A Multiresolution Analysis for Likelihoods: Theory and Methods.

Eric D. Kolaczyk
Boston University

Robert D. Nowak
Rice University

November, 2000

ABSTRACT

We describe here a theory for a certain class of multiscale likelihood factorizations wherein, in analogy to a wavelet decomposition of an L^2 function, a given likelihood function has an alternative representation as a product of conditional densities reflecting information in both the data and the parameter vector localized in position and scale. The underlying framework is developed as a set of sufficient conditions for the existence of such factorizations, formulated in analogy to those underlying a standard multiresolution analysis for wavelets. Hence, we establish a theory of multiresolution analysis for likelihoods, and we show how the associated conditions may be used to characterize families of distributions whose members permit a multiscale likelihood factorization. We then consider the use of these alternative representations in a class of problems in nonparametric, complexity penalized likelihood inference, and show that certain fundamental connections between our factorizations, recursive partitioning and a type of generalized Haar basis allow such problems to be solved using analogues of the best-subset and best-basis methods in the literature on wavelets.

AMS (1991) Subject Classification: Primary 62G05; Secondary 60E05.

Key Words and Phrases: Cuts, hereditary constraints, factorization, multiscale, recursive partitioning, reproducibility, unbalanced Haar bases, wavelets.

Acknowledgments: Supported by ARO Grant DAAD19-99-1-0349, NSF Grant MIP-9701692, and ONR Awards N00014-99-1-0219 and N00014-00-1-0390.

1 Introduction.

Much of the recent work in adaptive, nonlinear statistical methods employs a paradigm that begins with the traditional "signal plus noise" model and then seeks an alternative *representation,* often with respect to an orthonormal basis or redundant expansion, that is somehow more appropriate for the assumed underlying structure – Fourier and wavelet-based methods, of course, being canonical examples. However, in many nonparametric statistical problems the "signal plus noise" model can be less applicable or even inappropriate, while a more general likelihood-based model may suggest itself in a natural manner. In such cases, the concept of, say, an orthogonal basis decomposition may be replaced by analogy with a factorization of the data likelihood.

Our focus in this paper is in how the above-mentioned issues of representation pertain to "scale." Often the primary features of scientific interest in a studied phenomena are revealed through data analyses that allow for simultaneous consideration of local structure or behavior at varying levels of what alternately can be termed scale, resolution, or granularity. Commonly a phenomenon of interest is viewed as a function $f(\cdot)$, and observations modeled as samples of f in additive noise i.e., $Y_i = f(i/N) + Z_i$, for $i = 1, \ldots, N$. Wavelets frequently are used then to provide an alternative representation of this model in a manner sensitive to location/scale variations in f. That is, the relation

$$f(t) = \sum_{(j,k)\in\mathbf{Z}^2} \omega_{j,k}\, \psi_{j,k}(t) \tag{1}$$

provides a mechanism through which the details in f gained between certain approximations at scales j and $j+1$, in the vicinity of locations indexed by k, are captured by the coefficients $\omega_{j,k} \equiv \langle f, \psi_{j,k}\rangle$, where for example the $\psi_{j,k}(t) \equiv 2^{j/2}\psi(2^j t - k)$ are orthonormal dilations/translations of a wavelet function $\psi(t)$ satisfying the admissibility condition $\int \psi(t)dt = 0$.

Not surprisingly, nonparametric function estimation is the area of statistics that has been the most profoundly impacted by the introduction of wavelets into the literature (e.g., DeVore and Lucier 1992; Donoho and Johnstone 1994; and a variety of papers since, by numerous other authors). Yet for many statistical problems in which "scale" is relevant it is difficult or even inappropriate to entertain the "signal plus noise" model and conventional wavelet representations. For example, the data may arise with respect to a counting process (i.e., non-additive noise), or even be categorical in form (i.e., non-ordinal). Alternatively, and sometimes in addition, the data may derive from highly irregular sampling patterns, or the data space may be lattice in structure (rather than a continuum), two situations particularly common to many spatial contexts.

In such cases it often can be found that many of the standard statistical methods begin with the specification of a common likelihood function for the data. Hence we take as a premise to our work in this paper that, for a likelihood $p(\boldsymbol{X}|\boldsymbol{\theta})$, a natural analogue to the orthogonal *wavelet decomposition* of a function f in (1) is a *multiscale factorization*

$$p(\boldsymbol{X} \mid \boldsymbol{\theta}) = \prod_{j,k} p(\, D_{j,k}(\boldsymbol{X}) \mid \omega_{j,k}(\boldsymbol{\theta})\,) \quad , \tag{2}$$

where the random variables $D_{j,k}$ contain information in the original data $\boldsymbol{X}$ local to scale j and position k, and the parameters $\omega_{j,k}$ reflect similar information in the original parameter $\boldsymbol{\theta}$. We use the notation $p(\cdot)$ here and throughout the paper to denote either a probability density or a probability mass function, which we shall refer to generically as a density (i.e., p.d.f.) defined with respect to some appropriate dominating measure (which we typically will suppress).

Our interest here is in the existence and derivation of factorizations like that in (2), as well as in their usage for problems of nonparametric inference. In the case of wavelets these issues are connected intimately to the notion of a multiresolution analysis (MRA). Although the first systems of smooth orthonormal wavelets (Stromberg 1982; Meyer 1985), generalizing the construction of Haar (1910), in fact were derived prior to the development of the MRA formalism, arguably it was only with the development of the latter (Mallat 1989; Meyer 1992) that a framework was established for an intuitive understanding of these wavelets and the construction of others. The term "multiresolution," broadly speaking, refers to a method whereby an object of interest is studied at various, nested scales of resolution, and indeed such methods pre-date the appearance of wavelets and the wavelet MRA. For example, multiresolution approaches have been common for some time in the fields of image processing and computer vision, where they arise quite naturally (e.g., Burt and Adelson 1983; Witkin 1983).

Therefore, we describe here a theory of multiresolution analysis for likelihoods. Specifically, we begin in section 2 with the development of a set of sufficient conditions for the existence of a certain class of multiscale factorizations, which are formulated quite naturally in analogy to those underlying a standard MRA for wavelets. In section 3 we then demonstrate how these conditions may be used for characterizing families of distributions whose members permit a multiscale likelihood factorization. Examples are found to include the Gaussian and Poisson families, as well as a certain sub-class of the discrete natural exponential family distributions. Having thus considered issues of existence and characterization, we then study in section 4 the usage of our multiscale factorizations in a general class of nonparametric, complexity penalized likelihood problems. In particular, we show that fundamental connections between our factorizations, the underlying partitions, and a class of generalized Haar bases may be used to develop analogues for best-subset and best-basis methods found in the literature on wavelet and related methods. A brief numerical illustration of these methods is presented in section 5. Finally, some additional discussion closes the paper in section 6.

2 Multiresolution Analysis for Likelihoods.

2.1 The Example of Wavelets.

We begin this section with a brief review of the components underlying a multiresolution analysis for wavelets, based on the material in Daubechies (1992). The idea behind this method is to construct a sequence of subspaces $V_j \subseteq L^2(\mathbb{R})$, across various scales j, whose members serve as successively finer approximations to functions $f \in L^2(\mathbb{R})$. Consider requiring the following four characteristics of these subspaces.

A. Hierarchy of Nested Subspaces. The subspaces V_j satisfy the condition

$$\cdots V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \cdots$$

where $\cap_{j \in \mathbb{Z}} V_j = \{0\}$ and $\overline{\cup_{j \in \mathbb{Z}} V_j} = L^2(\mathbb{R})$.

B. Orthonormal Basis Within V_0. There exists a function ϕ such that the collection $\{\phi(\cdot - k)\}_{k \in \mathbb{Z}}$ forms an orthonormal basis for V_0.

C. Scaling Between Subspaces.

$$g \in V_j \Longleftrightarrow g(2^{-j}\cdot) \in V_0$$

D. Translation Within Subspaces.

$$g \in V_0 \Longrightarrow g(\cdot - k) \in V_0, \quad \forall k \in \mathbb{Z} \ .$$

Our interest in the above characteristics (gathered into these four labeled categories for our own later convenience of exposition) centers primarily on the fact that they form a set of sufficient conditions for the existence of a (wavelet) function $\psi \in L^2(\mathbb{R})$ for which the collection $\{\psi_{j,k}\}$ forms an orthonormal basis of $L^2(\mathbb{R})$, as in equation (1). In other words, these conditions assure a multiscale decomposition or "decoupling" of any given function $f \in L^2(\mathbb{R})$ into components of L^2 "energy" localized to certain combinations of scale j and position k. The fact that this decoupling is with respect to an orthonormal basis implies that knowledge of these components (i.e., the coefficients and their corresponding wavelets) is equivalent to knowledge of the function f itself – only the representation has changed.

As an important aside, we mention here that the conditions in (A) - (D) technically form a multiresolution analysis corresponding to most of what now have come to be called "first generation" wavelets. These include wavelets produced with the addition of more conditions (e.g., vanishing higher order moments, near-symmetry, etc.), as well as the slight loosening of certain of these original conditions (e.g., requiring the weaker condition of a Riesz basis for the collection $\{\phi_{0,k}\}$ in (B)). In contrast, members of the so-called "second generation" of wavelets (Sweldens 1997) in fact give up the basic translation and scaling structure (i.e., induced by conditions (C) and (D)) inherent in the $\psi_{j,k}$ in equation (1). In our development below we focus on crafting a multiresolution analysis for likelihoods more analogous to that associated with the original, "first generation" method for wavelets, while at various points throughout the paper we comment additionally where "second generation" analogues are or might be feasible.

2.2 Likelihood MRAs.

Let $\boldsymbol{X} \equiv (X_1, \ldots, X_N)$ be a vector of random variables, with p.d.f. $p(\cdot|\boldsymbol{\theta})$, for some parameter $\boldsymbol{\theta} \in \Theta^N \subseteq \mathrm{I\!R}^N$. We shall assume throughout that the use of indices $\{1, \ldots, N\}$ reflects an ordering in the underlying data space. (This condition is convenient but not necessary for the results in this section and section 3, but is necessary for the result of Theorem 9 in section 4.2.) Accordingly, this set of indices generally will be referred to as the (discrete) interval $I_0 \equiv (0, N]$, and subsets of indices referred to as subintervals (denoted generically as $I \subseteq I_0$). Note that while such an ordering certainly might correspond to something typical like time or distance (e.g., a financial time series), it may also correspond to a non-temporal index (e.g., spectral energy in astronomy), or even a non-physical index (e.g., opinions in a survey, ordered as "strongly agree," "agree,", etc.). The framework we develop next applies equally in each of these cases, as well as in others.

We begin by formalizing the expression in equation (2) with the following definition.

Definition 1 *By the term multiscale likelihood factorization we shall mean an equality of the form*

$$p\left(\boldsymbol{X} \mid \boldsymbol{\theta}\right) = \prod_{\alpha \in \mathcal{G}} p\left(D_\alpha\left(\boldsymbol{X}\right) \mid \omega_\alpha\left(\boldsymbol{\theta}\right)\right) \tag{3}$$

between the original data likelihood and a factorization with respect to some graph $\mathcal{G}$ reflecting a hierarchy of scales, where

(i) $\boldsymbol{X} \rightarrow \boldsymbol{D}(\boldsymbol{X})$ is a mapping to an N-length representation of statistically independent

components $D_\alpha(\boldsymbol{X})$, each containing information localized to a scale/position combination indexed by nodes $\alpha \in \mathcal{G}$ and distributed according to some member of a common parametric family of distributions, and

(ii) $\boldsymbol{\theta} \to \boldsymbol{\omega}(\boldsymbol{\theta})$ is an accompanying, one-to-one re-parameterization of $\boldsymbol{\theta}$ into multiscale components that are L-independent with respect to the multiscale likelihood function.

In requiring L-independence (Barndorff-Nielsen (1978, pg. 26)), we ask that the domain of variation of $\boldsymbol{\omega}$ be equal to the product of the domains of the components ω_α, and that the role of $\boldsymbol{\omega}$ in the likelihood $p(\boldsymbol{D} \mid \boldsymbol{\omega})$ can be separated into isolated roles as in (3) i.e., one and only one parameter component per probability component. In this sense we achieve a phenomenon like that which occurs in an orthonormal basis decomposition, as in equation (1), wherein a function f is represented as the aggregation of contributions from separate pairs of template functions $\psi_{j,k}$ and parameters $\omega_{j,k}$. In (3), the template functions are replaced by template distributions, as it were, and the common shape of the various wavelets is replaced by a requirement that the distributions fall in the same parametric family.

