003.pdf.

[15] L. Devroye and L. Gy orfi, *Nonparametric Density Estimation: The L_1 View*, Wiley, New York, 1985.

[16] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, Ann Arbor, MI, 2001.

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

[18] D. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley, New York, 1992.

[19] B. Prakasa-Rao, *Nonparametric Functional Estimation*, Academic Press, Orlando, 1983.

[20] R. Eubank, *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York, 1988.

[21] R. Tapia and J. Thompson, *Nonparametric Probability Density Estimation*, John Hopkins University Press, Baltimore, 1978.

[22] A. Antoniadis and R. Carmona, "Multiresolution analyses and wavelets for density estimation," Tech. Rep., University of California, Irvine, 1991.

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

[24] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.

[25] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[26] L. Devroye and C. S. Penrod, "Distrubtion-free lower bounds in density estimation," *Annals of Statistics*, vol. 12, pp. 1250–1262, 1984.

[27] A. R. Barron and C.-H. Sheu, "Approximation of density functions by sequencyes of exponential families," *Annals of Statistics*, vol. 19, no. 3, pp. 1347–1369, 1991.

[28] O. V. Lepski, E. Mammen, and V. G. Spokoiny, "Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors," *The Annals of Statistics*, vol. 25, no. 3, pp. 929–947, 1997.

[29] J. Fan, I. Gijbels, T. Hu, and L. Huang, "A study of variable bandwidth selection for local polynomial regression," *Statistica Sinica*, vol. 6, pp. 113–127, 1996.

[30] L. Wasserman, *All of Nonparametric Statistics*, Springer, New York, 2006.

[31] G. Lugosi and A. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *Annals of Statistics*, vol. 24, no. 2, pp. 687–706, 1996.

[32] A. Nobel, "Histogram regression estimation using data-dependent partitions," *Annals of Statistics*, vol. 24, no. 3, pp. 1084–1105, 1996.

[33] J. Klemelä, "Multivariate histograms with data-dependent partitions," 2001, `http://www.vwl.uni-mannheim.de/mammen/klemela/`.

[34] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-d Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Trans. Med. Imaging*, vol. 8, no. 2, pp. 194–202, 1989.

[35] P. J. Green, "Bayesian reconstruction from emission tomography data using a modified EM algorithm," *IEEE Trans. Med. Imaging*, vol. 9, no. 1, pp. 84–93, 1990.

[36] J. A. Fessler and A. O. Hero, "Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms," *IEEE Trans. Image Processing*, vol. 4, no. 10, pp. 1417–1429, 1995.

[37] A. R. Depierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Trans. Med. Imaging*, pp. 132–137, 1995.

[38] J. Liu and P. Moulin, "Complexity-regularized image denoising," *IEEE Transactions on Image Processing*, vol. 10, no. 6, pp. 841 –851, 2001.

[39] P. Moulin and J. Liu, "Statistical imaging and complexity regularization," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1762 –1777, 2000.

[40] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 1998.

[41] J. Starck, F. Murtagh, and A. Bijaoui, *Image Processing and Data Analysis: The Multiscale Approach*, Cambridge Univ. Press, 1998.

[42] A. Aldroubi and M. Unser, "Wavelets in medicine and biology," CRC Pr., Boca Raton FL, 1996.

[43] M. Bhatia, W. C. Karl, and A. S. Willsky, "A wavelet-based method for multiscale tomographic reconstruction," *IEEE Trans. Med. Imaging*, vol. 15, no. 1, pp. 92–101, 1996.

[44] E. D. Kolaczyk, "Wavelet shrinkage estimation of certain Poisson intensity signals using corrected thresholds," *Statistica Sinica*, vol. 9, pp. 119–135, 1999.

[45] N. Lee and B. J. Lucier, "Wavelet methods for inverting the radon transform with noisy data," *IEEE Trans. Image Proc.*, vol. 10, pp. 79 –94, 2001.

[46] J. Lin, A. F. Laine, and S. R. Bergmann, "Improving pet-based physiological quantification through methods of wavelet denoising," *IEEE Trans. Bio. Eng.*, vol. 48, pp. 202 –212, 2001.

[47] J. Weaver, Y. Xu, D. Healy, and J. Driscoll, "Filtering MR images in the wavelet transform domain," *Magn. Reson. Med.*, vol. 21, pp. 288–295, 1991.

[48] A. Barron and T. Cover, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1034–1054, 1991.

[49] A. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probability Theory and Related Fields*, vol. 113, pp. 301–413, 1999.

[50] G. Castellan, "Density estimation via exponential model selection," *IEEE Transactions on Information Theory*, vol. 49, no. 8, pp. 2052–2060, 2003.

[51] P. Reynaud-Bouret, "Adaptive estimation of the intensity of inhomogeneous poisson processes via concentration inequalities," *Probab. Theory Relat. Fields*, vol. 126, pp. 103–153, 2003.

[52] Q. Li, *Estimation of Mixture Models*, Ph.D. thesis, Yale University, 1999.

[53] Q. Li and A. Barron, *Advances in Neural Information Processing Systems 12*, chapter Mixture Density Estimation, MIT Press, 2000.

[54] D. Donoho, "Wedgelets: Nearly minimax estimation of edges," *Ann. Statist.*, vol. 27, pp. 859 – 897, 1999.

[55] A. P. Korostelev and A. B. Tsybakov, *Minimax theory of image reconstruction*, Springer-Verlag, New York, 1993.

[56] E. Candès and D. Donoho, "Curvelets: A surprisingly effective nonadaptive representation for objects with edges," To appear in Curves and Surfaces, L. L. Schumaker et al. (eds), Vanderbilt University Press, Nashville, TN.

[57] D. Donoho, "Sparse components of images and optimal atomic decompositions," *Constr. Approx.*, vol. 17, pp. 353–382, 2001.

[58] R. Castro, R. Willett, and R. Nowak, "Coarse-to-fine manifold learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing — ICASSP '04*, 17-21 May, Montreal, CA, 2004.

[59] E. Chicken and T. Cai, "Block thresholding for density estimation: local and global adaptivity," *Journal of Multivariate Analysis*, vol. 95, pp. 76–106, 2005.

[60] R. Willett and R. Nowak, "Multiresolution nonparametric intensity and density estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing — ICASSP '02*, 13-17 May 2002, Orlando, FL, USA, 2002.

[61] F. J. Anscombe, "The tranformation of poisson, binomial, and negative-binomial data," *Biometrika*, vol. 15, pp. 246–254, 1948.

[62] B. Alpert, "A class of bases in l2 for sparse representation of integral operators," *SIAM Journal of Mathematical Analysis*, vol. 24, pp. 246–262, 1993.

[63] D. Donoho, N. Dyn, D. Levin, and T. Yu, "Smooth multiwavelet duals of alpert bases by moment-interpolation, with applications to recursive partitioning," Tech. Rep., Department of Statistics, Stanford University, 1996.

[64] R. Willett and R. Nowak, "Fast multiresolution photon-limited image reconstruction," in *Proc. IEEE Int. Sym. Biomedical Imaging — ISBI '04*, 15-18 April 2004, Arlington, VA, USA, 2004.

[65] E. W. Cheney, *Introduction to Approximation Theory*, AMS Chelsea Publishing, Providence, Rhode Island, second edition, 1982.