A simple example of (3) occurs in the case where the data are described by the model $X_i = \theta_i + \sigma Z_i$, where the Z_i are i.i.d. Gaussian with zero mean and unit variance, and a discrete orthogonal wavelet transform is applied. The result, as is well known, is the transformed model $D_{j,k} = \omega_{j,k} + \sigma Z_{j,k}$, where $D_{j,k}$ are the empirical wavelet coefficients of the data, the $\omega_{j,k}$ are the wavelet coefficients of the vector $\boldsymbol{\theta}$, and the $Z_{j,k}$ are similarly i.i.d. Gaussian with zero mean and unit variance. The graph $\mathcal{G}$ in this case is a simple binary tree, induced by the underlying dyadic structure of the discrete wavelet transform, and the multiscale likelihood factorization is in fact like that in equation (2), where the components on the right-hand side are univariate Gaussian density functions with means $\omega_{j,k}$ and common variance σ^2. These results may be contrasted with those that follow in the case where, for example, the X_i instead are modeled as independent Poisson, with means θ_i, and a discrete wavelet transform similarly is applied. A factorization in terms of individual wavelet coefficients (2) is not possible in the Poisson case due to the statistical dependency among the $D_{j,k}$ (Kolaczyk 1999a).

In analogy to the four characteristics (A) – (D) outlined in section 2.1, we will now describe four characteristics sufficient to insure a certain multiscale factorization with structure like that in (3). The first three characteristics will be seen to play roles similar to those of (A), (B), and (C), while the last would seem to pertain uniquely to the context of likelihoods. To begin with, we suppose that the notion of multiple resolutions arises (or can be viewed as having arisen) through recursive partitioning (or, conversely, hierarchical aggregation) of the dataspace i.e.,

A*. Hierarchy of Recursive Partitions. A collection $\{\mathcal{P}_\ell\}_{\ell=1}^N$ of recursively defined partitions of the discrete interval $(0, N]$, resulting in the nesting

$$\mathcal{P}_1 \subset \cdots \subset \mathcal{P}_N \ ,$$

such that $\mathcal{P}_1 = (0, N]$ and $\mathcal{P}_N = \cup_{i=1}^N \{i\}$.

We assume that, following the coarsest partition $\mathcal{P}_1$, each intermediate partition in the hierarchy is obtained by splitting one and only one interval of its immediate predecessor into two subintervals (hence resulting in N such $\mathcal{P}_\ell$). In general, we will write $\mathcal{P}^* \equiv \{\mathcal{P}_\ell\}_{\ell=1}^N$, and refer to $\mathcal{P}^*$ as a *complete recursive partition* (C-RP). For example, if $N = 4$, two possible C-RP's of $(0, 4]$ are

$$\{\,\{1,2,3,4\}\,,\ \{\{1,2\},\{3,4\}\}\,,\ \{\{1\},\{2\},\{3,4\}\}\,,\ \{\{1\},\{2\},\{3\},\{4\}\}\,\}$$

and

$$\{\,\{1,2,3,4\}\,,\ \{\{1\},\{2,3,4\}\}\,,\ \{\{1\},\{2,3\},\{4\}\}\,,\ \{\{1\},\{2\},\{3\},\{4\}\}\,\}\ .$$

The former is a complete, recursive *dyadic* partition (C-RDP), while the latter is non-dyadic. Note that to any C-RP there corresponds a labeled, binary tree, which we denote $\mathcal{T}^* \equiv \mathcal{T}(\mathcal{P}^*)$, where the label α at each vertex of $\mathcal{T}^*$ refers to some corresponding subinterval I_α of I_0 with cardinality, say, N_{I_α}. If $\mathcal{P}^*$ is the C-RDP, for example, then we may adopt the standard (j, k) indexing of dyadic wavelet bases for our labels. Without loss of generality, we will continue to emphasize through our notation the role of intervals and partitions over that of the corresponding trees, except in cases where explicit use of the latter is particularly convenient and/or useful in making a given argument.

Hierarchies of partitions occur quite naturally in many contexts. For instance, with respect to our earlier examples, they might arise through the aggregation of time series or spectral data over successively wider bins of time or energy, or the collapsing of response categories in an opinion survey. To condition (A*) we next add

B*. Independence Within $\mathcal{P}_N$. The components of $\boldsymbol{X}$ are statistically independent, the components of $\boldsymbol{\theta}$ are L-independent with respect to the likelihood of $\boldsymbol{X}$ i.e.,

$$p\left(\boldsymbol{X} \,|\, \boldsymbol{\theta}\right) = \prod_{i=1}^{N} p\left(X_i \,|\, \theta_i\right) \quad ,$$

and the p.d.f for each X_i is a member of some common parametric family $\mathcal{F} \equiv \{p(\cdot|\theta) : \theta \in \Theta \subseteq \mathrm{I\!R}\}$.

This condition can be seen to be analogous to condition (B), in that the assumption of a likelihood factorization in the original data space, with respect to the index set $\{1, \ldots, N\}$, and in components identical up to the parameters θ_i, mirrors the spirit and function of the orthonormal basis $\{\phi(\cdot - k)\}$ in V_0. However, we should point out that, just as the condition of orthonormality can be weakened in the case of wavelets (as mentioned in section 2.1), and therefore is not a necessary condition, so too we shall see in Theorem 5 (Section 3) that condition (B*), while sufficient, is not a necessary condition either.

The net result of conditions (A*) and (B*) is a hierarchically defined scheme for grouping subsets of a vector of independent random variables, whose distributions are identical up to parameterization within some family $\mathcal{F}$. The manner of grouping derives from the C-RP $\mathcal{P}^*$, which describes how any given "parent" interval $I \equiv (i_1 \ldots i_{N_I}]$ is split into two "children" subintervals $I_{ch(I),l}$ and $I_{ch(I),r}$, denoted $l \equiv$"left" and $r \equiv$"right" with respect to the ordering underlying the original index set $\{1, \ldots, N\}$. For each interval I, it is desirable to combine the information in $\{X_{i_1}, \ldots, X_{i_{N_I}}\}$ into a single summary statistic, X_I, with the simplest approach being through direct summation i.e.,

$$X_I \equiv X_{i_1} + \cdots + X_{i_{N_I}}, \tag{4}$$

which will be focused on exclusively in the sequel. Our third condition is then

C*. REPRODUCIBILITY BETWEEN PARTITIONS. The family $\mathcal{F}$ is reproducible in θ, in the sense that $\forall N_I \geq 1$, and $\forall \boldsymbol{\theta} \in \Theta^{N_I}$, the p.d.f. of $\sum_{i \in I} X_i$ is $p(\cdot|\theta_I) \in \mathcal{F}$, where $\theta_I \equiv \sum_{i \in I} \theta_i$.

Conceptually, reproducibility simply dictates that the distributional family is in some sense "invariant" under summation (a scale-invariance, if you will). Practically speaking, however, there are in fact a number of similar definitions in use (see section 3). We use the very simplest definition here, found for example in Wilks (1962), which describes the well-known behavior of such distributions as the Gaussian, Poisson, Cauchy, and others.

We now have the following

Theorem 1 *Given conditions (A*), (B*), and (C*), there holds a factorization of the form*

$$p(\boldsymbol{X} \,|\, \boldsymbol{\theta}) = p(X_{I_0} \,|\, \boldsymbol{\theta}) \prod_{I \in NT(\mathcal{P}^*)} p(X_{ch(I),l} | X_I, \boldsymbol{\theta}) \ \ , \tag{5}$$

where $NT(\mathcal{P}^*)$ *is the set of all non-terminal intervals in* $\mathcal{P}^*$ *(i.e., non-singletons), and the probability components* $p(X_{ch(I),l}|X_I, \boldsymbol{\theta})$ *all belong to the same parametric family of distributions.*

Proof of Theorem 1: For notational simplicity, we suppress the role of $\boldsymbol{\theta}$. Given the independence of the X_i and the binary nature of the tree $\mathcal{T}^*$ underlying $\mathcal{P}^*$, for equation (5) it is sufficient to note the relation

$$p(X_{ch(I),l}, X_{ch(I),r}) = p(X_{ch(I),l}|X_I)\, p(X_I) \ \ ,$$

and apply it recursively on $NT(\mathcal{P}^*)$ moving from fine scales to coarse (i.e., from the leaves of $\mathcal{T}^*$

to the root). Alternatively, this result may be viewed as a consequence of the fact that conditions (A*) and (B*) imply a so-called directed, local Markov property for the graphical model given by $\{X_{I_\alpha}\}_{\alpha\in\mathcal{T}^*}$, when $\mathcal{T}^*$ is equipped with arrows denoting parent/child relationships. Equation (5) then follows as an example of a recursive factorization (e.g., Lauritzen 1997, theorem 3.27). The fact that the conditional distributions in the factorization are of the same family follows from condition (C*). $\square$

Note that equation (5) essentially achieves the form that we seek in equation (3) of Definition 1, with $D_{I_\alpha}(\boldsymbol{X}) =^d X_{ch(I_\alpha),l}\,|\,X_{I_\alpha}$, except that the components of $\boldsymbol{\theta}$ cannot necessarily be expected to be L-independent with respect to this factorization as well. Put another way, we have a factorization of $p(\boldsymbol{X}|\boldsymbol{\theta})$ into components that are functions of only *local* information in $\boldsymbol{X}$, but *global* information in $\boldsymbol{\theta}$. However, one may hope that there exists some corresponding reparameterization of $\boldsymbol{\theta}$ that achieves a similar localization. Such behavior may be assured through the following condition.

D*. "Decoupling" of Parameters with Partitions (i.e. Cuts).

For any $X_i \sim p(\cdot|\theta_i) \in \mathcal{F}$, $i = i_1, i_2$, there exists some re-parameterization $(\theta_{i_1}, \theta_{i_2}) \rightarrow (\theta_{I_{12}}, \omega_{I_{12}})$ such that

$$p(\,X_{i_1}, X_{i_2}\,|\,\theta_{i_1}, \theta_{i_2}\,) = p(X_{I_{12}}\,|\,\theta_{I_{12}})\,p(X_{i_1}\,|\,X_{I_{12}}, \omega_{I_{12}})\;\;,$$

where $X_{I_{12}} \equiv X_{i_1} + X_{i_2}$ and $\theta_{I_{12}} \equiv \theta_{i_1} + \theta_{i_2}$. That is, $X_{I_{12}}$ is a cut for (X_{i_1}, X_{i_2}).

Generally speaking, given a statistic $S \equiv S(\mathbf{Y})$ defined with respect to a random variable $\mathbf{Y}$ with distribution function F, S is said to be a *cut* if the mapping $F \rightarrow (F_S, F^S)$ is surjective, where F_S and F^S are the marginal distribution of S and the conditional distribution given S, respectively. In the event that F is parameterized, a cut refers to a separation or "decoupling" of the parameters (or some re-parameterization thereof) with the marginal and conditional distributions. Conceptually, a cut is similar to the idea of sufficiency; in fact, it is equivalent to the principle of so-called S-sufficiency. The reader is referred to Barndorff-Nielsen (1978) for a detailed exposition these ideas, their definitions, and their usage.

We are now able to state the main result of this section.

Theorem 2 *Assume that the conditions (A^*), (B^*), (C^*), and (D^*) hold. Then there exists a factorization of the form*

$$p(\boldsymbol{X} \,|\, \boldsymbol{\theta}) = p(X_{I_0} \,|\, \theta_{I_0}) \prod_{I \in NT(\mathcal{P}^*)} p(X_{ch(I),l} | X_I, \omega_I) \ , \tag{6}$$

with respect to some re-parameterization $\{\theta_{I_0}, \boldsymbol{\omega}\}$ *of* $\boldsymbol{\theta}$*, where* $\theta_{I_0} \equiv \sum_{i=1}^N \theta_i$.

Recalling the recursive, pairwise treatment of the random variables in the proof of Theorem 1, and noting the pairwise nature of condition (D^*), the proof of Theorem 2 is immediate.

We see that (6) is an example of (3). For a random variable $\boldsymbol{X}$ associated with family $\mathcal{F}$ for whom (A^*) – (D^*) are satisfied, we will say that $\mathcal{F}$ *allows a multiresolution analysis with respect to* $\boldsymbol{\theta}$. A side-by-side comparison of the conditions underlying a wavelet MRA and a likelihood MRA is given in Table 1. As remarked above, the first three conditions of each are quite analogous in form and function. In fact, since condition (C^*) arguably shares the spirit of not only condition (C) but (D) as well in a wavelet MRA, it would appear that condition (D^*) alone stands out as being particularly unique to the likelihood-based context. In considering the factorization in (6), note that the role of a wavelet coefficient/function pair $(\omega_{j,k}, \psi_{j,k})$, in capturing detail lost between scales $j+1$ and j in approximating $f \in L^2(\mathbb{R})$, is played here by the conditional density $p(X_{ch(I),l}|X_I, \omega_I)$, a natural form of expressing the information lost between the aggregations dictated by a partition $\mathcal{P} \subseteq \mathcal{P}^*$ and its immediate predecessor. In the next section we present results characterizing those families $\mathcal{F}$, within certain broad classes, that allow a multiresolution analysis and provide in the process examples of explicit forms for these conditional densities.

3 Characterization.

Consider again the conditions underlying our likelihood MRA. While (A^*) establishes a fixed C-RP $\mathcal{P}^*$ and (B^*) imposes fairly standard distributional assumptions on the X_i, conditions

(C*) and (D*) arguably are the determining factors as to whether or not a family $\mathcal{F}$ allows an MRA. There are relatively small but detailed literatures on both reproducibility and cuts, mainly pertaining to their roles in exponential family distributions. In this section we illustrate, through the proofs of Theorems 3 and 4, how results from this literature may be used for the purpose of characterization. By a one-parameter natural exponential family (NEF), we shall mean a family $\mathcal{F}$ with densities of the form

$$p(X_i|\eta_i) = a(\eta_i)b(X_i)\exp\{\eta_i X_i\}$$

with respect to some sigma-finite measure $\nu(\cdot)$, for $\eta \in \mathrm{E} \subset \mathbb{R}$. Such families will be assumed to be minimal throughout, and either steep or regular (as dictated by our use of existing results).

We begin by illustrating the role that may be played by the condition of reproducibility and, for reasons that will become clear in the proof that follows, we adopt a slightly more general version of that given in (C*).

(C**) REPRODUCIBILITY-N^β: The family $\mathcal{F}$ is reproducible-N^β in θ, in the sense that $\forall N_I \geq 1$, there exists a constant $\beta \in \mathbb{R}$ (independent of N_I) and a surjective mapping $g_{N_I} : \Theta \rightarrow \Theta$ such that the p.d.f. of $(N_I)^\beta \sum_{i \in I} X_i$ is given by $p(\,\cdot \mid g_{N_I}(\theta_I)\,) \in \mathcal{F}$.

This particular definition is analogous to that of Bar-Lev and Enis (1986), whose definition was for random variables independent and identically distributed. When (C**) is combined with conditions (A*), (B*), and (D*), the factorization of equation (6) again follows, but in the scaled random variables $Y_I \equiv N_I^\beta X_I$ (where the notation in (D*) is to be modified in the obvious fashion). We then have the following characterization result.

Theorem 3 *Suppose that $\mathcal{F}$ is a (minimal and steep) one-parameter NEF. Then it follows that*

(i) $\mathcal{F}$ allows a likelihood MRA with respect to the natural parameterization $\boldsymbol{\theta} \equiv \boldsymbol{\eta}$ if and only if $\mathcal{F}$ is the family of Gaussian distributions;

(ii) $\mathcal{F}$ allows a likelihood MRA with respect to the mean parameterization $\boldsymbol{\theta} \equiv \boldsymbol{\mu}(\boldsymbol{\eta})$ if and only if $\mathcal{F}$ is either the family of Gaussian distributions or the family of Poisson distributions.

Proof of Theorem 3: Assume a C-RP like that in (A*) and the independence structure of (B*). Next, note that the collection of NEF's $\mathcal{F}$ satisfying (C**) must be contained within the collection of such $\mathcal{F}$ which do so in the case of i.i.d. random variables i.e., where $\eta_1 = \ldots = \eta_N \equiv \eta$, for some $\eta \in \mathrm{E}$. A characterization of this latter case is provided in Bar-Lev and Enis (1986). Specifically, among various other results, these authors show that reproducibility implies that $\mathcal{F}$ must have a power variance function (PVF) i.e., of the form $\mathrm{var}(X) \equiv V(\mu(\eta)) = \sigma \times (\mu)^\gamma$, for some $\gamma \in \mathbb{R} - \{2\}$, in which case the parameter β in (C**) is given by $\beta = (1-\gamma)/(\gamma-2)$. Furthermore, they present a characterization of all such NEF-PVF families, listing them as Gaussian ($\gamma = 0$, $\beta = -1/2$), Poisson ($\gamma = 1$, $\beta = 0$), compound Poisson ($\gamma \in (1,2)$), and a class generated by certain stable distributions ($\gamma > 2$).

Now for (i) we can consult the cumulant generating function (cgf) $\Psi(t;\eta) \equiv \log E[e^{tX}]$, for which condition (C**) implies that we must have $\Psi(t, g_{N_I}(\eta_I)) = \sum_{i=i_1}^{i_{N_I}} \Psi(tN_I^\beta; \eta_i)$, where $\eta_I \equiv \sum_{i \in I} \eta_i$. Examination of the four such possible $\Psi(t,\eta)$, provided in the i.i.d. case by Bar-Lev and Enis (1986), quickly yields result (i), due to the fact that for all but the Gaussian distribution the cgf's are prohibitively nonlinear in η. For (ii), it is more convenient to work with the variance function $V(\mu)$, which (along with $\mu(\mathrm{E})$) uniquely characterizes its NEF within the class of all NEFs (Morris 1982). Since $E[N_I^\beta \sum X_i] = N_I^\beta \sum \mu_i$ and $\mathcal{F}$ must have PVF, we have the condition $\sigma(N_I^\beta \sum \mu_i)^\gamma = N_I^{2\beta} \sum \sigma \mu_i^\gamma$, which is satisfied only in the case of $\gamma = 0$ or 1.

Finally, it remains to show that the sum of independent random variables forms a cut for the joint distribution in the case of the Gaussian and Poisson families, hence satisfying condition (D*), which is straightforward (e.g., Barndorff-Nielsen 1978). □

We will find it useful in sections 4 and 5 to have recorded the specific forms of the multiscale factorizations for Gaussian and Poisson data in the following corollary.

Corollary 1 *Let* $\mu_I = \mu_{ch(I),l} + \mu_{ch(I),r}$ *denote the mean of* $X_I, \forall I \in \mathcal{P}^*$.

(i) Suppose that the original observations $X_i \sim Gaussian(\mu_i, \sigma^2)$. *Then*

$$\prod_{i=1}^{N} p(X_i|\mu_i,\sigma^2)dX_i = p(X_{I_0}|\mu_{I_0}, N_{I_0}\sigma^2)dX_{I_0} \prod_{I\in NT(\mathcal{P}^*)} p(X_{ch(I),l}|X_I, \omega_I, c_I\sigma^2)dX_{ch(I),l} \tag{7}$$

where

$$X_{I_0}|\mu_{I_0}, N_{I_0}\sigma^2 \sim Gaussian\left(\mu_{I_0}, N_{I_0}\sigma^2\right) \ \ , and$$

$$X_{ch(I),l}|X_I, \omega_I, c_I\sigma^2 \sim Gaussian\left(\frac{N_{ch(I),l}}{N_I}x_I - \omega_I \, , \, c_I\sigma^2\right) \ \ ,$$

with

$$\omega_I = c_I\left(\frac{\mu_{ch(I),r}}{N_{ch(I),r}} - \frac{\mu_{ch(I),l}}{N_{ch(I),l}}\right) \ \ ,$$

for $c_I = N_{ch(I),l}\, N_{ch(I),r}/N_I$.

(ii) Suppose that the original observations $X_i \sim Poisson(\mu_i)$. *Then*

$$\prod_{i=1}^{N} p(X_i|\mu_i) = p(X_{I_0}|\mu_{I_0}) \prod_{I\in NT(\mathcal{P}^*)} p(X_{ch(I),l}|X_I, \omega_I) \ \ , \tag{8}$$

where $X_{I_0}|\mu_{I_0} \sim Poisson(\mu_{I_0})$ *and* $X_{ch(I),l}|X_I, \omega_I \sim Binomial(X_I; \omega_I)$,
with $\omega_I = \mu_{ch(I),l}/\mu_I$.

We postpone more detailed discussion of the above expressions until section 4. Before moving onward with our characterization results, however, we make the following observations.

Remark 1: Parameterization matters – hence our care in writing that a given family $\mathcal{F}$ allows an MRA *with respect to* a given parameterization. For the Gaussian distribution, the natural and mean parameterizations are equal up to a constant scaling term, and so the two cases are essentially the same. For the Poisson, however, $\eta_i \equiv \log(\mu_i)$.

Remark 2: Choice of a *class* of families $\{\mathcal{F}\}$ matters. For example, the family of gamma distributions is not reproducible as a one-parameter NEF (where σ is fixed), but is so when viewed as a two-parameter family (Bar-Lev and Enis 1986). This latter perspective necessitates the broadening of the concept of "family" to include NEF-PVF's that have the same values of γ and $\mu(\mathrm{E})$, but potentially different support.

Remark 3: Similarly, from the perspective of reproducibility, there are families outside of NEF's that suggest themselves. A particularly attractive candidate is the Cauchy distribution, a member of the larger of class of stable distributions for which a generalization of our notion of reproducibility could be introduced. It is not difficult to show, however, that the X_I do not qualify as cuts under the Cauchy distribution (Kolaczyk 1999a), thus violating condition (D^*).

In the above approach we used reproducibility to reduce our class of candidate families $\mathcal{F}$ to just a handful, after which we applied the concept of cuts case-by-case. Now consider the converse, wherein the condition of cuts is applied first. By far the most explicit general statements regarding cuts can be made regarding discrete NEF's, and so we have the following alternative characterization of the Poisson distribution among families allowing a likelihood MRA.

Theorem 4 *Suppose that $\mathcal{F}$ is a (minimal and regular) discrete, one-parameter NEF, for which the sample space $\mathcal{X}$ is a subset of $\{0,1,2,\ldots\}$ containing the points 0 and 1. Then the family $\mathcal{F}$ allows a likelihood MRA if and only if $\mathcal{F}$ is Poisson.*

Proof of Theorem 4: As in the proof of Theorem 3, assume a C-RP like that in (A^*) and the independence structure of (B^*). Next, without loss of generality, consider any pair of random variables X_{i_1} and X_{i_2} with distributions in $\mathcal{F}$. The joint distribution will be a bivariate NEF (minimal and regular) with sample space containing the points $\{0,0\}$, $\{0,1\}$, and $\{1,1\}$. By theorem 4 of Barndorff-Nielsen (1976), which pertains to a multivariate version of our bivariate case, $X_{i_1} + X_{i_2}$ is a cut (and hence satisfies condition (D*)) if and only if the joint distribution of (X_{i_1}, X_{i_2}) is that of a sum-symmetric power series distribution (SSPSD). However, the independence of X_{i_1} and X_{i_2} dictated by condition (B*) is satisfied, among all SSPSD's, only by the multivariate Poisson distribution (Joshi and Patil 1968, property 6.3). Reproducibility of the Poisson distribution in the mean parameterization is well-known, and hence condition (C*) is satisfied. The conclusion of the theorem then follows. $\square$

On a final note, we see from the above proof that the family of SSPSD's plays a key role in the theory of cuts for discrete exponential families, but that the independence assumption of condition (B*) excludes all but one member from allowing a likelihood MRA. Interestingly enough, despite the general dependency possessed by its members, the entire class of SSPSD families can be shown to permit multiscale likelihood factorizations like that in (6).

Theorem 5 *Suppose that the vector $(X_1, X_2, \ldots, X_N)$ follows a sum-symmetric power series distribution i.e., with probability mass function of the form*

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = b(\boldsymbol{X})\, \frac{\theta_1^{X_1} \cdots \theta_N^{X_N}}{g(\boldsymbol{\theta})} \quad ,$$

where the generating function $g(\cdot)$ depends on $\boldsymbol{\theta}$ only through $\theta_1 + \cdots + \theta_N$. Then $p(\boldsymbol{X}|\boldsymbol{\theta})$ allows a multiscale factorization, as in equation (6), where $X_{I_0}|\theta_{I_0}$ follows the univariate power series distribution with parameter θ_{I_0} and generating function $g(\theta_{I_0})$, and $X_{ch(I),l}|X_I, \omega_I$ is distributed as Binomial$(X_I; \omega_I)$, with $\omega_I = \theta_{ch(I),l}/\theta_I$, where $\theta_I \equiv \sum_{i \in I} \theta_i$.

Proof of Theorem 5: Assume a C-RP like that in (A*). The proof follows by exploiting properties 4.5, 4.6, and 4.3 of Joshi and Patil (1968), while constructing the relevant conditional probabilities recursively in $I \in NT(\mathcal{P}^*)$, as in the proof of equation (6). Specifically, suppose that the underlying C-RP $\mathcal{P}^*$ dictates that the first grouping (or, conversely, the final partitioning) of indices occur between $\{\{i\}, \{i+1\}\}$ and $\{i, i+1\}$ (which, in our interval notation correspond to $\{(i-1, i], (i, i+1]\}$ and $(i-1, i+1]$, respectively). Then by property 4.5 of Joshi and Patil, it follows that the $N-1$ length vector $(X_1, \ldots, X_{i-1}, X_{i,i+1}, X_{i+2}, \ldots, X_N)$ obeys an SSPSD, with parameter $(\theta_1, \ldots, \theta_{i-1}, \theta_{i,i+1}, \theta_{i+2}, \ldots, \theta_N)$ and generating function

$g(\theta_1 + \ldots + \theta_{i-1} + \theta_{i,i+1} + \theta_{i+2} + \ldots + \theta_N)$. Additionally, by property 4.6 of Joshi and Patil, we have that, given $(X_1, \ldots, X_{i-1}, X_{i,i+1}, X_{i+2}, \ldots, X_N)$, the conditional distribution of $\boldsymbol{X}$ is the product of $N-1$ independent (singular) multinomial distributions. One of these multinomials corresponds to the distribution of $(X_i, X_{i+1})|X_i + X_{i+1}$, which is just binomial, while the rest correspond to the trivial distributions of $X_\ell|X_\ell$, for each $\ell \neq i, i+1$. Hence the result of this step is to factor the original data likelihood into the product of a single binomial and an SSPSD component of dimension $N-1$. This sequence of arguments may be continued recursively and, at the end, property 4.3 of Joshi and Patil can be used to conclude that the distribution of X_{I_0} is a PSD with the stated parameterization. $\square$

Comparing Theorem 5 and Corollary 1, we see that the factorization for the Poisson distribution holds generally for all SSPSD's, up to the term involving the marginal distribution of X_{I_0}. Also, in Theorem 5 we see indications of how the MRA defined in section 2 may generalize. For example, the independence assumption underlying condition (B*) is sufficient, but not necessary, as the components of $\boldsymbol{X}$ are in general dependent for members of the SSPSD family other than Poisson (e.g., multinomial, negative multinomial, multivariate logarithmic, etc.). On the other hand, as the concept of cuts and SSPSD's are closely intertwined, this further suggests the relevance of condition (D*). Additionally, property 4.5 of Joshi and Patil (1968), in its more general form than was used here, is essentially a property of reproducibility. That is, if $\mathcal{P} \equiv \{I_1, \ldots, I_m\}$ is a partition of $(0, N]$, for $m < N$, then the distribution of $(X_{I_1}, \ldots, X_{I_m})$ is again a member of the family of SSPSD's. Here reproducibility is to be interpreted in a broader sense, as the dimension of the relevant SSPSD changes with each recursion. The reader is referred to section 6 for additional discussion on potential extensions to the basic likelihood MRA of section 2.

4 Adaptive Methods of Inference.

So far in this paper our attention has been on the development of conditions under which multiscale likelihood factorizations of a certain type exist, and the characterization of examples within certain classes of distributional families. We now turn our focus to the role that may be played by such factorizations in problems of adaptive nonparametric inference. In particular, we show how fundamental connections between our multiscale factorizations, their underlying partitions, and a specific class of orthonormal wavelet bases may be used to develop analogues of the best-subset and best-basis methods popular in the literature on wavelet-based nonparametric function estimation.

4.1 Partitions, Bases, and Factorizations – Connections.

Motivating our work on factorizations was the goal of finding alternative, multiscale representations for the data likelihood, in analogy to those provided for functions f by orthonormal bases of wavelets. Key to our approach is the assumption of a hierarchical collection of recursively defined partitions i.e., $\mathcal{P}^*$, of the corresponding data space. Hence our particular multiscale factorizations possess an intimate connection with recursive partitions, which in turn are fundamental to a variety of adaptive statistical methods (e.g., CART).

At the same time, a C-RP $\mathcal{P}^*$ can be put in one-to-one correspondence with a certain type of orthonormal wavelet basis. Let the inner product of functions f and g defined on $(0, N]$ be given by $\langle f, g\rangle \equiv \sum_i f(i)g(i)$. We then have the following

Theorem 6 *To the specification of any C-RP $\mathcal{P}^*$ of $(0, N]$ there corresponds the construction of an orthonormal basis of $I\!R^N$, $\mathcal{B}(\mathcal{P}^*) \equiv \{\phi_{I_0}\} \cup \{\psi_I\}_{I\in NT(\mathcal{P}^*)}$, where $\phi_I(i) = \chi_I(i)/N_I^{1/2}$ and*

$$\psi_I(i) = c'_I \left[\frac{\chi_{ch(I),r}(i)}{N_{ch(I),r}} - \frac{\chi_{ch(I),l}(i)}{N_{ch(I),l}} \right] \quad . \tag{9}$$

Here χ_I is the characteristic function on I, and

$$c'_I = \left(N^{-1}_{ch(I),r} + N^{-1}_{ch(I),l} \right)^{-1/2} \tag{10}$$

is a normalizing constant.

Proof of Theorem 6: As each function ψ_I corresponds to the partitioning of one interval in a given partition $\mathcal{P} \subset \mathcal{P}^*$ into two components, and this operation is performed $N-1$ times in constructing a C-RP, it follows that the cardinality of $\mathcal{B}(\mathcal{P}^*)$ is N. Additionally, the elements of $\mathcal{B}(\mathcal{P}^*)$ are normalized and pairwise orthogonal by construction. The latter property follows from observing that between any pair of elements, one will be fully supported on a subinterval in which the other is constant, and all elements but ϕ_{I_0} have zero mean. $\square$

The system $\mathcal{B}(\mathcal{P}^*)$ is a simple example of an *unbalanced Haar basis* (UHB). Girardi and Sweldens (1997) derive UHBs in some generality for nested partitions of arbitrary measure spaces. Note that $\mathcal{B}(\mathcal{P}^*)$ is in fact a generalization of the standard *dyadic* Haar basis (Haar 1910). The latter corresponds to the special case of a complete recursive *dyadic* partition, say $\mathcal{P}^*_{Dy}$, in which each subpartition $\mathcal{P}$ is formed from its predecessor by splitting one of the candidate subintervals precisely in half. Engel (1994) exploits this connection between $\mathcal{P}^*_{Dy}$ and the dyadic Haar basis in one dimension, while Donoho (1997) makes use of an analogous result for anisotropic, dyadic Haar bases in two dimensions. The UHB's defined in Theorem 6 constitute an example of so-called "second generation" wavelet bases, in that the functions ψ_I cannot be expressed as the dilations/translations of a single "mother" wavelet function ψ. Nevertheless, they are accompanied by such standard properties as a multiresolution analysis and fast transform algorithms.

Now implicit in the relationship between $\mathcal{P}^*$ and $\mathcal{B}(\mathcal{P}^*)$ is a similar relationship between subpartitions $\mathcal{P} \subset \mathcal{P}^*$ and certain subsets of $\mathcal{B}(\mathcal{P}^*)$. To make this statement more concrete, consider the problem of representing the mean $\boldsymbol{\mu}$ of data $\boldsymbol{X}$ with respect to either a recursive partition $\mathcal{P} \subset \mathcal{P}^*$ or the UHB $\mathcal{B}(\mathcal{P}^*)$. Suppose that $\boldsymbol{\mu}$ is in fact piecewise constant on $\mathcal{P}$. Then we may write $\mu_i = \mu_I / N_I$, when $i \in I$, on the terminal intervals I of $\mathcal{P}$ i.e., for $I \in T(\mathcal{P})$. Noting that the binary tree $\mathcal{T} \equiv \mathcal{T}(\mathcal{P})$ corresponding to $\mathcal{P}$ is in fact a pruned version of $\mathcal{T}^*$ suggests that $\boldsymbol{\mu}$ can be represented in terms of the corresponding UHB by retaining only some structured subset of UHB coefficients. The following theorem makes this notion precise.

Theorem 7 *Let $\mathcal{P}^*$ be a C-RP of $(0, N]$, and suppose that $\boldsymbol{\mu}$ is piecewise constant with respect to some partition $\mathcal{P} \subseteq \mathcal{P}^*$ i.e.,*

$$\mu_i = \sum_{I \in T(\mathcal{P})} \langle \boldsymbol{\mu}, \phi_I \rangle \, \phi_I(i) \quad . \tag{11}$$

Then $\boldsymbol{\mu}$ has an equivalent representation in terms of the UHB $\mathcal{B}(\mathcal{P}^*)$ in the form

$$\mu_i = \bar{\mu} \;\; + \sum_{I \in NT(\mathcal{P})} \langle \boldsymbol{\mu}, \psi_I \rangle \, \psi_I(i) \quad . \tag{12}$$

That is, the presence or absence of UHB components $\langle \boldsymbol{\mu}, \psi_I \rangle \, \psi_I$ in the expansion of $\boldsymbol{\mu}$ can be expressed through a collection of indicators $\mathcal{M}(\mathcal{P}) \equiv \{\mathcal{M}_I\}$, where the $\mathcal{M}_I$ satisfy the following two conditions

(H1) Each $\mathcal{M}_I$ is either 0 or 1, $\forall I \in NT(\mathcal{P}^*)$.

(H2) $\mathcal{M}_I = 1 \Rightarrow \mathcal{M}_{I'} = 1$, $\forall I' \supset I$, while $\mathcal{M}_I = 0 \Rightarrow \mathcal{M}_{I''} = 0$, $\forall I'' \subset I$.

Proof of Theorem 7 is straightforward, following an argument similar to, for example, an analogous result of Engel (1994) in the context of certain nonparametric regression estimators on recursive dyadic partitions. Donoho (1997) provides a result identical to that of Engel, in two dimensions, and labels the analogues of conditions (H1) and (H2) as *hereditary constraints.* These constraints are a critical component in Donoho's demonstration of an equivalence between a certain restricted form of least-squares CART estimation (Breiman, Friedman, Olshen, and Stone 1983) and complexity-penalized least-squares best-basis estimators over a certain library of basis functions. In particular, the best-basis estimators are seen to result from application of the same "bottom-up" optimal tree-pruning algorithm as used in CART.

The results that we present next, in section 4.2, are of a similar spirit for complexity-penalized likelihood estimation. Before proceeding, however, it is important to consider the relationship implied by the above results between our factorizations and UHB's. Of most immediate relevance is the fact that for each of the three families permitting multiscale likelihood factorizations in section 3 – Gaussian, Poisson, and SSPSD (i.e., treating Poisson as separate from other SSPSD's) – the multiscale coefficients ω_I obey a set of hereditary constraints in one-to-one correspondence with those of the UHB coefficients. For example, in the Gaussian case $\omega_I = (c_I/c'_I)\langle \boldsymbol{\mu}, \psi_I\rangle$, and hence the $\mathcal{M}_I$ indicate whether $\omega_I = 0$ or not. Similarly, in the Poisson case we have that

$$\omega_I = \frac{\mu_{ch(I),l}}{\mu_I} = c'_I \left(\frac{c'_I}{N_{ch(I),r}} - \frac{\langle \boldsymbol{\mu}, \psi_I\rangle}{\mu_I} \right) \quad , \tag{13}$$

and therefore when $\mathcal{M}_I = 0$, indicating that $\langle \boldsymbol{\mu}, \psi_I\rangle = 0$, it indicates as well that $\omega_I = \rho_I \equiv N_{ch(I),l}/N_I$. Finally, for SSPSD's in general, since $\mu_i = (\mu_{I_0}/\theta_{I_0})\,\theta_i$, we have that $\langle \boldsymbol{\mu}, \psi_I\rangle = 0$ if and only if $\langle \boldsymbol{\theta}, \psi_I\rangle = 0$, and so $\mathcal{M}_I = 0$ indicates that $\omega_I = \theta_{ch(I),l}/\theta_I = \rho_I$ by a similar argument.

In other words, the re-parameterizations of $\boldsymbol{\mu}$ induced by the likelihood factorizations in section 3 behave analogously to the coefficients corresponding to the UHB $\mathcal{B}(\mathcal{P}^*)$. However, the fact that these re-parameterizations are not identical to the UHB coefficients is important. That is, if one instead begins with an analysis directly of the data $\boldsymbol{X}$ with respect to a UHB, a likelihood factorization like that in (6) generally does not follow. The optimization algorithms we describe next for complexity penalized inference, although analogous to best-subset and best-basis algorithms in the wavelet literature, are available in part only because such factorizations exist.

4.2 Complexity Penalized Likelihood Methods.

4.2.1 Fixed C-RP $\mathcal{P}^*$.

Consider now a class of complexity-penalized likelihood estimators that are of the general form

$$\hat{\boldsymbol{\theta}}(\boldsymbol{X}) \equiv \arg\max_{\mathcal{P}\subseteq\mathcal{P}^*} \left\{ \log p\left(\boldsymbol{X}|\boldsymbol{\theta}(\mathcal{P};\boldsymbol{X})\right) + pen\left(\boldsymbol{\theta}(\mathcal{P};\boldsymbol{X})\right)\right\} \quad , \tag{14}$$

where (i) $p(\boldsymbol{X}|\boldsymbol{\theta})$ admits a multiscale likelihood factorization, (ii) the estimators $\boldsymbol{\theta}(\mathcal{P};\boldsymbol{X})$ can be expressed in terms of the multiscale re-parameterization $(\rho,\boldsymbol{\omega})$, induced by $\mathcal{P}^*$, in a manner consistent with the hereditary constraints (H1) and (H2), and (iii) the penalty functional $pen(\cdot)$ is additive in the components of $\boldsymbol{\omega}$. As the optimization is over all subpartitions $\mathcal{P}$ within a given C-RP $\mathcal{P}^*$, this is analogous to a type of constrained, best-subset problem in regression analysis. We illustrate the types of estimators that we have in mind through the following three examples of data likelihood and penalty pairs.

1. Gaussian Data and Hard Thresholding.

   Suppose $X_i \sim \text{Gaussian}(\mu_i,\sigma^2)$, as in part (i) of Corollary 1. Note that $\boldsymbol{\theta} \equiv \boldsymbol{\mu}$. As remarked above, the multiscale parameters in this case are simply proportional to the UHB coefficients of $\boldsymbol{\mu}$. Therefore, a logical estimator here is that based on hard thresholding of the empirical UHB coefficients (in the spirit of standard wavelet-based estimators for this problem e.g., Donoho and Johnstone 1994), but subject to hereditary constraints. Under this strategy, a coefficient $\langle\boldsymbol{\mu},\psi_I\rangle$ is estimated by the empirical coefficient $\langle\boldsymbol{X},\psi_I\rangle$ when the latter exceeds a threshold λ in absolute value; otherwise, it is assigned the value zero. If we let $\mathcal{M}_I = 1$ denote the former case, and $\mathcal{M}_I = 0$ denote the latter case, the hereditary version of this strategy dictates that we consider only such thresholding patterns for which the $\{\mathcal{M}_I\}$ satisfy (H1) and (H2).

   This approach is equivalent to using a complexity-penalized least-squares estimator, with a penalty of the form $pen(\boldsymbol{\theta}(\mathcal{P};\boldsymbol{X})) = \lambda^2\,\#(\mathcal{P})$, where $\#(\mathcal{P})$ denotes the number of piecewise constant regions dictated by $\mathcal{P}$ (e.g., Donoho 1997, section 6.1). A simple argument shows that $\#(\mathcal{P}) = \sum_{I\in NT(\mathcal{P}^*)}\mathcal{M}_I$. Therefore, written with respect to the multiscale coefficients ω_I, it can be shown that the optimal hard-thresholding estimator is produced by solving the optimization problem

$$\max_{\mathcal{P}\subseteq\mathcal{P}^*}\left\{\sum_{I\in NT(\mathcal{P}^*)}[\omega_I^e]^2\,[\mathcal{M}_I(\mathcal{P})-1]-\lambda^2\left(\frac{c_I}{c_I'}\right)\mathcal{M}_I(\mathcal{P})\right\}\ , \tag{15}$$

where $\omega_I^e = c_I \left(X_{ch(I),r}/N_{ch(I),r} - X_{ch(I),l}/N_{ch(I),l} \right)$ is the empirical version of ω_I.

2. Poisson Data and Minimum Description Length.

   Take $X_i \sim$ Poisson(μ_i), as in part (ii) of Corollary 1. Note that, again, $\boldsymbol{\theta} = \boldsymbol{\mu}$, and in this case the multiscale parameters are related to the UHB coefficients of $\boldsymbol{\mu}$ according to (13). Complexity-penalized estimators based on Rissanen's Minimum Description Length (MDL) criterion have been devised in the context of dyadic multiscale analysis and Poisson data (Nowak and Figueiredo, 1999). MDL leads to a "keep-or-kill" estimator quite similar to the hard thresholding operation considered in the Gaussian case above. Although originally formulated for dyadic partitions, it is a straightforward matter to extend the MDL procedure of Nowak and Figueiredo to an arbitrary non-dyadic partition. Specifically, the MDL criterion leads to the following penalized likelihood optimization problem

$$\max_{\mathcal{P} \subseteq \mathcal{P}^*} \left\{ \sum_{I \in NT(\mathcal{P}^*)} \log p \left(X_{ch(I),l} | X_I, \rho_I \right) [1 - \mathcal{M}_I(\mathcal{P})] \; - \; \mathcal{M}_I(\mathcal{P}) \log(X_I + 1) \right\} \quad . \tag{16}$$

   Each term multiplied by a $[1 - \mathcal{M}_I]$ is a conditional log-likelihood of a "child" sum, given the "parent" sum, under a homogeneous (constant mean) model i.e., binomial$(X_I; \rho_I)$, where $\rho_I = N_{ch(I),l}/N_I$. Note that under the hereditary constraint $(H2)$, the total log-likelihood of the vector $(X_{i_1}, \ldots, X_{i_{N_I}})$, under homogeneity over all of I, is equal to the sum of such conditional binomial log-likelihoods corresponding to the complete partitioning of I. This in turn is equivalent to the log-likelihood of the vector conditional on its sum, which is a multinomial log-likelihood. The negative of this latter log-likelihood value is the so-called *Shannon code-length* for this vector given its sum. The terms multiplied by $\mathcal{M}_I$ are the penalty terms. Each non-trivial multiscale coefficient i.e., in which the proportion ρ_I in the homogeneous model is replaced with the empirical estimate $\omega_I^e = \frac{X_{ch(I),l}}{X_I}$ of ω_I, is encoded with $\log(X_I + 1)$ bits (the cost of encoding $X_{ch(I),l}$ given X_I). The objective of the MDL criterion is to minimize the description length (sum of the code-lengths for data and non-trivial multiscale coefficients) or equivalently, as stated above in (16), to maximize the negative of the total code-length.

3. SSPSD Data and Bayesian Priors.

Suppose the data $\boldsymbol{X}$ are distributed according to a SSPSD, as defined in Theorem 5, and consider taking a Bayesian approach wherein we estimate $\boldsymbol{\theta}$ by the maximum aposteriori (MAP) value. Analogous to most Bayesian wavelet shrinkage methods for Gaussian data (*e.g.,* see the recent volume edited by B. Vidakovic and P. Müller 1999) in which independent priors are specified for the wavelet coefficients, we "penalize" the likelihood here through the prior distribution on $\boldsymbol{\theta}$ induced by assigning independent prior distributions to the multiscale parameters $\omega_I = \theta_{ch(I),l}/\theta_I$.

The hereditary constraints (H1) and (H2) suggest using priors of the form

$$\omega_I \,|\, \mathcal{P} \;\sim\; \begin{cases} f_I^{(0)}(\omega), & \text{if } \;\mathcal{M}_I(\mathcal{P}) = 0, \\ f_I^{(1)}(\omega), & \text{if } \;\mathcal{M}_I(\mathcal{P}) = 1 \end{cases} , \tag{17}$$

where $f_I^{(0)}$ is narrow and concentrated around $\rho_I = N_{ch(I),l}/N_I$ and $f_I^{(1)}$ is broader and more dispersed. For example, because the ω_I arise as parameters of independent binomial distributions, the family of beta distributions (conjugate to the binomial) is an obvious choice for the f_I.

Regardless of the specific choice of f_I, the multiscale factorization of the SSPSD probability mass function and the use of independent priors in the multiscale parameter space lead to a similar factorization of the posterior $p(\boldsymbol{\mu}|\boldsymbol{X})$, and hence a MAP optimization problem of the form

$$\max_{\mathcal{P}\subseteq\mathcal{P}^*} \left\{ \sum_{I\in NT(\mathcal{P}^*)} \log p\left(\omega_I(\mathcal{P};\boldsymbol{X}) \,|\, X_{ch(I),l}, X_{ch(I),r}\right) \right\} , \tag{18}$$

where $\omega_I(\mathcal{P};\boldsymbol{X})$ is the MAP estimate for ω_I deriving from the "local" model with data likelihood $p(X_{ch(I),l}|X_I, \omega_I)$ and prior f_I.

Note that in each of the three examples above the estimator is linked to the solution $\hat{\mathcal{P}}$ for an optimization problem of the general form

$$\max_{\mathcal{P} \subseteq \mathcal{P}^*} \sum_{I \in NT(\mathcal{P}^*)} h\left(D_I(\boldsymbol{X})\,;\, \mathcal{M}_I(\mathcal{P})\right) \quad , \tag{19}$$

for some function $h(\cdot)$, where $D_I(\boldsymbol{X}) =^d X_{ch(I),l}|X_I$. That is, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\mathcal{P}}; \boldsymbol{X})$. Equivalently, we could pose the problem as a maximization over the multipliers $\mathcal{M}_I(\mathcal{P}) \in \{0, 1\}$, subject to the hereditary constraint. This problem may be solved in an efficient manner, as summarized in the following

Theorem 8 *The problem in (19) may be solved using an $O(N)$ "tree-pruning" algorithm.*

Proof of Theorem 8: The proof follows in a manner analogous to an argument in Donoho (1997), section 6.1. Due to additivity in $I \in NT(\mathcal{P}^*)$ and the nature of the hereditary constraints on the $\mathcal{M}_I$, the objective function in (19) has a so-called inheritance property, by which the optimal partition $\hat{\mathcal{P}}$ may be obtained in association with an optimal pruning $\hat{\mathcal{T}}$ of the full tree $\mathcal{T}^*$. Specifically, for any interval $I \in NT(\mathcal{P}^*)$, the optimal subpartition $\hat{\mathcal{P}}(I)$ on I is either (i) the union of the optimal subpartitions $\hat{\mathcal{P}}(I_{ch(I),l})$ and $\hat{\mathcal{P}}(I_{ch(I),r})$ on the children of I, or (ii) the trivial subpartition in which I is partitioned no further. Therefore, moving from fine scales to coarse (i.e., from the leaves of $\mathcal{T}^*$ to the root), we employ this strategy recursively at each stage which, upon completion (i.e., reaching the root of $\mathcal{T}^*$) leaves us with the optimal partition of $(0, N]$. But this is the same as the optimal pruning algorithm for CART (Algorithm 10.1, pg. 294), which requires on the order of N operations. $\square$

Remark 1. – In Donoho (1997), the same ideas as found above are used to compute a certain penalized least-squares estimator (analogous to that in equation (15)) using the CART algorithm. Proof of that result relies on the fact that the underlying basis (an anisotropic Haar basis, specifically) is orthogonal and that the ℓ_2 norm is invariant under transformation with respect to an orthogonal basis (i.e., Parseval's relation). The equivalent behavior is provided in our own setting by the multiscale likelihood factorizations. Both there and here the optimization problem, due to the connection with orthogonal bases, may be thought of as selection of a (hereditary) best-subset within a fixed basis.

Remark 2. – In the case that the optimization in (19) derives from a Bayesian framework, implicit in the use of an objective function like that in (18) is the assumption that all partitions $\mathcal{P} \subseteq \mathcal{P}^*$ are equally likely. If one instead wishes to place a non-uniform prior on this space, a convenient mechanism through which to do this is via the $\mathcal{M}_I$'s, for example using distributions of the form $\Pr(\mathcal{P}) = \prod_{I \in NT(\mathcal{P}^*)} \Pr(\mathcal{M}_{ch(I),l} | \mathcal{M}_I)$. However, although this choice of prior retains the tree-like indexing inherent in the data likelihood and parameter prior, the model now has the character of a hidden Markov model (on a binary tree), and the algorithm of Theorem 8 is no longer adequate. Instead, a type of "forward-backward" (or "upward-downward") algorithm can be used, but at the expense of an increase in computational complexity (see Nowak 1999 for an overview of hidden Markov models in the context of wavelet analysis).

4.2.2 Libraries of C-RP's $\mathcal{P}^*$.

Of course, use of a single, fixed C-RP $\mathcal{P}^*$ can be limiting in some problems. For example, suppose that interest is mainly in the optimal recursive partition $\hat{\mathcal{P}}$ itself, rather than the corresponding estimate $\hat{\boldsymbol{\theta}}$, and that accuracy in the locations of the points of partition is particularly desirable (e.g., in the sense of changepoints). In this case, there may be no subpartitions within a single, fixed C-RP (e.g., such as the common choice of $\mathcal{P}^*_{Dy}$) with sufficient flexibility to localize a given point of partition in an efficient manner. As an illustration, suppose that data observed on the interval $(0, 1]$ arise from a process with piecewise constant mean on the subintervals $(0, 1/3]$ and $(1/3, 1]$. Then, assuming sufficiently strong evidence in the data for this point of partition (i.e., changepoint) to be detected, a method based on recursive dyadic partitioning, say, necessarily will choose each of the points $1/2$, $1/4$, $3/8$, etc. in an effort to localize the point 1/3.

In response to this concern, it makes sense to consider working within the larger library $\mathcal{L}$ of all possible C-RPs. It is easy to check that there are $(N-1)!$ unique (as defined by the sequence of recursive partitionings) C-RPs of N points. Consider the optimization problem

$$\max_{\mathcal{P}^* \in \mathcal{L}} \left\{ \max_{\mathcal{P} \subseteq \mathcal{P}^*} \sum_{I \in NT(\mathcal{P}^*)} h\left(D_I(\boldsymbol{X})\,;\, \mathcal{M}_I(\mathcal{P})\right) \right\} \quad . \tag{20}$$

Rather remarkably, this problem too may be solved in an efficient manner, as summarized in the following

Theorem 9 *The problem in (20) may be solved using an $O(N^3)$ "message-passing" algorithm.*

Proof of Theorem 9: First note there are a total of $\frac{N(N+1)}{2}$ unique intervals $I \subseteq (0, N]$, and that of this total there are exactly $N - N_I + 1$ intervals of length N_I. Also note that any interval of cardinality $N_I = m$ may be partitioned into two children intervals in exactly $m-1$ ways. Therefore, in total, among the $(N-1)!$ possible C-RPs, there are only $\sum_{m=1}^{N} m(N-m) = \frac{N^2(N+1)}{2} - \frac{N(N+1)(2N+1)}{6} \sim \frac{N^3}{6}$ unique parent-child pairs. Our message passing algorithm exploits both this redundancy and the inheritance property underlying the algorithm of Theorem 8. Here, however, the end result of the algorithm is not only an optimal partition $\hat{\mathcal{P}}$, but an accompanying C-RP $\mathcal{P}^*(\hat{\mathcal{P}})$ as well. Beginning with intervals I of cardinality $N_I = 2$, and working recursively over $N_I = 3, 4, \ldots$, we compute $h\left(D_I(\boldsymbol{X})\,;\, \mathcal{M}_I\right)$ for each of $\mathcal{M}_I = 0$ and 1, and pass the optimal decisions (i.e., partition or not – $\mathcal{M}_I = 1$ or 0) for each interval "upward" (i.e., towards the coarsest interval $(0, N]$). That is, for a given interval I of length $N_I = m$ (which may or may not appear in the definition of the final C-RP $\mathcal{P}^*(\hat{\mathcal{P}})$), we determine and record the optimal subpartition $\hat{\mathcal{P}}(I)$ on I and the associated optimal sub-C-RP $\mathcal{P}^*(\hat{\mathcal{P}}(I))$, and we record the complexity value

$$\sum_{I' \in NT(\mathcal{P}^*(\hat{\mathcal{P}}(I))} h\left(D_{I'}(\boldsymbol{X})\,;\, \mathcal{M}_{I'}\right) \quad . \tag{21}$$

This optimal (sub)partition and its complexity value can be found via a maximization over $O(m)$ terms involving the corresponding optimal subpartitions and complexity values determined previously for each of the $m-1$ pairs of possible children of I. The maximization over all possible blocks in all possible C-RP's therefore requires roughly $N^3/6$ comparisons. Once we reach the top, we are left with $\hat{\mathcal{P}}$ and $\mathcal{P}^*(\hat{\mathcal{P}})$. A detailed description of the actual underlying algorithm is given in the appendix. $\square$

Remark 1: Non-uniqueness – There may be "ties" in the comparisons made in the maximization steps above. In such cases, multiple C-RPs (corresponding to the same final partition, but having different recursive refinements) will achieve the maximum (20).

Remark 2: Library of UHBs – Recalling the material in section 4.1, there is a one-to-one correspondence between the library $\mathcal{L}$ of C-RP's in our maximization above and the library of all possible UHBs. Furthermore, Theorem 9 implies that more general versions of the CART/Best-Ortho-Basis problem of Donoho (1997), in which the optimization with respect to a fixed partition is replaced by optimization over the library of all possible partitions, can be solved in $O(N^3)$ calculations.

Remark 3: Computational Complexity – Although the optimal partition can be calculated in $O(N^3)$ operations, in many applications one may be interested in sub-optimal solutions that require fewer calculations. Constraining the partition to be dyadic leads to an $O(N)$ optimization problem, according to Theorem 8. However, it also is possible to obtain mid complexity algorithms e.g., $O(N^2)$, by searching for (and then combining) optimal partitions in each of a handful of subregions of the data domain, which still allows for more flexibility than a strictly dyadic structure. Yet another possibility is to consider all possible partitions, but rather than search for the optimal partition, employ a greedy top-down algorithm that operates in a sequential, coarse-to-fine fashion. A greedy algorithm like this was proposed for partitioning Poisson data in Nowak and Figueiredo (1999). It is easy to show that greedy algorithms such as this require $O(LN)$ calculations, where $L \leq N$ is the number of intervals in the final partition.

Remark 4: Libraries of Partitions and Priors – Similar to Remark 2 following Theorem 8, for the optimization (20) to arise in a Bayesian context (i.e., through extension of equation (18) to maximization over $\mathcal{L}$), there is the implicit assumption that a uniform prior has been placed on the space $\mathcal{L}$ and on all partitions $\mathcal{P} \subset \mathcal{P}^* \in \mathcal{L}$. Formulation of a non-uniform (and possibly less informative) prior here is a non-trivial task, as evidenced by recent work on Bayesian CART (Denison, Mallick, and Smith 1998; Chipman, George, and McCulloch 1998) and related tree models (Chipman, George, and McCulloch 2000). At issue is the fact that many partitions may have very similar structures, and hence uniformly assigning prior mass may in fact lead to "clumps" of mass on nearly equivalent partitions. It may be possible (and desirable) to devise non-uniform priors that more evenly distribute prior mass over the space of partitions.

Remark 5: Finally, we comment on an intriguing opportunity for adapting our framework to the problem of density estimation, using our results for SSPSD's. Specifically, beginning with a sample of size M from an unknown density and the definition of N bins at some very fine resolution, by combining a multiscale likelihood factorization for the multinomial distribution with the algorithms of Theorems 8 or 9, there results a method of histogram-based density estimation, with adaptive selection of (potentially varying) bin-widths. Such estimators would seem to be multiscale analogues of the various generalized histogram estimators proposed in the literature, such as those of Rissanen, Speed, and Yu (1992).

4.3 Evaluation of Estimator Performance.

Our interest in this paper has been in the existence and derivation of a certain class of multiscale likelihood factorizations and their usage for problems of nonparametric inference.

In this section we have proposed a general class of estimators deriving from maximization of complexity-penalized likelihoods, all of which may be viewed as instances of a certain prototypical optimization over either a fixed C-RP $\mathcal{P}^*$ or the library $\mathcal{L}$. Through exploitation of fundamental connections between partitions, bases, and factorizations in our particular multiscale context, we were able to show that these optimizations may in fact be solved using computationally feasible algorithms. Recalling the characterization results of section 3, we see that this single, unified framework pertains equally to a variety of models that include cases of continuous, count, and categorical data.

In the same spirit, it is desirable to pursue a similarly unified approach to other related issues, such as the determination of risk bounds describing estimator performance. Of course, for the special case of a Gaussian model and hard thresholding, there are results in the literature on wavelets that may be adapted with little change. For example, the proof of Theorem 7.1 in Donoho (1997) is in fact general enough to yield an immediate oracle inequality for the estimator defined by the optimization in (15) and its obvious extension to (20). But for the level of generality we seek, a more natural approach is to use adaptations of the methods of Barron, Birgé, and Massart (1999), which overcome a number of technical difficulties in producing analogues of the Gaussian/ℓ^2 risk bounds in fairly general non-Gaussian settings (including, in particular, likelihood-based contexts). Risk bounds for certain estimators of the sort introduced above appear in the companion to this paper (Kolaczyk and Nowak 2000).

5 A Numerical Illustration.

As an illustration of the potential for applications inherent in the results of the previous section, consider the problem of segmenting a Poisson time series. Such time series are common to the field of high-energy astrophysics (e.g., X-ray and γ-ray), for example, where they are produced typically through aggregation of photon arrival times into intervals of equal length at some fine resolution (generally dictated by both instrumental and scientific considerations). Figure 1 shows an example generated by a so-called gamma-ray burst (GRB), recorded by the BATSE instruments on board NASA's late Compton Gamma Ray Observatory (CGRO). Widely described as one of modern astronomy's greatest mysteries, the physical mechanisms underlying GRB's still remain poorly understood. In fact, the time series resulting from these phenomena often have been compared to snowflakes, in that "no two are alike." Therefore, segmentation of these and other similar time series, under minimal model assumptions, can be an important part of their analysis, either as a pre-processing step (e.g., for detecting potentially interesting structure in long data streams) or at later stages (during which such structure might receive closer attention). See, for example, Scargle (1998) for a recent characterization of the segmentation problem in astronomy.

The time series in Figure 1 is shown with the optimal segmentation produced by the MDL method described in Example 2 of section 4.2, using the fully adaptive method of section 4.2.2 (i.e., library search over all C-RP's). This method in fact corresponds to a MAP estimator within the context of Example 3 as well, in the special case where the SSPSD is the Poisson distribution, and where $f_I^{(0)}$ is taken to be a point mass at ρ_I and $f_I^{(1)}$ is the uniform distribution on $[0,1]$ for all I. A number of closely spaced segmentation points have been chosen to describe the sharply varying structure in the interval between 0.5 and 1.0 seconds. Yet, at the same time, the relatively slow rate of change in the intervals prior to 0.5 seconds and after 1.0 seconds is captured through the use of just two points. This sort of flexibility in producing the overall segmentation results from the multiscale nature of the underlying modeling framework, as well as the use of a full library of C-RP's (i.e., as opposed to, say, simply a dyadic C-RP).

As a means of gaining some insight into the reliability of such segmentations, we performed a small simulation study. The standard 'Blocks' test signal of Donoho and Johnstone (1994) was scaled so as to have a dynamic range more similar to that underlying the data in Figure 1 (resulting in a minimum of 60, maximum of 132). Poisson times series were sampled from the resulting intensity using $n = 256$ equispaced bins, over a total of 1000 trials, and the location of declared segmentation points was noted for each trial. Figure 2(a) shows, for those locations having been declared a point of segmentation at least once during the 1000 trials, the relative frequency with which such a declaration was made. For comparison, in Figure 2(b) is shown the analogous results from the use of the same estimation method restricted to dyadic partitioning. The dyadic method uses a much greater proportion of locations much more frequently in arriving at a segmentation for the 'Blocks' signal than does the fully adaptive method. The fully adaptive method is found to be quite accurate, in that it finds almost all segmentation points in 'Blocks' the vast majority of the time, although it would seem in turn to have a greater degree of highly infrequent but spurious declarations than in the dyadic case. Both results arguably can be seen as stemming from the presence or lack of constraints associated with the use or not of a single C-RP.

6 Discussion.

6.1 A Basic Paradigm.

At the heart of this paper is the fundamental idea of pursuing factorizations of a likelihood in analogy to orthonormal wavelet decompositions. Such factorizations have been used previously by the authors and colleagues in designing specific methodologies for analysis of Poisson time series (Kolaczyk 1999b) and images (Timmerman and Nowak 1999), Poisson linear inverse problems (Nowak and Kolaczyk 2000), and the spatial analysis of continuous and count data in Geography (Kolaczyk and Huang 2000). Our aim in this present paper has been to establish a general conceptual foundation for such methods.

In doing so, we clearly have been influenced not only by the technical structure associated with wavelet decompositions, but also by the basic underlying paradigm, something that has come to be called by some authors the *computational harmonic analysis (CHA)* paradigm (e.g., Donoho 1999). The principle put forth by advocates is that, given a class of "objects," one pursues an "optimal" representation for these objects and accompanying "fast" algorithms for computational tasks. In the field of CHA, "objects" generally are members of some functional class, and therefore representations/algorithms deriving from this community (e.g., wavelets, wavelet packets, etc.) have been put to immediate use by statisticians, signal/image processors, and the like in various tasks associated with the standard "signal plus noise" model. In a similar spirit, we have worked from the premise that, if the specification of a likelihood is a natural first step in a given statistical problem (surely not an uncommon event!), then arguably that itself is the relevant "object" when considering alternative representations – particularly if the efficiency of algorithms is an accompanying concern.

On a related note, we point out the interesting distinction between models deriving from *analysis,* and those from *synthesis* (borrowing terminology from the signal processing literature). For example, from this perspective, a method beginning with the standard "signal plus noise" model and then transforming to the "wavelet domain" (i.e., in the spirit of CHA) is intrinsically a method of "analysis". On the other hand, one may specify a model in the "time domain" implicitly as the result of applying an inverse wavelet transform to a collection of random coefficients – i.e., the model is "synthesized" (e.g., Nason, von Sachs, and Kroisandt 2000). In the same spirit, the framework described herein may be viewed as a formalism for obtaining multiscale probability models on graphs by "analysis" while, conversely, similar structures may be derived via *a priori* specification as well. Such models by "synthesis" have been described in a number of papers by Alan Willsky and colleagues (e.g., see Basseville *et al.* 1992), and by Huang and Cressie (2000).

6.2 Connections to Other Areas.

Our multiscale likelihood factorizations, and their accompanying MRA, arise at the intersection of ideas and structures from a variety of areas and hence, in turn, would seem to have the potential for some degree of reciprocal impact in those areas. For example, the role played by recursive partitioning and tree-based structures in our framework, and the existence of simple, scalable algorithms for searching over spaces of varying complexity (with inversely varying efficiency), suggest the potential for developing similar likelihood-based methods of relevance to the machine learning and data mining communities.

Similarly, if the multiple changepoint problem is viewed from the perspective of partitioning (e.g., Barry and Hartigan 1993), the extraction of a set of changepoints from a dataset is implicit in our framework, as illustrated in section 5. The multiscale perspective can be viewed in this context as "decoupling" the decision process inherent in choosing successive changepoints. In fact, it is interesting to note that the asymptotic method proposed by Akman and Raftery (1986) for the detection of a single changepoint in a Poisson process actually is based on the size of what turns out to be the maximum empirical UHB coefficient of the process among all such coefficients that can be formed at the coarsest scale (i.e., from UHB wavelets of full support on $(0, N]$). As noted in section 4.2, however, to obtain the decoupling that accompanies our likelihood factorizations, one is pointed toward certain functions of the UHB coefficients instead of the coefficients themselves (e.g., equation (13)).

Finally, likelihood factorizations in general, of course, have an extensive history in the graphical models literature (e.g., see Lauritzen 1996, and the references therein). From the perspective of this literature, our MRA may be viewed as a set of sufficient conditions for ensuring a particular type of recursive factorization. And on a related note, if our factorizations are paired with an appropriate prior probability structure, one may view the result as a type of so-called Bayesian network, with an inherent multiscale structure.

6.3 Extensions of the Basic MRA.

As mentioned previously, just as the notion of an MRA for wavelets has undergone various generalizations, so too might one entertain variations on the conditions underlying our MRA for likelihoods. For example, the notion of reproducibility in condition (B*) might be loosened, such as was the case implicitly with the factorization of SSPSD likelihoods detailed in Theorem 5. Alternatively, one might envision replacing the hierarchy of partitions of $(0, N]$ in condition (A*) by a similar hierarchy of coverings (i.e., allowing overlap of the intervals I). In conjunction with such a move might go as well the replacement of the definition of the X_I as simple summations of data X_i with indices $i \in I$ by more general schemes of weighted summation (or even nonlinear aggregations).

An obvious example in which such a framework would arise is the case in which a likelihood factorization is sought with respect to the wavelet coefficients of a vector $\boldsymbol{X}$, for wavelets smoother than those of a Haar system. The overlapping support of orthogonal wavelets of neighboring locations and scales generally implies a statistical dependency among the coefficients (though not in the special case of independent Gaussian data, due to orthogonality of the wavelets). Hence, we cannot expect a factorization of the likelihood with respect to a graph whose nodes are indexed by the (j,k) location-scale indexing of the wavelets. Instead, a natural extension is to pursue factorizations with respect to a so-called *clique tree* (e.g., Pearl 1988), in which the nodes now are occupied by subsets of coefficients with "localized" non-empty intersections.

There also remains the notion of "cuts" in condition (D*). It is this condition that induces the multiscale parameterization to "decouple" in parallel with the likelihood factorization, and from which in part, therefore, the efficiency of the algorithms in section 4.2 derives. Although it is not clear that such a complete decoupling is necessary (i.e., in one-to-one correspondence with the components in the factorization), it would seem that some sort of restriction to only localized coupling (i.e., with respect to the topology of the underlying graph) of parameter components is desirable.

Lastly, although we have concentrated on one-dimensional (sequential) settings here, much of our framework can be extended to two or more dimensions (e.g., images). For example, the recursive partitioning of the interval $(0, N-1]$ has a straightforward extension to the two-dimensional lattice $(0, N-1] \times (0, N-1]$, in which each rectangular block (rather than interval) may be partitioned into four smaller rectangles (Donoho 1997 considers the special dyadic case in detail). With appropriate modifications of (A*)-(D*) (replace intervals I with blocks B), a factorization analogous to that in Theorem 2 holds in the two-dimensional case, with the scalar parameters $\{\omega_I\}$ replaced by three-dimensional vector parameters $\{\boldsymbol{\omega}_B\}$, reflecting the three degrees of freedom at one's disposal in splitting the total sum over a block into four partial sums over the four sub-blocks. A CART-type algorithm analogous to the one described in Theorem 8 can also be devised for higher dimensional problems. In this case, we have an $O(M)$ complexity, where $M = N \times N$, the total number of pixels in the image. Another intriguing possibility is to take advantage of the additional flexibility in splitting that one has in higher dimensions. For example, the *wedgelet* analysis developed by Donoho (1999) allows one to incorporate non-rectangular splitting (split a square block into two half "wedges" rather than four square sub-blocks) at the terminal nodes of a dyadic partition structure. The CART-based wedgelet analysis algorithm developed there can be employed in conjunction with our multiscale likelihood factorizations as well. Moreover, the optimization over the library of all possible (interval) partitions in equation (20) also can be extended to higher dimensions. For instance, in two dimensions one has a library of partitions containing a total of $(N(N+1)/2)^2$ unique rectangles. Arguing along similar lines to the proof of Theorem 9, its not difficult to see that an $O(M^3)$ algorithm can be used to find the optimal two-dimensional partition.

7 Appendix A — Message Passing Algorithm

We describe here in detail the algorithm underlying Theorem 9, which computes the optimal partition in the set of all possible partitions. We adopt a two-step procedure. In Step 1, the maximization in (20) is computed for every possible subinterval, and the maximum value is recorded. By working from short subintervals to long subintervals (bottom-up), the maximization can be computed for each subinterval $I \subset (0, N]$ using results previously computed for smaller subintervals contained within a given subinterval. This reduces the computation to $O(N)$ operations for each subinterval and, as shown in the proof of Theorem 9, there are $O(N^2)$ subintervals. Thus, the overall complexity of this first step is $O(N^3)$. In Step 2, we work from top-down, extracting the optimal partition by comparing the maximal values computed in Step 1. The algorithm is reminiscent of a message passing algorithm (Pearl 1988); in Step 1 we pass information *up* from short subintervals to long subintervals, and in Step 2 we work back *down* to extract the optimal partition. Pseudo-code for the algorithm is given below.

Up Step $O(N^3)$:

Initialize: For $i = 1, \ldots, n$, set $C_0(X_i) = 0$, $C_1(X_i) = 0$, and $C(X_i) = 0$. For all $1 \leq i \leq k < j \leq n$, set $I = [i, \ldots, j]$ and $ch(I) = [i, \ldots, k]$, and define

$$l_0(i:k|i:j) \equiv h(D_I(\boldsymbol{X}); \mathcal{M}_I = 0),$$

$$l_1(i:k|i:j) \equiv h(D_I(\boldsymbol{X}); \mathcal{M}_I = 1), \tag{22}$$

where the precise form of $h(D_I(\boldsymbol{X}); \mathcal{M}_I)$ depends on the specific optimization problem at hand (see (19) and related discussion). Also define

$$\boldsymbol{X}_{i:j} \equiv \{X_i, \ldots, X_j\}.$$

Compute: The (local) maximization over each subinterval $I = [i, \ldots, j]$ can be computed recursively (bottom-up) as follows. $C(\boldsymbol{X}_{i:m})$ denotes the corresponding maximum value.

For $m = 1, \ldots, n-1$ $\quad$ ($m =$ "scale" $=$ length of subinterval)

For $i = 1, \ldots, n-m$ $\quad$ ($i =$ "position" $=$ start of subinterval)

Inhomogeneous Case: The "block" $\boldsymbol{X}_{i,m}$ is partitioned into two or more "sub-blocks". The optimal partition point is given by

$$k^*_{i,m} \equiv \arg \max_{k=i:m-1} l_1(i:k|i:i+m) + C(\boldsymbol{X}_{i,k}) + C(\boldsymbol{X}_{k+1,m}).$$

where $C(\boldsymbol{X}_{i,k})$ and $C(\boldsymbol{X}_{k+1,m})$ are the maximum values for the two sub-blocks computed at scales $< m$, as defined below in (23). The maximum value achieved by splitting $\boldsymbol{X}_{i,m}$ into two sub-blocks is then

$$C_1(\boldsymbol{X}_{i,m}) \equiv l_1(i:k^*_{i,m}|i:i+m) + C(\boldsymbol{X}_{i,k^*_{i,m}}) + C(\boldsymbol{X}_{k^*_{i,m}+1,m}),$$

Homogeneous Case: The maximization over k is unnecessary since the block $\boldsymbol{X}_{i,m}$ will not be partitioned. Thus, the value we assign for not splitting $\boldsymbol{X}_{i,m}$ is given by

$$C_0(\boldsymbol{X}_{i,m}) \equiv l_0(i:i|i:i+m) + C_0(\boldsymbol{X}_{i,k}) + C_0(\boldsymbol{X}_{k+1,m}).$$

Note that we do not make use of $l_0(i:k|i:i+m)$, $k \neq i$, and so these quantities need not be computed in (22) above.

Maximum of Two Cases: Define

$$C(\boldsymbol{X}_{i,m}) \equiv \max_{q=0,1} C_q(\boldsymbol{X}_{i,m}). \tag{23}$$

Down Step $O(N)$:

Initialize: `partition_points` = empty

Extract: `partition_points` = get_partition_points(`partition_points`, $1, n$)

The function get_partition_points recursively extracts the optimal partition points. get_partition_points starts by finding the first optimal partition point for full sequence $\boldsymbol{X}_{1,n} = \{X_1, \ldots, X_n\}$ (unless it is the $C_0(\boldsymbol{X}_{1,n}) \geq C(\boldsymbol{X}_{1,n})$, in which case the function terminates), and then calls itself to test the resulting sub-blocks. get_partition_points(`partition_points`, $1, n$) returns a vector of the optimal partition points (unordered).

function **get_partition_points:**

`partition_points` = get_partition_points(`partition_points`, i, m)

if $C_0(\boldsymbol{X}_{i,m}) < C(\boldsymbol{X}_{i,m})$

`partition_points` = [`partition_points`, $B(\boldsymbol{X}_{i:m})$]

`partition_points` =

get_partition_points(`partition_points`, $i, B(\boldsymbol{X}_{i:m})$)

`partition_points` =

get_partition_points(`partition_points`, $B(\boldsymbol{X}_{i:m}) + 1, m$)

end

References.

Akman, V.E. and Raftery, A.E. (1986). Asymptotic inference for a change-point Poisson process. *Ann.Statist.* **14** 1583 - 1590.

Bar-Lev, S.K. and Enis, P. (1986). Reproducibility and natural exponential families with power variance functions. *Ann. Statist.* **14** 1507-1522.

Barndorff-Nielsen, O. (1976). Factorization of likelihood functions for full exponential families. *J. Roy. Statist. Soc. Ser. B.* **38** 37-44.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families in Statistical Theory.* Wiley: New York.

Barron, A.R., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory and Relat. Fields* **113** 301-413.

Barry, D. and Hartigan, J. A. (1993). "A Bayesian Analysis for Change Point Problems." *J. Am. Stat. Assoc.* **88** 309-319.

Basseville, M., Benveniste, A., Chou, K.C., Golden, S.A., Nikoukhah,, R., and Willsky, A.S. (1992). "Modeling and Estimation of Multiresolution Stochastic Processes," *IEEE Trans. Inform. Theory* **38** 766-784.

Burt, P. and Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Trans. Comm.* **31** 482-540.

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

Chipman, H., George, E.I., and McCulloch, R.E. (1998). Bayesian CART model search (with discussion). *J. Am. Stat. Assoc.* **93** 935-960.

Chipman, H., George, E. I. and McCulloch, R. E. (2000). Bayesian Treed Models, Technical Report, Department of Statistics, University of Texas-Austin.

Daubechies, I. (1992), *Ten Lectures on Wavelets,* Philadelphia, Pennsylvania: SIAM.

Denison, D.G.T, Mallick, B.K., and Smith, A.F.M. (1998). A Bayesian CART algorithm. *Biometrika* **85** 363-377.

DeVore, R.A. and Lucier, B.J. (1992). Fast wavelet techniques for near-optimal image processing. IEEE-ICASSP-93, pgs. 48.3.1–48.3.7. IEEE Military Communications Conference, New York, NY.

Donoho, D.L. (1997), "CART and best-ortho-basis selection: A connection." *Ann. Statist.* **25** 1870-1911.

Donoho, D.L. (1999). Wedgelets: Nearly minimax estimation of edges. *Ann. Statist.* **27** 859 - 897.

Donoho, D.L. and Johnstone, I.M. (1994), Ideal spatial adaptation via wavelet shrinkage, *Biometrika* **81** 425-455.

Engel, J. (1994), "A simple wavelet approach to nonparametric regression from recursive partitioning schemes." *J. Multivariate Anal.* **49** 242-252.

Girardi, M. and Sweldens, W. (1997), "A New Class of Unbalanced Haar Wavelets that Form an Unconditional Basis for L_p on General Measure Spaces." *J. Fourier Anal. Appl.* **3:4** 457-474.

Haar, A. (1910). Zur theorie der orthogonalen funktionen-systeme. Math. Ann. **69** 331-371.

Huang, H-C. and Cressie, N. (2000). Multiscale graphical modeling in space: applications to command and control. In *Spatial Statistics and Applications,* Moore, M. (ed.). Springer, NY.

Joshi, S.W. and Patil, G.P. (1968). A class of statistical models for multiple counts. In *Random Counts in Scientific Work,* Vol. 2, pp. 189-203. Pennsylvania State University Press.

Kolaczyk, E.D. (1999a). Some observations on the tractability of certain multi-scale models. In *Bayesian Inference in Wavelet-Based Models,* Vidakovic, B. and Müller, P. (eds). Springer-Verlag: New York.

Kolaczyk, E.D. (1999b). Bayesian multiscale models for Poisson processes. *J. Am. Stat. Assoc.* **94** 920-933.

Kolaczyk, E.D. and Huang, H. (2000). Multiscale statistical models for hierarchical spatial aggregation. *Geographical Analysis,* (to appear).

Kolaczyk, E.D. and Nowak, R.D. (2000). Complexity penalized estimation using multiscale likelihood factorizations: risk bounds and asymptotics. In preparation.

Lauritzen, S. L. (1996). *Graphical Models.* Oxford University Press: New York.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. PAMI* **11** 674 - 693.

Meyer, Y. (1985). *Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs.* Séminaire Bourbaki, 1985-1986, no. 662.

Morris, C.N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.* **10** 65-80.

Nason, G.P., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J. Roy. Statist. Soc. Ser. B.* **62** 271-292.

Nowak, R.D. (1999). Multiscale hidden Markov models for Bayesian image analysis. In *Bayesian Inference in Wavelet-Based Models,* Vidakovic, B. and Müller, P. (eds). Springer-Verlag: New York.

Nowak, R. and Figueiredo, M. (1999), "Unsupervised progressive parsing of Poisson fields using minimum description length criteria," in *Proceedings of IEEE Conference on Image Processing*, Kobe, Japan.

Nowak, R.D. and Kolaczyk, E.D. (2000). A Bayesian multiscale framework for Poisson inverse problems. *IEEE Trans. Inform. Theory* **46:5** 1811-1825.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann: San Mateo, CA.

Rissanen, J., Speed, T.P., and Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. Inform. Theory* **38** 315-323.

Scargle, J.D. (1998). Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, a New Method to Analyze Structure in Photon Counting Data. *The Astrophysical Journal,* 504, 405 - 418.

Sweldens, W. (1997). The lifting scheme: A construction of second generation wavelets. *Siam J. Math. Anal.* **29:2** 511-546.

Stromberg, J.O. (1982). A modified Franklin system and higher order spline systems on $\mathbb{R}^n$ as unconditional bases for Hardy spaces. Conf. in honor of A. Zygmund, Vol. II, W. Beckner et al., ed., Wadsworth Math. Series, 475-493.

Timmerman, K. and Nowak, R.D. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Tran. Inform. Theory* **45** 846-862.

B. Vidakovic and P. Müller, Editors (1999). *Bayesian Inference in Wavelet Based Models.* Springer-Verlag.

Wilks, S.S. (1962). *Mathematical Statistics.* Wiley: New York.

Witkin, A. (1983). Scale space filtering. *Proc. Internat. Joint Conf. Artificial Intelligence.*

DEPARTMENT OF MATHEMATICS
AND STATISTICS
BOSTON UNIVERSITY
BOSTON, MA 02215
kolaczyk@math.bu.edu

DEPARTMENT OF ELECTRICAL
AND COMPUTER ENGINEERING
RICE UNIVERSITY
HOUSTON, TX 77251-1892
nowak@rice.edu

Caption for Table 1.

Table 1. – Side-by-side comparison of conditions for a wavelet MRA (left) and likelihood MRA (right).

Wavelet MRA	**Likelihood MRA**
(A) Hierarchy of Nested Subspaces	(A*) Hierarchy of Recursive Partitions
(B) Orthonormal Basis within V_0	(B*) Independence within $\mathcal{P}_N$
(C) Scalability Between Subspaces	(C*) Reproducibility Between Subpartitions
(D) Translation within Subspaces	(D*) "Decoupling" of Parameters with Partitions (i.e. Cuts)

Table 1:

CAPTIONS FOR FIGURES.

Figure 1. – Fully adaptive segmentation of an astronomical gamma-ray burst (GRB). Time series results from aggregation of original photon arrival times into $n = 256$ equispaced bins for burst #845 (photon energies 25 keV - 1 MeV), collected by the BATSE instruments of NASA's Compton Gamma Ray Observatory. Segmentation is based upon a Poisson model and minimum description length criterion, as in Example 2 of section 4.2, using the library search described in section 4.2.2.

Figure 2. – Stem plots of the relative frequency with which locations were declared a change-point, over a total of 1000 trials, in a simulation study of the segmentation of 'Blocks' from Poisson observations. (a) Results from fully adaptive method (i.e., library search); (b) results from dyadic method. On both plots there is overlaid a scaled plot of the 'Blocks' function (dotted) for reference.

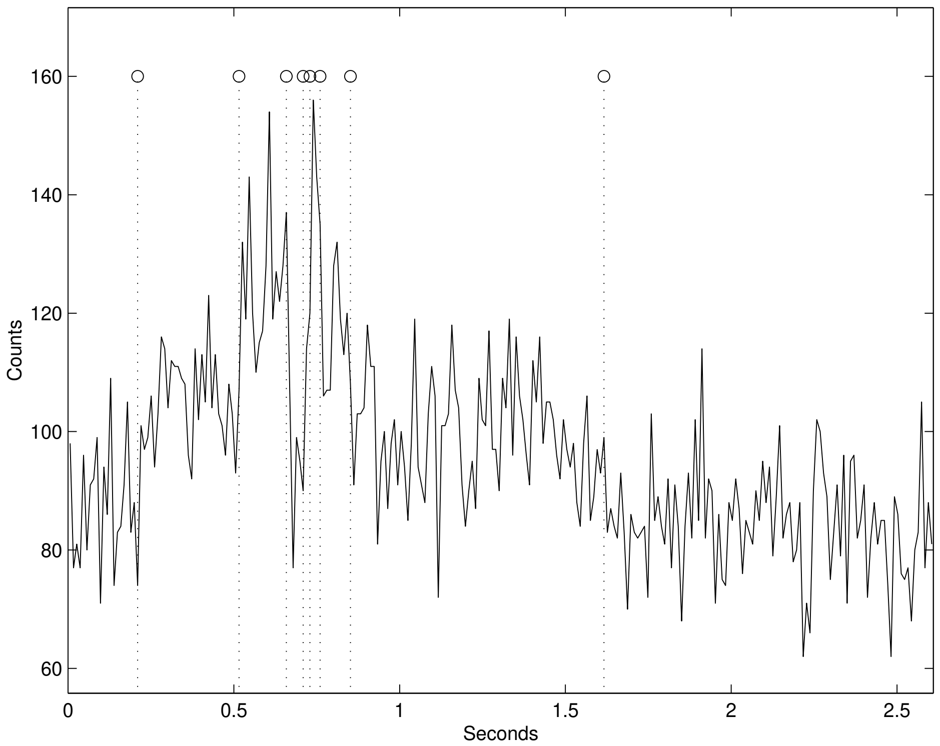

Figure 1:

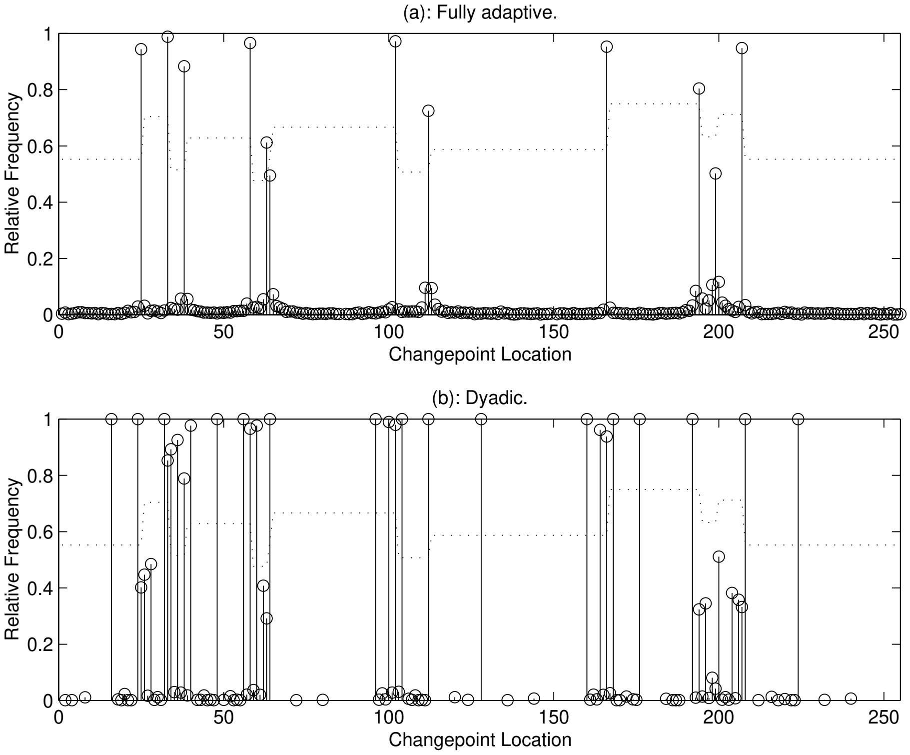
(a): Fully adaptive.
Relative Frequency
Changepoint Location
(b): Dyadic.
Relative Frequency
Changepoint Location

Figure 2: