# Multiresolution Markov Models for Signal and Image Processing

ALAN S. WILLSKY, FELLOW, IEEE

*Contributed Paper*

*This paper reviews a significant component of the rich field of statistical multiresolution (MR) modeling and processing. These MR methods have found application and permeated the literature of a widely scattered set of disciplines, and one of our principal objectives is to present a single, coherent picture of this framework. A second goal is to describe how this topic fits into the even larger field of MR methods and concepts—in particular, making ties to topics such as wavelets and multigrid methods. A third goal is to provide several alternate viewpoints for this body of work, as the methods and concepts we describe intersect with a number of other fields.*

*The principle focus of our presentation is the class of MR Markov processes defined on pyramidally organized trees. The attractiveness of these models stems from both the very efficient algorithms they admit and their expressive power and broad applicability. We show how a variety of methods and models relate to this framework including models for self-similar and $1/f$ processes. We also illustrate how these methods have been used in practice.*

*We discuss the construction of MR models on trees and show how questions that arise in this context make contact with wavelets, state space modeling of time series, system and parameter identification, and hidden Markov models. We also discuss the limitations of tree-based models and algorithms and the artifacts that they can introduce. We describe when these are of concern and ways in which they can be overcome. This leads to a discussion of MR models on more general graphs and ties to well-known and emerging methods for inference on graphical models.*

***Keywords*—*Autoregressive processes, Bayesian networks, data assimilation, data fusion, estimation, fractals, geophysical signal processing, graphical models, hidden Markov models, image enhancement, image processing, image segmentation, inverse problems, Kalman filtering, machine vision, mapping, Markov random fields, maximum entropy methods, multiresolution (MR) methods, quadtrees, signal processing, sparse matrices, state space methods, stochastic realization, trees, wavelet transforms.*

Manuscript received November 2, 2001; revised April 4, 2002. This work was supported in part by the Air Force Office of Scientific Research under Grant F49620-00-0362, the Office of Naval Research under Grant N00014-00-0089, and Office of the Deputy Under Secretary of Defense for Laboratories MURI under ARO Grant DAAD19-00-0466.

The author is with the Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: willsky@mit.edu).

Publisher Item Identifier 10.1109/JPROC.2002.800717.

## NOMENCLATURE

| | |
|---|---|
| $\mathcal{G}$ | Graph. |
| $\mathcal{V}$ | Node or vertex set of a tree or graph. |
| $\mathcal{V}_s$ | Set of nodes in the subtree rooted at node $s$ (i.e., node $s$ and all its descendents). |
| $\mathcal{E}$ | Edge set of a graph. |
| $\mathcal{A}, \mathcal{U}, \mathcal{W}$ | Subsets of nodes in a graph. |
| $\mathcal{C}$ | Clique in a graph. |
| $\mathbf{C}$ | Set of all cliques of a graph. |
| $s,\ t,\ u$ | Nodes on trees and graphs. |
| $s\alpha_1,\ s\alpha_2,\ s\alpha_i$ | Children of node $s$ on a tree. |
| $s\overline{\gamma}$ | Parent of node $s$ on a tree. |
| $s \wedge t$ | Closest common ancestor to nodes $s$ and $t$ on a tree. |
| $m$ | Index for scale in a MR representation. |
| $m(s)$ | Scale of node $s$ in a tree. |
| $A(s), C(s), Q(s), R(s)$ | Matrices used to define MR models on trees. |
| $x(s), y(s), w(s), v(s)$ | Random variables or vectors at node $s$ in a tree or graph. |
| $x_{\mathcal{A}}$ | Collection or vector of the variables $\{x(s) \vert s \in \mathcal{A}\}$. |
| $x, y, w, v$ | Collection or vectors of variables over an entire tree or graph. |
| $P_x(s)$ | Prior covariance of $x(s)$ in an MR model. |
| $\hat{x}_s(s)$ | Smoothed estimate of $x(s)$ in an MR model. |
| $P_e(s)$ | Covariance of the error in the estimate $\hat{x}_s(s)$. |
| $\hat{x}(s \vert s)$ | Estimate of $x(s)$ based on data in $\mathcal{V}_s$. |
| $P(s \vert s)$ | Covariance of the error in the estimate $\hat{x}(s \vert s)$. |
| $\hat{x}(s \vert s-)$ | Estimate of $x(s)$ based on all of the data in $\mathcal{V}_s$ except the measurements at node s. |

| | |
|---|---|
| $P(s\|s-)$ | Covariance of the error in the estimate $\hat{x}(s\|s-)$. |
| $\hat{x}(s\|s\alpha_i)$ | Estimate of $x(s)$ based on data in $\mathcal{V}_{s\alpha_i}$. |
| $P(s\|s\alpha_i)$ | Covariance of the error in the estimate $\hat{x}(s\|s\alpha_i)$. |
| $P_x$ | Prior covariance of the vector $x$. |
| $\hat{x}$ | Optimal estimate of $x$. |
| $P_e$ | Covariance of the error in the estimate $\hat{x}$. |
| $\boldsymbol{r}$ | Spatial variable in two or three dimensions. |
| $\boldsymbol{I}$ | Two-dimensional planar region. |
| $z(\boldsymbol{r})$ | Height of a surface over the 2-D region $\boldsymbol{I}$. |
| $(p(\boldsymbol{r}),\, q(\boldsymbol{r}))$ | Gradient of the surface $z(\boldsymbol{r})$. |

## I. Introduction

Multiresolution (MR) concepts and methods for the statistical analysis of phenomena and data have been and remain topics of tremendous interest in a wide variety of disciplines (see, for example, two special issues devoted to this subject [87] and [194] as well as the book [304]). The reasons for the intensity of activity and the dizzying variety of methods that have been developed are myriad, and it is not the intent of this paper to put this entire subject into one simple and coherent picture. Rather, our objective is to provide an introduction to one significant component of this vast field that has provided fertile ground for both theory and application. Moreover, we have personally found that the perspective that this framework yields provides a very useful platform for organizing one's understanding of the broader field of MR analysis and processing.

One of the distinguishing characteristics of the framework we describe is that it does not start with *algorithms* for processing or analyzing phenomena at multiple resolutions—e.g., as in the use of wavelet transforms to produce decompositions of signals at multiple resolutions—but rather begins with the *modeling* of phenomena at multiple resolutions. Much as in the development of methodologies for modeling time series or random fields, the intent is to construct statistical models that: 1) are rich enough to capture large and important classes of phenomena of broad interest; 2) possess structure that can be exploited both to gain insight into these phenomena and to design powerful classes of algorithms; and 3) provide statistical tools for analyzing with precision both when these models are appropriate and how well the resulting algorithms perform.

### A. What is it That is MR?

A principle objective of this modeling framework is capturing the several important ways in which a data analysis or signal processing problem can have MR characteristics. The first is that the *phenomenon* that is to be modeled can exhibit distinctive behavior over a range of scales or resolutions. For example, many physical processes—e.g., geophysical fields such as atmospheric or oceanographic phenomena—possess behavior over vast ranges of spatial or spatio–temporal scales [31], [112], [198], [219], [220], [352], [354]. Studies of large classes of natural imagery also show characteristic variability at multiple scales [46], [47], [140], [157], [218], [243], [250], [261], [268], [281], [297]–[300], [333], as do mathematical models of self-similar or fractal processes [288] such as fractional Brownian motion (fBm) [30], [83], [116], [232], [313], motivating examinations of the properties of the wavelet transforms of such signals and images [69], [83], [102], [114], [117], [154], [176], [191], [235], [273], [293], [320], [346]–[350], [359].

Second, whether the phenomenon displays MR behavior or not, it may be the case that the *available data* are at multiple resolutions. While this might be the simple result of transforming the data—e.g., using wavelet transforms—there are also many problems in which the collected data directly measure the quantities of interest at multiple resolutions. For example, large-scale data assimilation problems in the geosciences quite frequently involve the fusion of several distinct sources of data, representing not only very different measurement phenomenologies but also probing a geophysical medium at very different resolutions. One example is the fusion of satellite measurements [112], [113], [354] of oceanographic variables with measurements made from surface ships [162], [227], [242], [331] and perhaps also data from oceanacoustic tomographic collections [241], [253]. Similar examples can be found in a variety of other problems involving remotely sensed or probed data including the fusion of synthetic aperture radar (SAR) imagery [77] and geophysical inversion and data fusion [84], [139], [198], [247]. In addition, advances in biomedical sensing [317] require the development of new methods for fusing data sets with very different characteristics (e.g., positron emission tomography (PET) and magnetic resonance imaging (MRI) images).

Third, whether or not the phenomenon or the data are MR, it may be the case that the *objectives* of the user or users may be at multiple resolutions. This is certainly the case in large-scale geophysical mapping in which different scientific studies focus on behavior over different ranges of scales so that the variability of concern to one scientist is simply "noise" to another. In addition, in many contexts other than those of pure scientific inquiry, the objective of data assimilation can be stated at very high levels: the mapping of an oil reservoir to assess production rates and total yield or the characterization of the threat of subsurface contaminants to populated areas. One also finds this in military applications in which maps of both environmental and situational variables (e.g., maps of terrain elevation and vegetation for the former and of the disposition of friendly and unfriendly forces for the latter) are required by multiple users: typically large-scale maps at comparatively coarse scales by strategic planners and much more localized finer-scale maps by tactical forces.

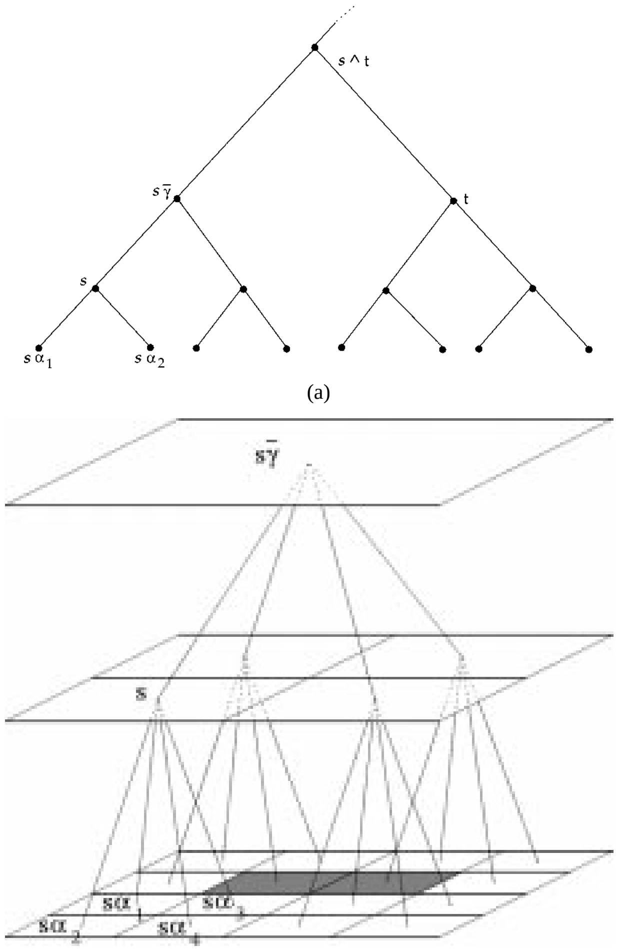
Finally, whether or not the phenomenon, the data, or the objectives are naturally described at multiple resolutions, there may still be compelling reasons to consider developing *algorithms* at multiple resolutions. In particular, MR algorithms offer the promise of computational efficiency. This can be seen in a variety of methods for the solution of large systems of equations [e.g., representing discretizations of partial differential equations (PDEs)]. Multigrid methods

[44], [45], [109], [190], [319] represent one class of examples in which coarser (and hence computationally simpler) versions of a problem are used to guide (and thus accelerate) the solution of finer versions, with finer versions used in turn to correct for coarsening or aliasing errors in the coarser versions. Multipole algorithms [115], [256], [280] approximate the effects of distant parts of a random field with coarser aggregate values, providing substantial computational gains for many problems. Similarly, wavelet-based methods [37], [38], [89], [95], [215], [228], [247], [264], [276], [286], [329], [335], [361] provide potentially significant speed-ups for a variety of computationally intensive problems.

### *B. Our Starting Point*

A key characteristic of MR methods or models is that they introduce a one-dimensional (1-D) quantity, namely scale or resolution, that can be exploited to define recursions or dynamics much as time is used for temporal phenomena. The point of departure for this paper, and for our exploitation of recursions in scale, is the investigation of statistical models defined on MR trees. Two examples of such trees are depicted in Fig. 1. The dyadic tree in Fig. 1(a) is a prototypical structure used in MR representations of 1-D signals and processes, i.e., of signals that are functions of a single independent variable. Here, each level in the tree corresponds to a distinct resolution of representation with finer representations at the lower levels of the tree. Similarly, the quadtree in Fig. 1(b) is one example of tree structures used in the MR representation of two-dimensional (2-D) signals, images, or phenomena. While these two figures represent the structures that are most widely used, much of what we describe here does not require such regular tree structure and could, for example, work equally well on trees in which the number of branches descending from each node was different from either two [as in Fig. 1(a)] or four [Fig. 1(b)] and, in fact, might vary from node to node and resolution to resolution.

In the models we describe, each node $s$ has associated with it a random variable or random vector $x(s)$. Roughly speaking, each such variable represents some set of information relevant to the phenomenon or available data at the resolution and location corresponding to that node. However, what these variables actually are and how they are related to the signals, images, phenomena, or data of interest varies considerably from application to application. For example, in some situations, all of the fundamental, physical variables, i.e., both the signals that are observed and the variables that we wish to estimate or about which we wish to reason, reside at the finest scale only. The coarser scale variables in such a case might simply represent decompositions of the finest scale variables into coarser scale components, e.g., as in the use of wavelet decompositions or Laplacian pyramid [6], [47] representations of images. In other problems, some of these coarser scale variables may be measured directly, as occurs in problems in which we wish to fuse data sets collected at differing resolutions. More generally, the coarser scale variables may or may not be directly observed and may or may not be deterministic functions of the finest scale variables, and their inclusion in the representation may serve purposes such as exposing the statistical structure of the phenomenon under study and/or capturing more global quantities whose estimation is desired. For example, in analogy with stochastic realization theory [8], [9], [214] and the concept of state for dynamic systems, such variables may simply play the role of capturing the intrinsic memory in the signals that are observed or of primary interest. The models we describe also have close ties to hidden Markov models (HMMs) [80], [222], [261], [265], [272], [281], [302], in which the hidden variables may represent higher level descriptors which we wish to estimate, as in speech analysis, image segmentation, and higher level vision problems [42], [53], [59], [175], [179], [180], [183], [199], [283], [323].



(a)

(b)

**Fig. 1.** Examples of MR trees, organized into resolutions. (a) A dyadic tree, typically used for the MR representation of 1-D signals, including notation for nodes on the tree that is used in this and subsequent sections of the paper. (b) A quadtree, frequently used for MR representations of 2-D imagery and random fields. Here, we have used a pictorial representation that emphasizes that each node on the tree represents a "pixel" or spatial region of spatial resolution and at a spatial location corresponding to that node. The shading of the two fine-scale pixels in this figure is associated with a discussion in Section VI-B1.

Whatever the nature of the variables defined on such a tree, there is one critical property that they must satisfy, namely, that collectively they define a Markov process on the tree, a concept we discuss in more detail in subsequent sections. As we will see, MR processes possessing such a Markov property make contact with standard Markov processes in time, with Markov random fields (MRFs) and with the large class of Bayes' nets, belief networks, and graphical models [35], [36], [89], [108], [123], [128], [143], [168]–[170], [197], [204], [236], [267], [294], [295], [302], [337], [339], [357]. It is the exploitation of this Markovian property that leads to the efficient algorithms that we describe.

### C. Getting Oriented

A fair question to ask is: for whom is this paper written? A reply that is only partially frivolous is: for the author. The reason is not self-promotion (although the author pleads guilty to frequently resorting to notation and examples from work with which he is most familiar) but rather an ambitious set of personal goals for this paper. In particular, the field of MR analysis is sufficiently involved and interconnected (forming something much more complex than the singly-connected graphical structures in Fig. 1) and makes contact with so many other disciplines that the writing of this paper has provided an opportunity for the author to sort some of this for himself from a particular point of reference (namely MR models on trees). The result is this paper, which is intended to reach several overlapping but distinct audiences: scientists and engineers interested in applying these methods for problems of complex data analysis; researchers in signal and image processing who are interested in understanding the current state of this active area of research as well as its relationship to others; and researchers in other fields who may find the connections to their specialties of intellectual interest.

To meet this rather ambitious objective, our presentation makes several detours along the way in order to touch on topics ranging from graphical models to stochastic realization theory to solution methods for large systems of linear equations. On several occasions, we also step back and provide additional "navigation tools" for the reader, in particular by explaining how the methods we describe relate to other MR frameworks, most notably wavelets, multigrid/renormalization methods, and MR methods for inverse problems. In addition, throughout the paper, we provide pointers to areas of current research and pointers, both forward and backward, to relationships among the concepts we describe that cannot be accommodated within the severe constraints of linearly ordered text. As a result, the path followed in this paper is not optimized for any of the audiences we have in mind, but we hope that each finds the detours, pointers, and navigation aids interesting, or at least minimally distracting, diversions.

In the next section, we begin by providing an initial look at a sampling of applications that provide context, motivation, and vehicles for illustrating the methods that we describe in later sections of the paper. In Section III, we then introduce the class of MR models on which we focus, provide a few initial simple examples of processes described by such models, and take a first look at ties to graphical models, MRFs, factoring sparse matrices, and recursive modeling of time series. As is the case in much of the paper, our discussion focuses in most detail (but not exclusively) on linear and often linear/Gaussian models. Our reasons for doing this include: the importance of these models in many applications; the simple and explicit form of many computations that allow certain points to be made more clearly; and the relationships that this setting provides to fields such as linear state space modeling of time series and linear algebra. Of course, not everything that we describe for the linear case extends quite so nicely to more general nonlinear models, and we have attempted to make clear what concepts/algorithms extend directly to more general models and what, if any, other issues arise in such cases.

In Section IV, we describe the structure and illustrate the application of the very efficient inference algorithms that these MR models admit. We also take a detour to examine in a bit more detail why these models do admit such powerful algorithms, making contact again with graphical models and with the solution of large, sparse linear systems of equations. In Section V, we take a step back and examine the question of how the models and methods of Sections III and IV relate to wavelet-based methods and multigrid algorithms and in the process also describe relationships with research in inverse problems and image reconstruction. Our intent in so doing is not to provide reviews or tutorials for these very important and substantial lines of research but rather to make clear where these methods intersect with those on which we focus and where, how, and why they diverge.

One of the principal conclusions from Section IV is that *if* a problem can be modeled within the MR framework of Section III, then a very efficient solution can be constructed. That, of course, begs the question of what *can* be modeled effectively within this MR framework and how such models can be constructed. Examination of that question in Section VI uncovers further connections with a number of topics including state-space realization theory, HMMs, graphical models, wavelets, maximum entropy modeling, and algorithms for constructing sample paths of processes that have found use in both the theory of stochastic processes and in fractal generation.

As with any useful modeling framework, this one has a nontrivial and extensive domain of applicability, and we return on several occasions to the applications introduced in Section II in order to provide insight into the classes of processes that can be effectively modeled with MR models and also to illustrate the power of these methods and how they can be used in practice. In addition, as must also be the case for any truly useful modeling framework, its utility is not universal, and, in Sections IV and VI, we provide insights into some of these limitations. In Section VII, we then take a brief look at one of the characteristics that, not surprisingly, is critical both to the power of these models and to their limitations, and, by making ties to the richer class of graphical models not restricted to trees, we provide a brief glimpse into recent and emerging extensions of our framework that expand its domain of applicability. Section VIII concludes our paper with some perspective on this framework and some prospective thoughts.

## II. A Sampling of Applications

The methods we describe in this paper have been employed in a wide variety of applications, including: low-level computer vision and image processing problems (image denoising [59], [67], [80], [261], [281], deblurring [19], edge detection [292], optical flow estimation [10], [223], surface reconstruction [111], texture classification [225], and image segmentation [42], [58], [199], [212], [324], to name a few); higher level recognition and vision problems [183], [323]; photon-limited imaging [188], [261], [263], [322]; network traffic modeling [279]; oceanographic, atmospheric, and geophysical remote sensing, data assimilation, and data fusion [112], [113], [158], [184], [242], [255], [326]; speech [42], [162], [175], [241], [249], [331]; multisensor fusion for hydrology applications [84], [139], [198]; process control [18], [196], [306], [327]; synthetic aperture radar image analysis and fusion [77], [119], [160], [185], [309]; geographic systems [93], [189]; medical image analysis [290]; models of neural responses in human vision [274]; and mathematical physics [15], [94], [136]. In this section, we introduce several of these applications which serve to provide context, motivation, and illustrations for the development that follows, as well as to indicate the breadth of problems to which these methods can be applied.

### A. Ocean Height Estimation

A first application, described in detail in [112] and [113] is to the problem of mapping variations in sea level based on satellite altimetry measurements (from one or several satellites). Fig. 2 (from [112]) shows an example of a region of the Pacific Ocean and the tracks over which the TOPEX/POSEIDON satellite provides measurements to be used to estimate sea-level variations.[1] The challenges in this as well as in other oceanographic data assimilation problems [227] are several. First, the dimensionality of such mapping problems can be enormous, involving estimates on grids of $10^5$–$10^7$ points. Second, as Fig. 2 illustrates, the data that are collected have an irregular sampling pattern. Third, there are substantial nonstationarities both in sea-level variations and in the fidelity of the measurements derived from the altimetry data. For example, the statistical structure of sea-level variations in regions of strong currents, such as the Gulf Stream or Kuroshio Current in the Pacific, are quite different than they are in other ocean regions. Also, since the quantity to be estimated is actually the variation of sea level relative to the geoid (the equipotential surface of the Earth's gravitational field), the raw satellite data must be adjusted to account for spatial variations in the geoid. Since the geoid is not known with certainty and, in fact, can have significant errors near large features such as the Hawaiian Islands and extended subsurface sea mounts and trenches, the resulting adjusted altimetry measurements have errors with spatially varying uncertainties. Fourth, there is a need to compute not only estimates of sea-level variations but also the statistical quality of these estimates (e.g., error variances), as such statistics are needed to fuse these estimates with other information (e.g., ocean circulation models) and to identify statistically significant anomalies. Finally, oceans display variations over an extremely large range of scales—indeed, the typical wavenumber spectral models for sea-level variations have fractal, $1/f^\gamma$ spectra [130], [354].[2]

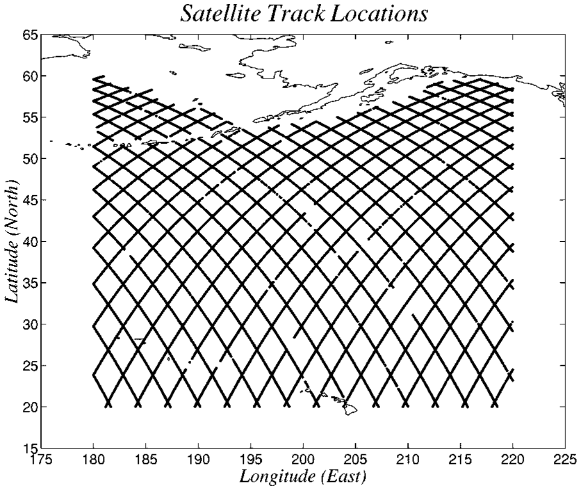


**Fig. 2.** A set of TOPEX/POSEIDON measurement tracks in the north Pacific Ocean. (Reprinted from [112].)

The dimensionality of the sea-level estimation problem and our desire to compute error variances as well as estimates present a daunting computational task, precluding brute force solution methods. Further, because of the nonstationarity of the phenomenon, the varying quality of the data, and the sampling pattern of measurements, efficient methods, such as those based on the fast Fourier transform (FFT), are not applicable. However, as we will see in Section IV, by taking advantage of the fractal character of sea-level variations, a surprisingly simple MR model yields an effective solution.

### B. Surface Reconstruction

A second closely related problem is one that has been widely studied in the field of computer vision, namely that of reconstructing surfaces from regular or irregularly sampled measurements of surface height and/or of the normal to the surface (as in the shape-from-shading problem [34], [53], [152]). One well-known approach to reconstruction problems such as this involves the use of a *variational formulation*. In particular, let $\boldsymbol{I}$ denote the 2-D planar region over which the surface $z(\boldsymbol{r})$, $\boldsymbol{r} = (r_1, r_2) \in \boldsymbol{I}$ is defined, and let

$$\nabla z(\boldsymbol{r}) = (p(\boldsymbol{r}), q(\boldsymbol{r})) = \left(\frac{\partial z(\boldsymbol{r})}{\partial r_1}, \frac{\partial z(\boldsymbol{r})}{\partial r_2}\right) \qquad (1)$$

denote the gradient of the surface. Similarly, let $y(\boldsymbol{r})$ denote measurements of the surface and $g(\boldsymbol{r}) = (u(\boldsymbol{r}), v(\boldsymbol{r}))$ denote measurements of the gradient.[3] Given that these measurements are likely to be noisy and may also be available only at irregular locations (or have spatially varying quality), we

[1] What is actually estimated is sea height, relative to the geoid, with additional corrections to remove effects such as tidal variations and, quite frequently, the overall temporally averaged ocean circulation pattern (see [353] and [354]).

[2] As these references indicate, the power law exponent, $\gamma$, can and generally does vary with frequency and spatially (i.e., the statistics of ocean height variation are only locally stationary in space).

[3] More generally, we might have a measurement of the dot product of this normal with a known vector, a situation that requires only a minor variation in the variational formulation.

take as our estimate of the surface and its gradient the quantities $z(\boldsymbol{r})$ and $(p(\boldsymbol{r}), q(\boldsymbol{r}))$ that minimize the following functional:

$$\int_I \alpha_1(\boldsymbol{r})\,[y(\boldsymbol{r})-z(\boldsymbol{r})]^2\,d\boldsymbol{r} + \int_I \alpha_2(\boldsymbol{r})\|g(\boldsymbol{r})-\nabla z(\boldsymbol{r})\|^2\,d\boldsymbol{r} + \int_I \alpha_3(\boldsymbol{r})\,\|\nabla z(\boldsymbol{r})\|^2\,d\boldsymbol{r} + \int_I \alpha_4(\boldsymbol{r})\left\{\|\nabla p(\boldsymbol{r})\|^2 + \|\nabla q(\boldsymbol{r})\|^2\right\}\,d\boldsymbol{r}. \qquad (2)$$

The nonnegative coefficients $\alpha_1(\boldsymbol{r})$ and $\alpha_2(\boldsymbol{r})$ in the first two terms in (2) allow us to control how closely we wish the reconstruction to follow the measurements,[4] while the third and fourth terms in (2) represent *smoothness penalties* on the reconstructed surface. In particular, the first of these terms is often referred to as a *thin membrane* penalty, as it penalizes nonzero surface gradients, while the last term is referred to as a *thin plate* penalty, as it penalizes curvature or bending of the surface. By adjusting $\alpha_3(\boldsymbol{r})$ and $\alpha_4(\boldsymbol{r})$, we can adjust the relative strengths of these penalties.

mentOne further complication is the integrability constraint, namely, that $(p(\boldsymbol{r}),\ q(\boldsymbol{r}))$ is, in fact, the gradient of a surface. In particular, from (1), it is clear that we must have that

$$\frac{\partial p(\boldsymbol{r})}{\partial r_2} = \frac{\partial q(\boldsymbol{r})}{\partial r_1}. \qquad (3)$$

Minimizing (2) with the constraint (3) or variations of this problem, where, for example, the hard constraint (3) is relaxed and replaced by a quadratic penalty on the difference between the two sides of this equation, is a classic variational problem [152].

Alternatively, as discussed in [111], [223], and [312] (see also Section VI-B1), optimization problems such as this can also be interpreted as estimation problems with “fractal” priors. Computing the optimal estimates for such problems involves solving PDEs [152], a computationally intensive but not overwhelming task in itself. However, the computation of the statistics of the errors in these estimates *is* a daunting task. As we discuss and illustrate in Section VI-B1, (and as is developed in much greater detail in [111]), an alternative is to replace the smoothness penalties in (2), which correspond to a prior model on the surface to be reconstructed, with a different MR prior model which has the same qualitative fractal characteristics but which leads to very efficient algorithms for the computation of estimates *and* error statistics.

### C. Image Denoising

The problem of removing additive noise from images is one that has been the subject of a vast number of studies. Linear methods such as Wiener filters (for spatially stationary models) or those based on Gaussian MRF models have a long history (see, e.g., [13], [163], [343], and [345]). However, such methods, which generally aim to minimize mean-squared estimation error using only second-order statistics of the images to be restored, have serious limitations for many applications in which the image or field to be restored has edges or areas of substantial high-frequency/high contrast behavior. In particular, the generally low-pass nature of linear methods implies that they will reduce noise at the expense of blurring or distorting such important features. For example, Fig. 3(b) depicts a noisy image of a scene [shown without noise in Fig. 3(a)] with a great deal of edge-like, high-frequency behavior. As can be seen in Fig. 3(c) and (d), performing linear Wiener filtering offers a comparatively poor tradeoff in the amount of noise rejection versus the amount of blurring of features.

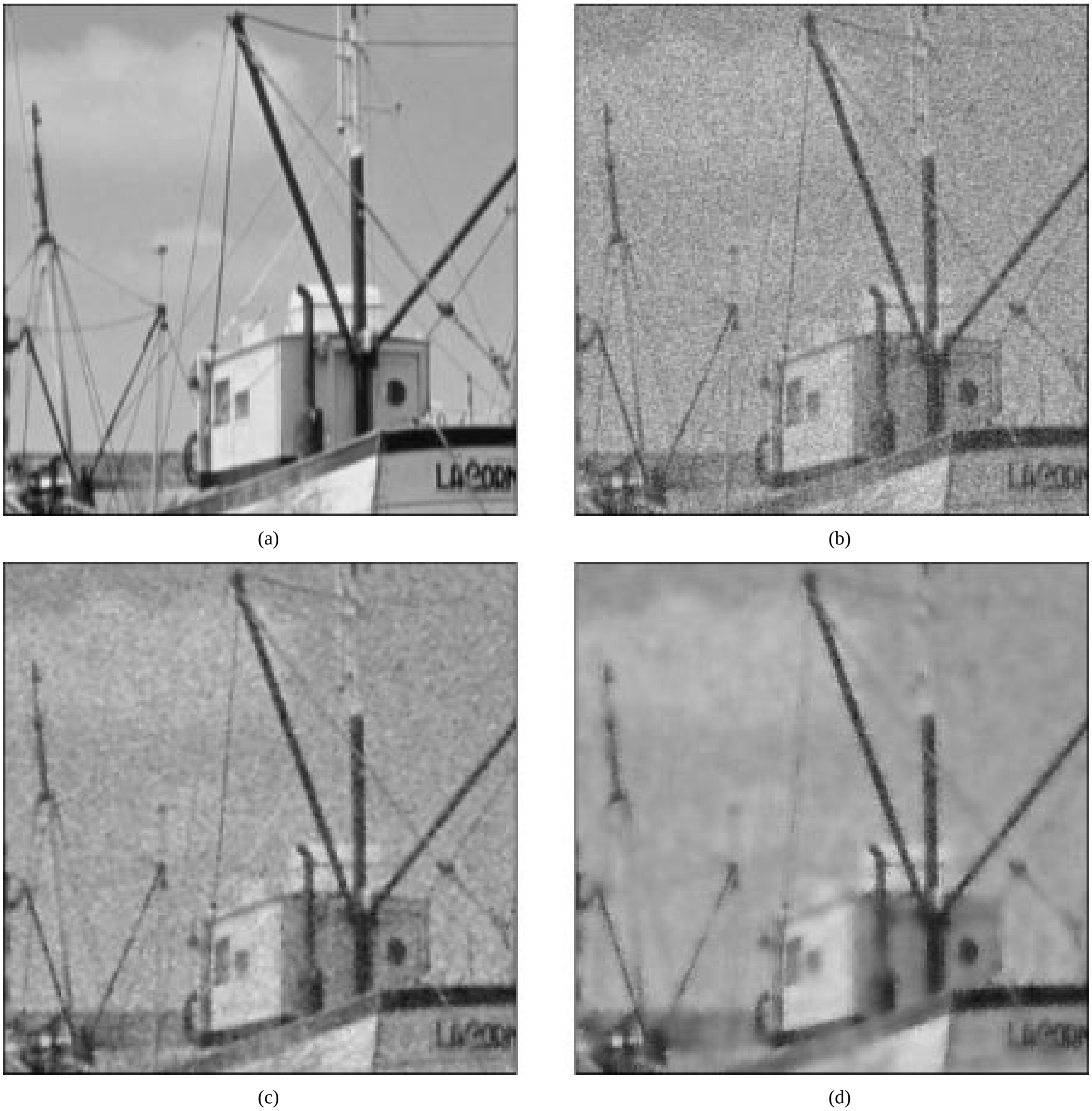
Numerous approaches have been developed to combat such problems—in essence, attempting to remove noise in regions of images away from such features while preserving those features with minimal distortion. Included in the literature are methods based on explicit modeling of edges and other boundary-like features (see, for example, [132] and [234]), approaches that use non-Gaussian models in order to better capture the “heavy tail” nature of imagery (for example, the generalized Gaussian models studied in depth in [41]) and an array of procedures using wavelet transforms (e.g., [2], [57]–[59], [68], [80], [104], [192], [193], [261], [281], [301], [330], and [333]). For this latter set of methods, the general idea is to exploit the localization properties of wavelets to allow much easier and more transparent adaptive processing in order to minimize distortion of important image features while removing noise. As we will see, some of these methods explicitly involve the modeling framework developed in this paper, while many others have close ties to it.

### D. Texture Discrimination

Another problem of importance in computer vision and in other image processing applications is that of texture discrimination. One well-known class of statistical texture models is that based on MRFs [50], [71], [178], [233]. For example, Fig. 4 shows two synthetic MRF textures, one modeling pigskin and one sand. The problem of discriminating textures such as these given noisy measurements is a standard hypothesis testing problem whose solution hinges on the computation of the likelihood ratio for the two textures based on the observed imagery. However, calculating these likelihoods can be a prohibitively complex operation if the data correspond to irregularly spaced samples, if the region over which the data are available has an irregular shape, or if the data have spatially varying statistics (so that FFT methods are not applicable).

As we discuss in Section IV, likelihood calculations for the class of MR models on trees are far simpler and remain tractable even for very high-dimensional image processing problems. Further, as we describe in Section VI, it is possible to develop MR models that capture the statistical variability of textures such as in Fig. 4. These alternate models are *not* identical to the MRF models used to generate these examples, but they are sufficiently close so that they represent equally valid mathematical models for real textures, at

[4] For example, if we do not have measurements of one type or the other over subregions of $\boldsymbol{I}$, we simply set $\alpha_1(\boldsymbol{r})$ or $\alpha_2(\boldsymbol{r})$ to be zero over those subregions.



(a) (b) (c) (d)

**Fig. 3.** (a) A noise-free image. (b) Noisy version of the image. (c) Restored version of this image using optimal Wiener filtering over 3 × 3 image blocks. (d) Restored version using Wiener filtering over 7 × 7 image blocks. (Reprinted from [282].)

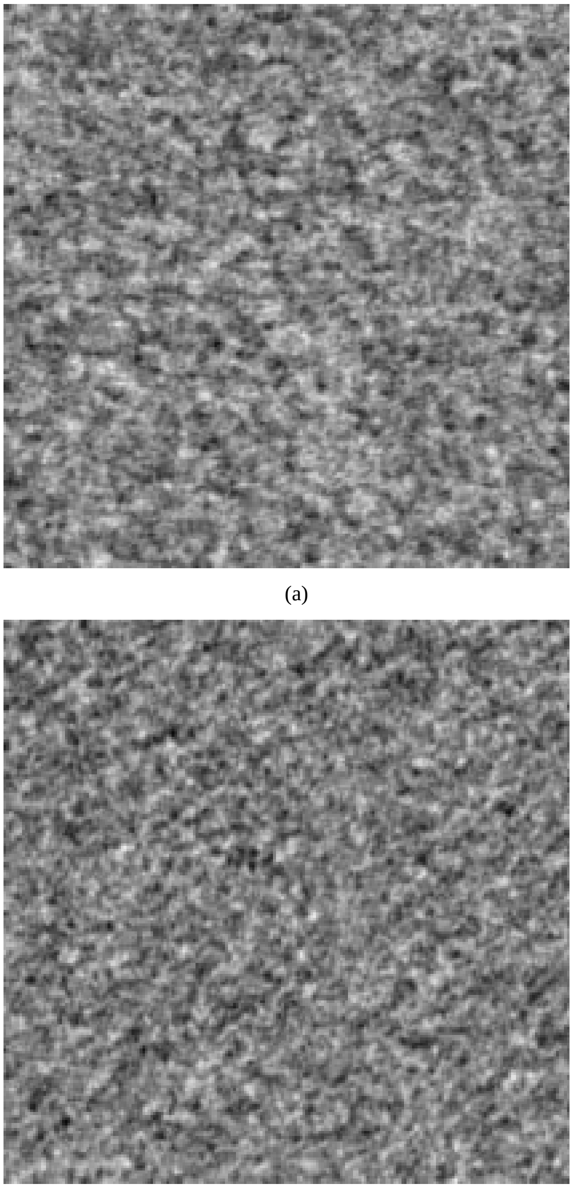
least for the task of discrimination for which they admit very efficient solutions.

### *E. Image Segmentation*

Another image processing and low-level computer vision problem that arises in many applications is that of segmentation. Segmentation of images such as the multispectral image or document page shown in Fig. 5 is a challenging and computationally intensive task, as it involves both accounting for image variability within each class as well as the potentially combinatorially explosive set of candidate segmentations that must be considered. For example, MRF models such as those described in [132] and [234] include discrete hidden label variables whose estimation corresponds to the specification of a segmentation. However, the search for the optimal estimates for such models is computationally demanding, requiring methods such as simulated annealing for their solution or leading to suboptimal methods such as iterated conditional mode (ICM) [36]. These problems have led a variety of authors to consider MR algorithms and models [14], [40], [42], [48], [53], [58], [59], [135], [144], [179], [180]. We describe how some of these methods fall directly into the framework on which we focus and how others relate to it.

### *F. Multisensor Fusion for Groundwater Hydrology*

As we mentioned in Section I, one of the motivations for using MR methods comes from applications in which the available measurements are at multiple resolutions and/or in which the variables to be estimated may also represent aggregate, coarser-scale variables. One application in which this has been examined is in the field of groundwater hydrology [84].



(a)

(b)

**Fig. 4.** Illustration of two textures based on Markov random field models. (a) Pigskin. (b) Sand. (Reprinted from [225].)

The objective in this application is to estimate (and characterize the estimation errors for) the travel time of solutes (e.g., contaminants) traveling in groundwater. This travel time is highly uncertain because of the considerable uncertainty, large dynamic range, and spatial variability in hydraulic conductivity, which controls the spatially varying transport behavior of a groundwater system (see [84]). Specifically, let $f(\boldsymbol{r})$ denote the log-conductivity field[5] as a function of spatial location. Then, the basic governing equation is

$$\nabla \cdot \left[ e^{f(\boldsymbol{r})} \nabla h(\boldsymbol{r}) \right] = \boldsymbol{Q_{rc}}(\boldsymbol{r}) \tag{4}$$

where $h(\boldsymbol{r})$ is the potential field known as *hydraulic head* and $\boldsymbol{Q_{rc}}(\boldsymbol{r})$ is the so-called *recharge rate*, which is often (and, in particular, in [84]) assumed to be known.[6] Further, the local groundwater velocity $u(\boldsymbol{r})$ is a function of conductivity and head, and, in particular, is proportional to the product of conductivity and the gradient of the potential field

$$u(\boldsymbol{r}) \propto e^{f(\boldsymbol{r})} \nabla h(\boldsymbol{r}). \tag{5}$$

The data that are available generally come from a sparse and irregularly spaced set of wells in which both log-conductivity and hydraulic head are measured. From these measurements, we wish to estimate the travel time between two specified points (e.g., a point representing a central contaminant source location and a point on the boundary of a containment region). That travel time is in turn determined by the velocity field in (5).

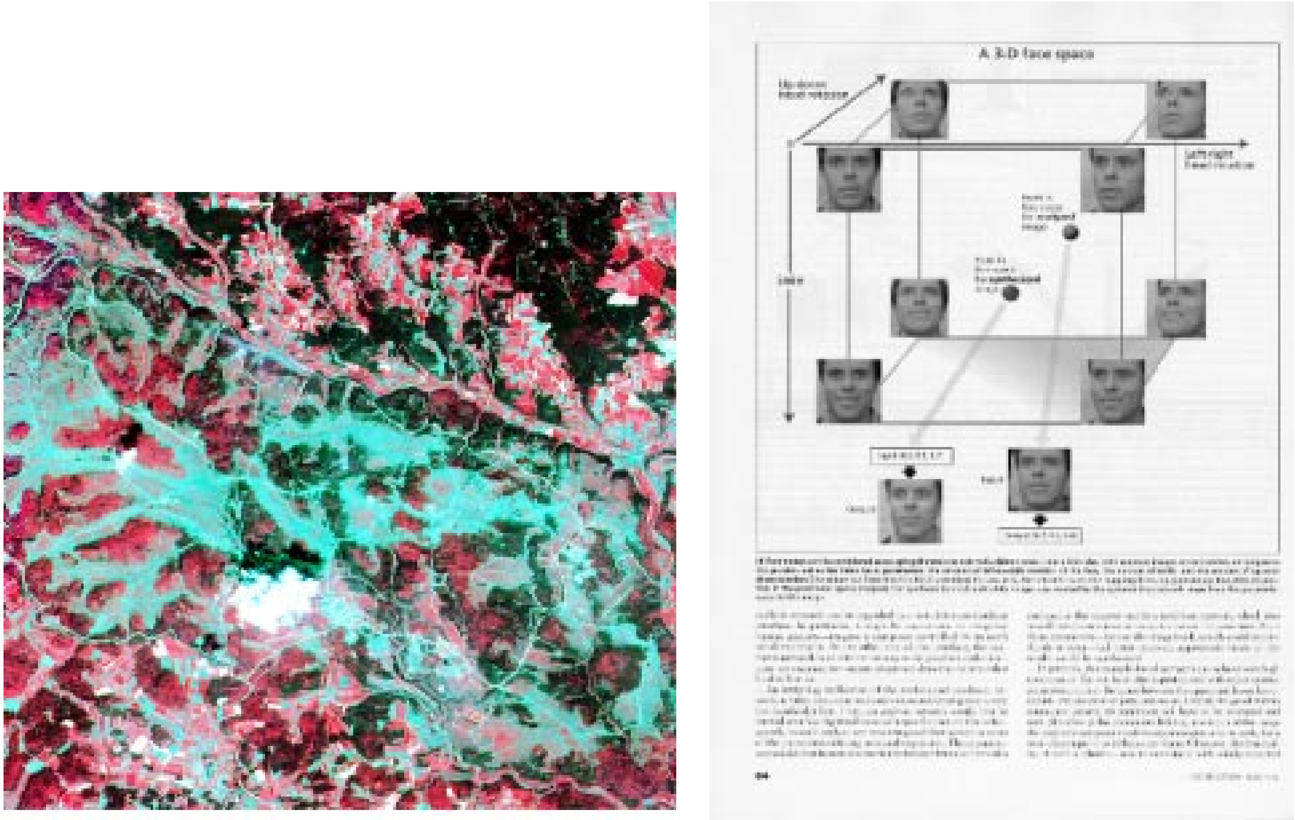
The complexity of this problem should be evident. As discussed in [84], while our measurements of log-conductivity represent point measurements of the random field $f(\boldsymbol{r})$ at the well locations, the measurements of hydraulic head are related to $f(\boldsymbol{r})$ in a much more complicated and nonlocal manner through (4). In Section VI, we will see how the MR framework we describe can be used to capture both the statistical structure of $f(\boldsymbol{r})$ as well as the nonlocal head measurements. We will also see that the MR methodology provides two alternative methods for the fusion of these measurements for the estimation of travel time. In the first of these, we simply model travel time as another nonlocal quantity included explicitly in the MR model and thus estimated directly by the MR estimation methodology described in Section IV. An alternative approach involves the widely used geostatistical concept of *conditional simulation* [171]–[173], in which samples of the entire log-conductivity field are drawn from the distribution for the field conditioned on the available measurements. As we will see, drawing samples from MR models is also extremely efficient—comparable in complexity to generating sample outputs of a time-series model driven by white noise and much more efficient than corresponding methods for many other random field models.

### *G. Image Reconstruction and Inverse Problems*

In the preceding section, we described an application in which the data to be fused included both local measurements of the quantities of interest and nonlocal measurements resulting from indirect probing of the medium or field to be imaged. Data of this latter type are the rule rather than the exception in many applications, including tomographic reconstruction and deblurring or deconvolution problems. In the former, the observed data correspond to projections or sets of line integrals through the field of interest. In the latter, the field to be estimated or reconstructed is blurred by the measurement process. Such image reconstruction or inverse problems present challenges for a variety of reasons. One such reason is purely computational: Performing a reconstruction is a nontrivial task. Another is the ill-posedness of many such problems. For example, operations that involve integration or smoothing (as both tomography and convolution do) can significantly attenuate high-frequency features,

[5] Because of the large dynamic range of conductivity, it is common to use log-conductivity as the fundamental variable.

[6] See [84] for a discussion of the boundary conditions that accompany (4).



(a)

(b)

**Fig. 5.** (a) Remotely sensed multispectral SPOT image (from [42]). (b) Document page (reprinted from [58]).

and, as a result, some operators of this type may not be invertible or their inverses may have very undesirable properties (in particular, amplification of high-frequency noise). As a result, regularization methods, often interpretable as specifying a prior statistical model on the field to be estimated (as in Section II-B), are often employed. In Section V, we will see an example of such a reconstruction algorithm based on an MR model of the type on which we focus in this paper.

## III. MR Models on Trees

### A. *Basic Model Structure*

The general class of models of interest to us are Markov processes defined on trees organized into levels or resolutions, such as in Fig. 1. In Section III-C, we review the concept of Markovianity for more general graphs, but it suffices here to point out that the Markov property for trees is particularly simple. If we condition on the value of the process at any node $s$ on the tree other than a leaf node (e.g., other than one of the nodes at the finest scale in Fig. 1), the sets of values of the process on each of the disconnected components formed by removing node $s$ are mutually independent.

One way in which to specify the complete probabilistic description of such a process is the following generalization of the specification of a temporal Markov process in terms of an initial distribution and its transition probabilities. Specifically, let 0 denote the *root node*, namely the single node at the "top" of the tree, i.e., at the coarsest resolution. For this node, we specify a marginal distribution $p(x(0))$.[7] For each node $s$ on the tree other than 0, let $s\bar{\gamma}$ denote its parent (i.e., the node to which it is connected at the next coarser scale—see Fig. 1), and we then specify the one-step coarse-to-fine transition probability $p(x(s)|x(s\bar{\gamma}))$. The initial distribution together and the full set of these transition probabilities at all nodes other than 0 completely together specify the joint probability distribution for $x(\cdot)$ over the entire MR tree.

One such class of MR models, which we will use to illustrate many of the concepts in this paper, is the class of linear-Gaussian models in which $x(0)$ is a Gaussian random vector and the values of the process at finer scale nodes are specified via coarse-to-fine linear stochastic dynamics as follows:

$$x(s) = A(s)x(s\bar{\gamma}) + w(s) \tag{6}$$

where $A(s)$ is a matrix, specified at each node other than 0 and possibly varying from node-to-node, and where $w(s)$ is a Gaussian white noise process, i.e., a set of mutually independent Gaussian random vectors defined at each node other than 0. Such a model is a simple generalization of the usual linear state space model for temporal processes and systems.[8]

[7] Here, $p(.)$ denotes a probability density function if $x(0)$ is a continuous variable, a discrete probability mass function if $x(0)$ takes on only a discrete set of values, and a combination of the two if $x(0)$ is a hybrid continuous-discrete quantity.

In an analogous manner, one can define other classes of processes such as the generalization of finite-state Markov chains to trees [42], [80], [199], [261]. On numerous occasions, we will find the comparison with temporal Markovianity useful both to interpret results and to identify places in which the extension to trees introduces issues not encountered for time series.

### B. A First Few Examples

To help gain some initial intuition about these models and about their breadth and variability, we present a few initial examples.

*Example 1:* Perhaps the first example of an MR model of the form of (6) is that introduced in [67] in the context of image denoising (see also [355]). Specifically, suppose we are interested in modeling a 2-D random field defined over a square region, where, for simplicity, we assume that the number of pixels along each edge of the square is a power of 2, allowing us to use the simple quadtree structure of Fig. 1(b). In this case, the index $s$ at each node can be thought of as a 3-tuple, $(m(s), i(s), j(s))$, where $m(s)$ denotes the scale of node $s$, and the pair $(i(s), j(s))$ specifies the spatial coordinates of the coarsened spatial region corresponding to $s$ (note that the root node node 0 does not need spatial coordinates; also, we number resolutions consecutively, with $m(0) = 0$, and with increasing scale corresponding to finer resolutions). Let $x(0)$ be a scalar, Gaussian random variable, and define the entire process via the following tree recursion:

$$x(s) = x(s\bar{\gamma}) + w(s). \tag{7}$$

Here, $w(s)$ is a scalar Gaussian white noise process on the tree, and $x(s)$ can intuitively be thought of as a coarse-scale representation of the random field being modeled at the scale and spatial location corresponding node $s$.

Even this very simple model allows us to introduce some of the concepts and issues that arise with MR models. The first concerns the choice of the variance of $w(s)$. One simple choice is a constant variance over the entire tree. Note that, in this case, if we examine the 1-D sequence of values corresponding to the path in the tree from the root node to any leaf node, we see that this sequence is a simple constant-variance Gaussian random walk. If, on the other hand, we choose the variances of $w(s)$ to be constant at each scale but to decrease geometrically from scale to scale (e.g., variances that decrease with scale by a factor of $4^{-m}$), the resulting process has a rudimentary type of self-similarity or fractal character: the variance of the variation at any scale has a power law dependence on scale. We will have more to say about self-similarity and fractal processes later and will also see a somewhat different example next.

A second observation is that (7) corresponds to perhaps the simplest coarse-to-fine interpolation process: interpolation consists simply of copying the value at the coarser node [the first term on the right-hand side of (7)] and then adding independent "detail." This interpretation of multiscale dynamics as coarse-to-fine interpolation combined with the addition of new detail at each resolution clearly rings of concepts common in other areas of multiresolution analysis, most notably wavelets. We will have more to say about relationships to wavelets later. However, as the next point makes clear, the tie to wavelets or other ideas in interpolation requires considerably more thought.

In particular, for any standard MR decomposition of a signal or image, the values of the variables at coarser nodes are simply functionals (i.e., weighted averages or smoothed differences) of the values at finer scales. Indeed, that is certainly the case for wavelet analysis. However, note that for the process defined by (7), it is certainly *not* the case that $x(s\bar{\gamma})$ is the average of its four descendent values (since the four white noise values added to these children are independent). As a consequence, if our primary interest is in the random field at the finest scale, the values at coarser scales in this model represent true hidden variables, as they are not deterministic functions of that finest scale process. As in other contexts, among the reasons for building models with such hidden variables is that they lead to efficient algorithms. In addition, later in this paper we will also discuss the class of so-called *internal* MR models in which the coarser scale variables are *not* hidden.

*Example 2:* In [224], a class of MR models is introduced for 1-D Gauss–Markov processes and for 2-D Markov random fields. The simplest example of this uses Paul Lévy's construction of Brownian motion via midpoint deflection [211].[9] Specifically, suppose that we wish to construct a sample path of a Brownian motion process $B(t)$ over a time interval, say $[T_0, T_1]$, of length $T = T_0 - T_1$. To begin, we first generate samples of the 2-D Gaussian random vector $(B(T_0), B(T_1))^T$. As illustrated in Fig. 6(a), we then draw a straight line between the generated values of our process at these two endpoints. This represents the best estimate of the values of the process at every point between $T_0$ and $T_1$ given the values at the endpoints. Consequently, the error in this estimate at any specific point is independent of the values at the end points. At the midpoint of the interval, $T_0 + T/2$, we then generate an independent, zero-mean random variable $D(T_0 + T/2)$ with variance equal to the error variance in the estimate of $B(T_0 + T/2)$ based on the two endpoint values. If we then "deflect" the straight line at this midpoint by adding this new random variable, we now have *three* samples, $B(T_0)$, $B(T_1)$, and $B(T_0 + T/2)$, that have the desired joint distribution of a sample path of Brownian motion.

The process continues, taking advantage of a critical fact. Because Brownian motion is a Markov process, conditioned on the value at the midpoint, the values of the Brownian motion on the two half-intervals are mutually independent,

[8] As is also true for standard temporal models and signals, while for simplicity we assume that the variables in the model (6) are Gaussian, all of the results and concepts for these models that make use only of second-order properties (means, covariances) hold more broadly as wide-sense concepts. For example, if we only assume that $w(s)$ is uncorrelated from node to node (i.e., is wide-sense white noise on the tree), then the estimation algorithm in Section IV-B represents the best linear estimator.

[9] See [129] for related constructions for the so-called Brownian bridge.

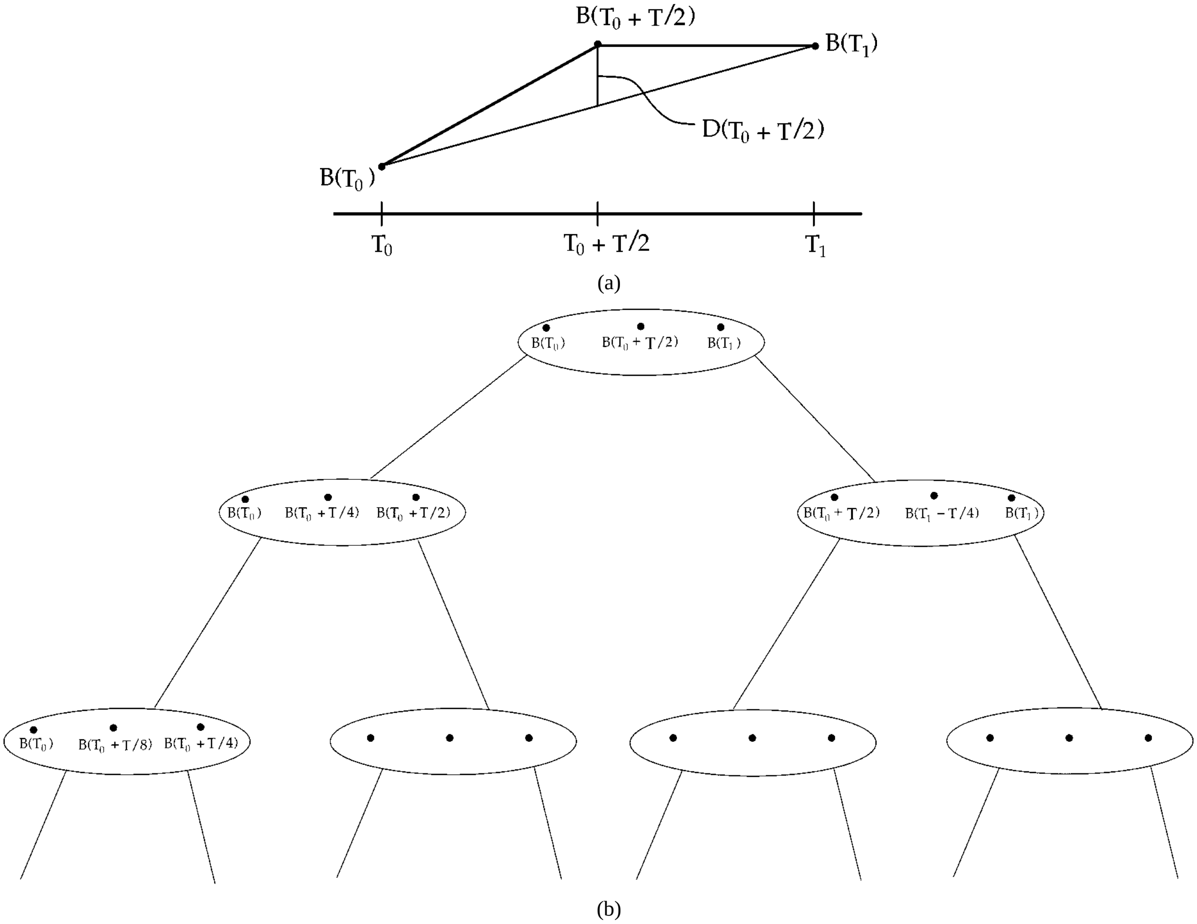

B(T0 + T/2)
B(T1)
D(T0 + T/2)
B(T0)
T0
T0 + T/2
T1
(a)
B(T0) B(T0 + T/2) B(T1)
B(T0) B(T0 + T/4) B(T0 + T/2)
B(T0 + T/2) B(T1 − T/4) B(T1)
B(T0) B(T0 + T/8) B(T0 + T/4)
(b)

**Fig. 6.** (a) Illustrating the midpoint deflection construction of samples of Brownian motion. (b) The MR tree model structure corresponding to the midpoint deflection construction.

and, thus, the subsequent deflection of midpoints of each of these half-intervals can be carried out independently. The result is a procedure for generating denser and denser samples of Brownian motion, which is depicted in Fig. 6(b). As this figure suggests, the procedure we have described corresponds to a linear-Gaussian MR model of the form in (6): here, the three-dimensional (3-D) "state" $x(s)$ at any node $s$ consists of the two endpoint and midpoint values of $B(t)$ over the subinterval identified with node $s$. The coarse-to-fine dynamics are precisely the midpoint deflection scheme we have just described. Each node corresponds to half of the interval associated with its parent node. As a result, two of the components of the 3-D state at each child node are simply copied from the parent node (namely, one of the two endpoints of the parent interval and its midpoint value), and a new midpoint value is generated for the child interval by taking the average of its endpoints and adding an independent zero-mean Gaussian random variable with variance equal to that of the error in the estimate of that new midpoint given the endpoint values.

It is a straightforward calculation to write down the dynamics of (6) for this example (see [224]), but even without doing that explicitly we can make several important observations. The first is that the procedure we have just described works equally well for other Gauss–Markov processes, including those of higher order. The only difference is that the best estimate of a midpoint value given the two endpoint values will in general be a more complex linear function of the endpoint values, depending on the correlation structure of the field. Note also that because of the fact that increments of Brownian motion have variances that scale linearly with the length of the interval over which the increment is taken, the MR model depicted in Fig. 6 has self-similar scaling behavior (i.e., the variances of the midpoint deflections decrease geometrically as we move to finer scales). In addition, as in Example 1, each step in the Brownian motion construction does indeed involve coarse-to-fine interpolation plus the addition of independent detail (to deflect midpoints), although the nature of the interpolation and the detail are very different in this example, as is the fact that the state of the MR process at each node does not represent a spatial average of the process but rather a different type of coarse-scale representation, namely a simple three-point piecewise linear approximation to the Brownian motion sample path, as illustrated in Fig. 6(a). Finally, note that, in contrast to Example 1, the MR model for Brownian motion *is* internal, as the state at each node is a completely deterministic function of its children.

*Example 3:* A class of nonlinear MR models that plays just as important a role in theory and practice as the linear model in (6) is the class of MR Markov chains on trees. In such a model, each of the variables $x(s)$ on the tree takes on one of a finite set of values (where the nature and cardinality of that set may vary from node to node or from scale to scale). As described previously, such a model can be completely specified in terms of the distribution $p(x(0))$ at the root node and the parent–child transition distributions $p(x(s)|x(s\overline{\gamma}))$ for every node $s \neq 0$.

Such models have a long history, extending back to studies in statistical physics [26], dynamic programming [32], artificial intelligence and other investigations of graphical models [7], [89], [128], [169], [267], [294], [295], and signal and image processing [42], [58], [59], [80], [175], [199], [213], [261], [281], [283]. Later in this paper we will illustrate examples of such models for two different purposes. One is a class of image segmentation problems [42], [58], [199], as introduced in Section II-E in which the discrete variable at each node represents a coarse-level label for the image region corresponding to the resolution and location of that node. A standard example used in such problems is a multiscale variant of the Potts model [26], [132] in which each child node takes on the same value of its parent with some probability and is equally likely to take on any value different from its parent, i.e.,

$$p\left(x(s) = i | x(s\overline{\gamma}) = j\right) = \begin{cases} \theta_{m(s)}, & i = j \\ \dfrac{1 - \theta_{m(s)}}{N - 1}, & i \neq j \end{cases} \tag{8}$$

where the label index set for each node is $\{1, \ldots, N\}$ and where we allow the probability $\theta_m$ of the child label equaling the parent to vary with scale $m$ (e.g., as described in [42], one may wish to increase this probability at finer scales).

The second example in which we will see such discrete models is in the context of wavelet-based image denoising problems [59], [80], [261], in which the MR Markov chain represents hidden variables used at each node to control the distribution of a wavelet coefficient at that same node. As we will see, such a model can capture the "cascade" behavior seen in real imagery in which large wavelet coefficients occur in localized patterns across scale corresponding to the locations of abrupt changes, edges, or other high-frequency, high-contrast signal or image features. Including these hidden variables then leads to denoising algorithms that automatically adapt to the presence of edges, alleviating the blurring that occurs if space-invariant linear filtering is performed.

### *C. Some First Ties to Graphical Models, Time Series, and Matrix Factorization*

As we have indicated, MR models on trees are a special class of graphical models [35], [36], [89], [108], [123], [128], [143], [168]–[170], [197], [204], [236], [267], [294], [295], [302], [337], [339], [357]. With an eye toward some of the generalizations we describe later and to lay the foundation for relating our framework to other work, we briefly summarize some of the basic graph-theoretic concepts associated with this larger class of models.

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set $\mathcal{V}$ of vertices and a set $\mathcal{E}$ of edges between pairs of vertices (i.e., $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$). In general, one can distinguish between directed graphs in which an edge $(s, t)$ is directed from node $s$ to node $t$ (so that the edges $(s, t)$ and $(t, s)$ represent different objects) or an undirected graph in which $(s, t)$ and $(t, s)$ do not represent different objects (so that inclusion of one of these in $\mathcal{E}$ is equivalent to inclusion of the other or both). For our purposes, it is sufficient to focus on the latter for the moment and to make a few comments about the former shortly.[10]

Consider an undirected graph $\mathcal{G}$ and a random process $x(s),\ s \in \mathcal{V}$, defined over the index set $\mathcal{V}$. Of particular importance to us is the class of MRFs over the graph $\mathcal{G}$. Specifically, for each node $s \in \mathcal{V}$, let $\mathcal{N}(s)$ denote the set of *neighbors* of $s$, i.e., the set of all nodes other than $s$ itself that are connected to $s$ by an edge. Then $x(s)$ is Markov on $\mathcal{G}$ if for each node $s$

$$p\left(x(s) | \left\{x(t),\, t \in \mathcal{N}(s)\right\}\right) = p\left(x(s) | \left\{x(t),\, t \neq s\right\}\right). \tag{9}$$

That is, conditioned on the values of the process at all its neighbors, $x(s)$ is independent of the remaining values of the process at other nodes. An alternative characterization of Markovianity requires a bit more graph-theoretic terminology. A *path* in the graph $\mathcal{G}$ is a sequence of nodes in $\mathcal{V}$ such that there is an edge corresponding to each successive pair in this sequence. A subset $\mathcal{A}$ of $\mathcal{V}$ *cuts* the graph if the remaining nodes in $\mathcal{V}$ (i.e., $\mathcal{V}/\mathcal{A}$) can be partitioned into two *disconnected subsets* $\mathcal{U}$ and $\mathcal{W}$, i.e., two subsets so that any path in $\mathcal{G}$ from any $u \in \mathcal{U}$ to any $w \in \mathcal{W}$ includes an element of $\mathcal{A}$. Also, we introduce the notation $x_{\mathcal{S}}$ for the set of values $\{x(s)|s \in \mathcal{S}\}$ for any subset $\mathcal{S}$ (although, for $\mathcal{S} = \mathcal{V}$, we will generally denote $x_{\mathcal{V}}$ simply as $x$). Then, $x(s)$ is Markov if, for any subset $\mathcal{A}$ that cuts $\mathcal{V}$ into disconnected subsets $\mathcal{U}$ and $\mathcal{W}$, then

$$p\left(x_{\mathcal{U}},\, x_{\mathcal{W}} | x_{\mathcal{A}}\right) = p\left(x_{\mathcal{U}} | x_{\mathcal{A}}\right) p\left(x_{\mathcal{W}} | x_{\mathcal{A}}\right). \tag{10}$$

That is, conditioned on the values of $x(.)$ on $\mathcal{A}$, the set of values of $x(.)$ on $\mathcal{U}$ is independent of the set of values of $x(.)$ on $\mathcal{W}$. Note that, if $\mathcal{A}$ actually separates $\mathcal{V}$ into $K$ disconnected subsets and if $x(.)$ is Markov, then the sets of values of the process over each of these subsets are mutually independent given the values on $\mathcal{A}$.

The specification of Markov models on general graphs requires some care. In particular, in contrast to temporal Markov processes (or, as we will see, tree models as well), Markov models on graphs are not, in general, specified in terms of the marginal density at a single node and transition probabilities between pairs or small groups of nodes, thanks to the fact that a general graph has *loops or cycles*, i.e., nontrivial paths that begin and end at the same node. Such loops imply that there are constraints (typically complex and numerous) among such marginal and transition probabilities, so that they do not represent a simple parametrization of Markov distributions.

The Hammersley–Clifford theorem [35], however, provides such a parametrization in terms of so-called clique potentials. In particular, a *clique* $\mathcal{C}$ of $\mathcal{V}$ is a fully connected

[10] For simplicity, we assume throughout that $\mathcal{G}$ is *connected*, i.e., that there exist paths of edges that connect every pair of nodes in $\mathcal{V}$.

subset of $\mathcal{V}$ (so that $(s, t) \in \mathcal{E}$ for every pair of distinct nodes $s, t \in \mathcal{C}$). Let $\mathbf{C}$ denote the set of all cliques in $\mathcal{G}$. Then, the Hammersley–Clifford theorem states that $x(.)$ is Markov with respect to $\mathcal{G}$ if its probability density can be written in the following form:[11]

$$p(x) = \frac{1}{Z}\exp\left\{\sum_{\mathcal{C}\in\mathbf{C}}\varphi_{\mathcal{C}}(x_{\mathcal{C}})\right\}. \tag{11}$$

Here, $\varphi_{\mathcal{C}}(x_{\mathcal{C}})$ is a function of the values of $x(.)$ over the clique $\mathcal{C}$ and is known as a *clique potential*. Also, $Z$ is a normalizing constant, often referred to as the *partition function*.

Several points are worth noting. First, if $x(.)$ is a Gaussian process, then we know that the exponent in (11) is a quadratic form in the vector $x$ minus its mean, where the matrix appearing in that quadratic form is the inverse of the covariance matrix $P$ of $x$. An examination of the Hammersley–Clifford theorem for such a process yields the observation that $x(.)$ is Markov with respect to $\mathcal{G}$ if and only if

$$P^{-1}(s, t) = 0 \qquad \forall\,(s, t) \notin \mathcal{E} \tag{12}$$

where, for any matrix $M$ whose blocks are indexed by the nodes in $\mathcal{V}$, $M(s, t)$ denotes its $(s, t)$-block.

In general, the specification of a Markov process according to (11), while providing a natural and unconstrained parametrization, leads to significant computational challenges. For example, recovering the marginal probability distributions of the process at any individual node from this specification has complexity that can grow explosively with the size of the graph, even in the Gaussian case.[12] Similarly, estimating parameters of such models or performing estimation of the process given measurements can also be extremely complex.

The situation, however, is far simpler if $\mathcal{G}$ is acyclic (i.e., loop-free), such as the tree illustrated in Fig. 7(a). One way in which to see why this is the case is to consider the relationship between *directed* graphical models and undirected ones. In a directed graphical model, the quantities that must be specified include the conditional distribution at each node $s$ given the values of all of its parents (where $t$ is a parent of $s$ if there is a directed edge from $t$ to $s$). It is straightforward to convert a directed graphical model into an undirected one (e.g., see [169]), but the construction of a directed graphical model equivalent to an undirected one is generally very complex and, in fact, requires defining new node and edge sets where the nodes consist of entire cliques of nodes of the original undirected graph (see the references on graphical models and also Section VI-A for more insight into this). For a tree, however, the construction of a directed graphical model from an undirected one is straightforward[13] and in fact does not change the nodes of the graph nor the graphical structure (except that edges become directed rather than undirected).

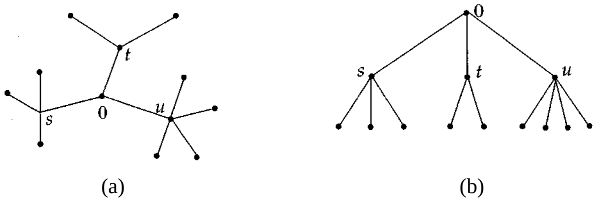


**Fig. 7.** (a) A typical example of a tree. (b) The tree of part (a) redrawn as it appears when the node labeled "0" is taken as the root node.

Specifically, consider an undirected graphical model over a tree and choose any node to designate as the "root" node. Consider then "hanging" the tree from this node—i.e., redraw the graph with the root node at the top level, with its neighbors at the next level, etc. For example, in Fig. 7(a), we have labeled one node, 0, as the root node, with its neighbors denoted as nodes $s$, $t$, and $u$. In Fig. 7(b), we have redrawn the tree as it appears when we hang it from 0. It is then straightforward to see that the overall distribution for this graphical model can be specified, exactly as we did in Section III-A, in terms of the marginal distribution at the root node and the set of parent–child transition distributions. In particular, note that for an acyclic graph any single node other than a leaf node cuts the graph into disconnected components. As a result, for an MRF on the graph in Fig. 7, the processes on each of the subtrees rooted at $s$, $t$, and $u$ are mutually independent when conditioned on the value of $x(0)$. Thus, the overall probability distribution for $x(.)$ can be factored in terms of the individual marginal distribution for $x(0)$ and the conditional distributions for each of the three subtrees rooted at $s$, $t$, and $u$ conditioned on $x(0)$. Continuing this process, each of these subtree conditional distributions can be specified in terms of the individual transition densities for $x(s)$, $x(t)$, and $x(u)$ conditioned on $x(0)$ and the transition densities for each of the leaf nodes conditioned on its parent.

While the preceding discussion is framed in graph-theoretic language, the ideas here become quite familiar to those in the signals and systems community if we describe them in terms of time series and matrix factorizations. Specifically, consider a discrete-time Gauss–Markov process (scalar-valued for simplicity), $x[n]$ defined over the interval $[N_0, N_1]$, and form the vector $x$ by ordering the values of $x[.]$ sequentially. In this case, the graph of interest is simply the set of integers in the interval $[N_0, N_1]$, with edges between consecutive integers. Thanks to (12), we know that the inverse of the covariance $P$ of $x$ is tridiagonal. Such a tridiagonal inverse covariance corresponds to an undirected representation of the statistical structure of this process. However, we also know that such a process has a simple sequential, i.e., a *directed*, representation with the same

[11] Note that the exponential form in (11) implies that $p(x) > 0$ for all $x$ and, in this case, (11) is a necessary and sufficient condition for Markovianity. There are conditions for Markovianity that can be stated if that is not the case; however, that detail is unnecessary for our exposition. We refer the reader to the references at the start of this section for more on this and other aspects of graphical models.

[12] For discrete-state processes, the complexity can be combinatorially explosive, while in the linear-Gaussian case the complexity of the linear-algebraic computations grows polynomially. In either case, the required computations can be prohibitive for Markov processes on arbitrary graphs. We will have more to say about this in subsequent sections and also refer the reader to the references at the start of this section.

[13] Discussions of this can be found in or inferred from many of the graphical model references given at the start of this section. Other discussions of this can be found in [156] and in the discussion of so-called reciprocal processes on trees in [101].

graphical structure connecting each time point to its successor. Specifically, if we take the point $N_0$ as the root node of the (acyclic) graph for this process, the corresponding directed representation of this Markov process is the familiar first-order autoregressive (AR) model

$$x[n] = a[n]x[n-1]+w[n], \qquad n = N_0+1, \ldots, N \quad (13)$$

where $w[N_0 + 1], \ldots w[N_1]$ are a set of independent Gaussian random variables, which are also independent of the initial condition, i.e., the value of $x[N_0]$ at the "root" node. The representation in (13) is precisely in the form of a directed model.

The matrix interpretation of this representation is equally simple. Specifically, define the vector $w = [x[N_0], w[N_0 + 1], \ldots w[N_1]]^T$, which as we have just seen has a diagonal covariance which we denote by $Q$. If we then collect the set of equations in (13), together with the trivial equation $x[N_0] = x[N_0]$, we obtain a vector equation of the form

$$Fx = w \quad (14)$$

where the matrix $F$ is lower bidiagonal, reflecting the fact that each equation in (13) involves a single value of $x[n]$ and its predecessor (while the trivial first equation we have added involves only the initial value $x[N_0]$). Note that a simple calculation using (13) reveals that

$$P^{-1} = F^T Q^{-1} F \quad (15)$$

which corresponds to a very simple *square root-factorization* of the tridiagonal inverse covariance matrix (which also in this case is in the form of a *UDL-factorization*). Several points are of particular note. The first is that the upper and lower triangular factors in (15) are bidiagonal and thus have *no fill*[14] compared to the tridiagonal structure of $P^{-1}$, which is equivalent to the statement that the graph for the corresponding causal recursion in (13) has the same first-order directed graphical structure as that for the original undirected graphical model (corresponding to the tridiagonal inverse covariance). Further, the computation of these factors is very simple, as can be seen from (13): the calculation of $a[n]$ and the variance of $w[n + 1]$ involve only the joint statistics of $x[n]$ and $x[n + 1]$.

In contrast, for a general Gaussian graphical model, calculating a square-root factorization of $P^{-1}$ is computationally involved and results in additional fill in the square root (implying in particular that a directed version of such a model has a more complicated graphical structure). However, for a Gaussian–Markov model on a tree, the procedure we outlined for hanging the tree from a root node and then proceeding recursively down the tree implies that: 1) the calculation of the parameters analogous to those in (13) from one node to its child are as simple as those for a temporal Markov process and 2) there is again no fill. It is precisely these special properties of tree models that lead to efficient algorithms such as those we describe in the next section.

As a final point, it is interesting to note that the procedure we have outlined here to convert an undirected graphical model on a tree into a directed representation of the type we specified in Section III-A allowed us to choose *any* node as the root node and then to define recursions relative to that choice. One of the implications of this for standard temporal Markov processes is well known: we can define a recursive model either forward or backward in time. However, what is perhaps not as widely known or at least as widely used is the fact that we also can define a recursive model that proceeds from the center (or any interior point) out toward the two ends of the interval (see [321]).

## IV. Estimation and Inference Algorithms for MR Models on Trees

As the preceding discussion suggests and as is well known in fields such as graph theory, theoretical computer science, artificial intelligence, and linear algebra, computations on tree structures can be performed very efficiently. In this and subsequent sections, we will see that the implications of this efficiency for statistical signal and image processing and large-scale data assimilation are substantial, and this in turn leads to our asking different questions than those that typically arise in other contexts involving tree-structured computations.

### *A. Computation of Prior Statistics and Simulation of MR Models*

Before discussing optimal estimation and other inference problems for MR models, we examine two related problems, namely the computation of the prior statistics of an MR process $x(.)$ and the generation of a sample path—i.e., the simulation—of such a process. These computations are not only important in their own right but also provide an initial look at the computational challenges in performing statistical calculations and how these challenges are met effectively if we have an MR model on a tree. For each of these problems, we focus primarily on the linear-Gaussian model (6) and comment on the analogous issues that arise for discrete-state models.

As discussed in Section III, the specification of a Gaussian model that is Markov on a general graph corresponds directly to specifying the inverse of the covariance of that process. However, calculating the actual elements of the covariance from such a specification, i.e., calculating the marginal and joint statistics of the values of $x(.)$ at individual or pairs of nodes, is far from a computationally easy task for a general graph. In particular, a naive approach to this would simply be to invert the inverse covariance, a computation that has complexity possibly as large as $O((Nd)^3)$, where $N$ is the number of nodes in the graph and $d$ is the dimension of the "state" $x(s)$ at any node in the graph.[15] Such complexity,

[14] That is, there is no element of $F$ that is nonzero for which the corresponding element of $P^{-1}$ is zero.

[15] In some situations, the dimension of states at different nodes may vary. In this case, the dimension of the vector of all states is simply the sum of the dimensions of these variables (which reduces to $Nd$ if all states have the same dimension $d$).

however, can be prohibitive and, in fact, is for the applications we describe in this paper and for many contexts in which graphical models are used. For example, consider an MR model on a quadtree as in Fig. 1(b). A simple calculation shows that $N$ in this case is roughly 4/3 times the number of pixels at the finest scale. Thus, for a 512 × 512 image, $N$ is on the order of 350 000, while for remote sensing problems, such as that introduced in Section II-A, $N$ can easily be in the millions. For such applications, computations that scale any worse than linearly with $N$, i.e., that have more than constant complexity per pixel for spatial estimation problems, are prohibitive. Moreover, for applications such as in remote sensing, one would never be able to store or even look at the full covariance which contains billions or trillions of distinct elements. However, what we *would* like to be able to do is to compute, in an efficient manner, selected elements of the covariance, e.g., the diagonal blocks corresponding to the covariances of the variables at individual nodes and perhaps a small number of off-diagonal blocks capturing the correlation in space and scale among selected variables.

To be sure, more efficient methods can be devised that exploit the structure of particular graphs, but it is for trees that we obtain especially simple and scalable algorithms [23], [60], [61], [225], [226] for such computations. In particular, consider the linear model (6), where $w(s)$ is a white noise process, with covariance $Q(s)$, independent of the state $x(0)$ at the root node whose covariance we denote by $P_x(0)$. From (6), we then see that the covariance of $x(s)$ satisfies a coarse-to-fine recursion itself as follows:

$$P_x(s) = A(s)P_x(s\bar{\gamma})A^T(s) + Q(s) \tag{16}$$

which is nothing more than the generalization of the usual Lyapunov equation for the evolution of the state covariance of temporal state space systems driven by white noise [12], [174], [182]. Note that this computation directly produces the diagonal blocks of the overall covariance matrix for $x$, and the total complexity of this calculation is $O(Nd^2)$. The quadratic dependence on the dimension of each $x(s)$ reflects the matrix multiplies and additions in (16), while the linear dependence on $N$ reflects the fact that the recursion passes through each node on the tree only once.

Calculation of any individual off-diagonal block of the covariance of $x$ can also be performed in an efficient manner. In particular, for any two nodes $s$ and $t$ on the tree, let $s \wedge t$ denote the closest common ancestor to $s$ and $t$. Then, using the statistical structure of the model, we find that the covariance between $x(s)$ and $x(t)$ is given by

$$P_x(s, t) = \Phi(s, s \wedge t)P_x(s \wedge t)\Phi^T(t, s \wedge t) \tag{17}$$

where for any two nodes $s$ and $u$, in which $u$ is an ancestor of $s$, we have that $\Phi(s, u)$ denotes the state transition matrix from node $u$ to its descendent $s$, which satisfies a recursion in scale analogous to that for the usual state transition matrix for state space models

$$\begin{aligned} \Phi(s, s) &= I \\ \Phi(s, u) &= A(s)\Phi(s\bar{\gamma}, u). \end{aligned} \tag{18}$$

It is interesting to note that (17) is a strict generalization of the formula for temporal models. In the temporal context in which the index set is completely ordered, $s \wedge t$ equals either $s$ or $t$, and, as a result, a nonidentity state transition matrix appears on only one side or the other in (17). However, for MR models on more general trees, both of these will appear in general. Also, note that the calculation of (17) for any particular value of $s$ and $t$ is computationally simple,[16] with complexity bounded by $O((\log N)d^2)$, where the factor of $\log N$ comes from the fact that the path from $s$ or $t$ to $s \wedge t$ can have this length.[17]

While we have described the computation of statistics for the linear-Gaussian case, the same concepts and conclusions hold as well for more general models. For example, consider the case in which $x(.)$ is a finite-state process, taking on any of $d$ values at each node. As we discussed in the preceding section, the computation of marginal distributions at individual nodes or joint distributions at small sets of nodes can be extremely complex for a loopy graph (i.e., a graph with cycles). In particular, in this case, the distribution of the process over the entire graph involves a state set of size $d^N$, i.e., that grows *exponentially* with $N$, and explicit computation of projections of this distribution corresponding to particular marginals or joints has been shown to be NP-Hard for general graphs [76]. However, for an MR process on a tree, $x(.)$ represents a generalization of a Markov chain, and computations of marginals at all nodes can be computed by a coarse-to-fine tree recursion generalizing the usual Chapman–Kolmogorov equation for recursive computation of distributions in Markov chains [266]. Similarly, joints for one node and several of its descendants can be calculated efficiently, and then by averaging over that ancestor node we can obtain joints for any set of nodes [yielding the counterpart to (17)]. In each of these cases, the complexity of computation grows at most linearly with $N$ and exponentially in the number of nodes whose joint distribution is required. As in the linear case, computing or even storing all such joints is prohibitively complex. Typically, one is interested in calculating only a modest number of very low-dimensional joint probabilities, and such computations can indeed be performed efficiently.

MR models also admit very efficient simulation methods. For example, generation of a sample path for the linear MR model (6) is a simple generalization of the straightforward simulation of a linear state space model driven by white noise. We need only to generate a sample of the Gaussian random vector corresponding to the root node $x(0)$ and independent samples of each of the $w(s)$, and then perform the coarse-to-fine computation corresponding to (6). Similarly, for a discrete-state MR model, we draw a sample from the distribution for $x(0)$ and then, in a coarse-to-fine manner, draw samples from the distribution for each node $x(s)$

[16] Note that, essentially with a bit of additional storage and a modest level of additional computation, the calculation of (17) for a particular pair of nodes $s$ and $t$ also yields the values for the covariances for any other pairs of nodes on the path from $s$ to $t$.

[17] This assumes a more or less balanced tree such as in Figs. 1 or 7 in which the diameter of the tree (i.e., the length of the longest direct path between any pair of nodes) is $O(\log N)$.

conditioned on the previously drawn sample value for its parent. In contrast, the simulation of MRFs on graphs with loops can be very complex. For example, for discrete-state processes, iterative procedures such as the Metropolis or Gibbs sampling algorithms [35], [132], [234], [339] must often be employed. Such procedures generally require many revisits of each node of the graph, compared with the single pass through each node for MR models.[18]

As a final comment, it is important to note that the complexity of the MR algorithms we have described here and that we will describe in the rest of this section scale extremely well with problem size as measured by $N$, the number of nodes in our MR tree. However, it is also the case that these algorithms scale polynomially in $d$, which measures the "size" of the variables stored at each node of the tree, e.g., the dimension of the state in a linear model or the cardinality of the state set in a finite-state model. Consequently, the utility of all of these methods depends critically on $d$ being quite small compared to $N$, and it is this observation that in essence provides the "acid test" to see if any particular problem can be successfully addressed using the methods described in this section. The issues of constructing models with manageable state sizes and characterizing processes for which that is or may be possible is the subject of Section VI.

### B. Two-Pass Estimation Algorithms

In this section, we consider the problem of estimating an MR process given noisy measurements of some or all of its values. As we did in the previous section, we begin with a discussion of the linear case, i.e., with an MR model as specified by (6), where, for simplicity of exposition only, we assume that $x(s)$ is zero mean. The problem to be considered is that of estimating this MR process given a set of linear measurements

$$y(s) = C(s)x(s) + v(s). \tag{19}$$

Here, $v(s)$ is a zero-mean white noise process on the tree, independent of $w(s)$ and with covariance $R(s)$, while the matrix $C(s)$ specifies what is measured at each node of the tree. Note that, in principle, this model allows measurements at multiple resolutions and, thus, the estimation algorithm we describe here provides a means for seamlessly fusing such MR data. In addition, even if we only have measurements at the finest scale [i.e., even if $C(s) = 0$ for all nodes other than those at the finest scale], the algorithm we describe has substantial computational advantages.

Let $\hat{x}_s(s)$ denote the optimal estimate[19] of $x(s)$ given all of the data in (19) throughout the tree [i.e., $\{y(s)|s \in \mathcal{V}\}$, where $\mathcal{V}$ denotes the set of all nodes in the MR tree], and let $P_e(s)$ denote the covariance of the error in this estimate. As developed in detail in [60], the computation of these quantities throughout the entire tree can be accomplished using a two-pass algorithm analogous to the two-pass Rauch–Tung–Striebel (RTS) smoother [12], [174], [277] for temporal state space models.[20] That smoother consists of a two-sweep algorithm. The first sweep, forward in time, yields the optimal causal estimate (i.e., the optimal estimate at each time $t$ given all data before and including time $t$), a computation performed using a Kalman filter [12], [174], [182]. At the end of the time interval, $T$, the forward sweep yields the optimal estimate at that final point given all of the data, i.e., this is the optimal smoothed or noncausal estimate at this terminal point as well, since there are no data beyond time $T$. That smoothed estimate then serves as the initial condition for a second sweep backward through the data to compute the optimal smoothed estimate at every point in time. At time $t$, this backward sweep combines the optimal causal estimate at time $t$, computed during the first sweep, with the smoothed estimate just computed at time $t+1$, in order to determine the optimal smoothed estimate at that time $t$, together with the covariance of the error in this estimate.

For more general trees, such as those in Fig. 1, the situation requires a modest amount of additional care and notation. First of all, while the RTS algorithm for time series can be equally well applied from either end of the interval (i.e., we could just as easily start with a Kalman filter that runs in reverse time, starting from time $T$, followed by a sweep forward in time), there is an asymmetry between the fine-to-coarse and coarse-to-fine directions in MR trees, as shown in Fig. 1. As a result, the generalization of the RTS smoother to such a tree must begin with a fine-to-coarse, child-to-parent sweep, starting at the finest nodes, followed by a coarse-to-fine, parent-to-child sweep.[21] The first fine-to-coarse sweep, whose computational flow is illustrated in Fig. 8(a), is a generalization of the temporal Kalman filter. The objective of this sweep is the computation, at each node $s$ of $\hat{x}(s|s)$, the optimal estimate of $x(s)$ based on all of the data in $\mathcal{V}_s$, the subtree rooted at node $s$ (i.e., node $s$ and all of its descendants), together with $P(s|s)$, the covariance of the error in this estimate. As in the temporal Kalman filter, the recursive computation of these estimates involves several steps and intermediate quantities. In particular, suppose that we have computed the best estimate $\hat{x}(s|s-)$ and corresponding error

[18] There are classes of Gaussian graphical models on loopy graphs for which simulations can be performed with efficiency approaching or comparable to the $O(N)$ complexity for tree models. For example, FFT-based methods can be used to simulate spatially stationary random fields with $O(N \log N)$ operations (this follows directly from the fact that the Fourier transform whitens stationary processes [266]). Alternatively, efficient sparse matrix methods (e.g., those used to solve elliptic PDEs) can often be exploited for loopy but sparsely connected graphs, resulting in procedures for sample generation that also have $O(N)$ complexity (see Section IV-D for a related discussion).

[19] Optimality here is defined in the least-squares sense, so that the optimal estimate is simply the conditional mean based on the available data.

[20] In addition, there are other smoothing algorithms for graphical models on trees which have somewhat different computational structures, with the same general complexity. See, for example, [169], [259], [267], [285], [332], [334], [338], and, in particular, [110] for the general characterization of any algorithm that yields the optimal smoothed estimates in a finite number of steps.

[21] Actually, one can equally well define a smoothing algorithm that takes any node as its "root" and which then first sweeps from leaf nodes to the root node, followed by a root-to-leaf sweep. Note that any such procedure has the property that data from each node do in fact find their way into the computations of the smoothed estimate at every other node in the tree.

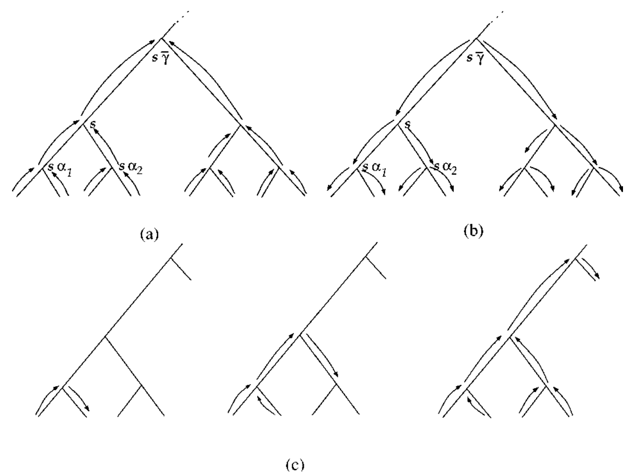


**Fig. 8.** Illustrating the recursive structure of statistical processing algorithms on MR trees. (a) The fine-to-coarse "upward" sweep of the optimal estimation algorithm [see (20)–(32)]. (b) The coarse-to-fine "downward" sweep producing optimal smoothed estimates [see (33)–(37)]. (c) The hybrid recursive structure for the whitening of MR data (see Section VI-C); the upward, fine-to-coarse portions of these computations comprise the Kalman filtering upward sweep shown in (a) to produce partially whitened measurement residuals; the downward portion of these computations complete the whitening based on a particular total ordering of nodes on the tree that is compatible with the partial order implied by the tree itself. (Adapted from [225].)

covariance, $P(s|s-)$, at node $s$, given all of the data in $\mathcal{V}_s$ except for the measurement at node $s$ itself. The computations to produce the updated estimate (and associated error covariance) that incorporates the measurement at node $s$ are identical in form to the analogous equations for the usual Kalman filter.

*Measurement Update:* Given $\hat{x}(s|s-)$, $P(s|s-)$, and $y(s)$, we have

$$\hat{x}\left(s|s\right) = \hat{x}\left(s|s-\right) + K(s)\nu(s) \tag{20}$$

where $\nu(s)$ is the measurement innovations

$$\nu(s) = y(s) - C(s)\hat{x}\left(s|s-\right) \tag{21}$$

which is zero-mean with covariance

$$V(s) = C(s)P\left(s|s-\right)C^T(s) + R(s) \tag{22}$$

and where the gain $K(s)$ in (20) and the updated error covariance $P(s|s)$ are given by

$$K(s) = P\left(s|s-\right)C^T(s)V^{-1}(s) \tag{23}$$

$$P\left(s|s\right) = \left[I - K(s)C(s)\right]P\left(s|s-\right). \tag{24}$$

The second component of the fine-to-coarse recursion is a step that has no counterpart in temporal Kalman filtering, as it involves the fusion of estimates that come from all of the immediate children of node $s$. Specifically, let $\hat{x}(s|s\alpha_i)$ denote the optimal estimate for node $s$ based on all of the data in $\mathcal{V}_{s\alpha_i}$, the subtree rooted at node $s\alpha_i$, and let $P(s|s\alpha_i)$ denote the corresponding error covariance. Fusing all of these estimates produces the estimate (and error covariance) at node $s$ based on all of the data at nodes descendent from $s$ as follows.

*Fusion of Subtree Estimates:* Given $\hat{x}(s|s\alpha_i)$ and $P(s|s\alpha_i)$ for all $i$ (where we let $K_s$ denote the number of descendants of node $s$), we have

$$\hat{x}\left(s|s-\right) = P\left(s|s-\right)\sum_{i=1}^{K_s} P^{-1}\left(s|s\alpha_i\right)\hat{x}\left(s|s\alpha_i\right) \tag{25}$$

$$P^{-1}\left(s|s-\right) = P_x^{-1}(s) + \sum_{i=1}^{K_s}\left[P^{-1}\left(s|s\alpha_i\right) - P_x^{-1}(s)\right] \tag{26}$$

where $P_x(s)$ is the prior covariance at node $s$, computed from (16).

The third step of the recursion involves the computation of the estimates $\hat{x}(s|s\alpha_i)$ (and error covariances) for each child of node $s$. This step is identical in nature to the one-step prediction step in the usual Kalman filter (in which we predict the state at time $t$ based on data through time $t-1$). The only difference in detail is that the "prediction" we must do here is from fine to coarse, while the MR model (6) is specified in a coarse-to-fine manner. As a result, the form of the following step involves a so-called "backward" model analogous to that for temporal models [328].

*Fine-to-Coarse Prediction:* Given $\hat{x}(s\alpha_i|s\alpha_i)$ and $P(s\alpha_i|s\alpha_i)$, we have

$$\hat{x}\left(s|s\alpha_i\right) = F(s\alpha_i)\hat{x}\left(s\alpha_i|s\alpha_i\right) \tag{27}$$

$$P\left(s|s\alpha_i\right) = F(s\alpha_i)P\left(s\alpha_i|s\alpha_i\right)F^T(s\alpha_i) + U(s\alpha_i) \tag{28}$$

where

$$F(s) = P_x(s\bar{\gamma})A^T(s)P_x^{-1}(s) \tag{29}$$

$$U(s) = P_x(s\bar{\gamma}) - P_x(s\bar{\gamma})A^T(s)P_x^{-1}(s)A(s)P_x(s\bar{\gamma}). \tag{30}$$

Finally, this recursion must be initialized, where, in contrast to the temporal Kalman filter, we must provide initial conditions at *all* of the finest scale leaf nodes of the tree. This is done by setting the initial estimate at each leaf node to the prior mean (here assumed to be 0) and the initial covariance to the prior covariance.

*Initialization at the Finest Scale:* For each finest scale leaf node $s$, we have

$$\hat{x}\left(s|s-\right) = 0 \tag{31}$$

$$P\left(s|s-\right) = P_x(s). \tag{32}$$

Note also that as for temporal Kalman filters the gain and covariance matrices can be precomputed, and, in fact, (22)–(24), (26), (28)–(30), and (32) together form the MR tree generalization of the Riccati equation for the error covariance [61].

When the fine-to-coarse sweep reaches the root node, the estimate and covariance computed at that node provide initial conditions for the second coarse-to-fine sweep, exactly as in the temporal RTS algorithm as follows:

$$\hat{x}_s(0) = \hat{x}\left(0|0\right) \tag{33}$$

$$P_e(0) = P\left(0|0\right). \tag{34}$$

As derived in [60], the computations in this second sweep are identical in form to those in the temporal RTS algorithm.

In particular, the computation at node $s$ in the tree involves fusing together the optimal smoothed estimate and covariance just computed at its parent $s\bar{\gamma}$ with the statistics computed at node $s$ during the first Kalman filtering sweep. The only difference in the case of trees is that the node $s\bar{\gamma}$ has several children, so that the following computation is carried out in parallel [as illustrated in Fig. 8(b)] at each of the children of node $s\bar{\gamma}$:

$$\hat{x}_s(s) = \hat{x}\,(s|s) + J(s)\,[\hat{x}_s(s\bar{\gamma}) - \hat{x}\,(s\bar{\gamma}|s)] \qquad (35)$$

$$P_e(s) = P\,(s|s) + J(s)\,[P_e(s\bar{\gamma}) - P\,(s\bar{\gamma}|s)] \qquad (36)$$

where

$$J(s) = P\,(s|s)\,F^T(s)P^{-1}\,(s\bar{\gamma}|s)\,. \qquad (37)$$

Note again that the covariance computations (34), (36), and (27) can be precomputed.

As with the computation of prior statistics described in Section IV-A, the smoothing algorithm just described has very significant computational advantages. In particular, note that the computations at any node on either the upward or downward sweep involve matrix–vector and matrix–matrix multiplies, as well as matrix inversions, where the matrices and vectors involved have dimension $d$ (or perhaps less for those involving the measurement $y(s)$ and its associated matrices). In addition, each node in the tree is visited twice—once in each sweep. Consequently the total complexity of this algorithm is at worst $O(Nd^3)$, which does scale linearly with the number of nodes in the tree. Furthermore, the result of this computation produces both estimates *and* their error covariances.

To understand the significance of this result a bit more deeply, consider an alternate vector form of the estimation equations. Specifically, as before, let $x$ denote the vector of values of $x(s)$ throughout the tree, and let $P_x$ denote its covariance. Similarly, let $y$ and $v$ denote the corresponding vectors of measurements and measurement noises, respectively, so that

$$y = Cx + v \qquad (38)$$

where $v$ has covariance $R$ and where $C$ and $R$ are block-diagonal matrices formed, respectively, from the values of $C(s)$ and $R(s)$ throughout the tree. Then one form for the equations for the optimal estimate is the following:

$$\left(P_x^{-1} + C^T R^{-1} C\right)\hat{x}_s = C^T R^{-1} y \qquad (39)$$

where $\hat{x}_s$ denotes the vector of optimal smoothed estimates, which has corresponding error covariance given by

$$P_e^{-1} = P_x^{-1} + C^T R^{-1} C. \qquad (40)$$

Thanks to the fact that $x(s)$ is Markov on the MR tree, $P_x^{-1}$ has a tree-structured pattern of nonzero elements. Further, since $C$ and $R$ are block-diagonal, the matrix on the left-hand side of (39), namely $P_e^{-1}$, also has the same tree structure. There are two implications of this observation.

The first is that (39) can be solved very efficiently via the tree-structured generalization of Gaussian elimination (the fine-to-coarse Kalman filtering sweep) followed by back-substitution (the RTS smoothing step), yielding the $O(N)$ complexity discussed previously. To be sure, other methods of numerical linear algebra (e.g., conjugate gradient or multipole methods [138], [256], [280]) could be used to solve this equation with this same order complexity. However, what is particularly important about the MR algorithm are both its noniterative nature and especially the fact that it also yields the diagonal blocks of the error variance matrix $P_e$ as part of this same computation. Since these error statistics are extremely important in many applications (see Examples 4, 5, 8, and 9 to follow in this and subsequent sections), this is a major benefit of the use of MR tree models.

The second implication of the tree structure of $P_e^{-1}$, which directly generalizes known results for temporal models [17], [27], [28], [226], is that this implies that the error process $e(s) = x(s) - \hat{x}_s(s)$ is an MR process on the same tree, with parameters (i.e., matrices analogous to $A(s)$ and $Q(s)$ for the original process) that are automatically available as a result of the RTS smoothing computations. Since those computations already yield the covariance of the error at each individual node, the method described in Section IV-A (in particular, equations analogous to (17) and (18) for the smoothing error model) can be used to compute any of the covariances between errors at different nodes. More importantly, we see that the result of the smoothing process produces a model for the remaining errors that has the same form as the original model and which therefore can be used directly for the fusion of any subsequent measurements that become available. Moreover, this error model provides the basis for extremely efficient conditional simulation of MR processes, a feature that is illustrated in Example 9.

It is interesting to note also that there are connections of the preceding development (and, for that matter, much of the discussion in this paper) to problems arising in the field of decentralized control. In such problems, different “nodes” in a network correspond to different “agents” or controllers who observe and can influence the behavior of a dynamical system. If all of the data collected by these agents could be centralized, in principle, one could determine the optimal estimate of the system given all of the measurements as well as determine the optimal coordinated control policy for all of the agents. However, because of constraints that might include computation but in many cases are dominated by other issues (such as communication constraints or the geographic separation among agents), such centralization of information is not possible, and instead one must consider alternative strategies for coordination that, in particular, may produce estimates that are suboptimal. While there are many issues in decentralized control that do not arise in the processing problems considered here,[22] it is worth noting that one case in which relatively simple solutions can be found is that in which the agents have what are referred to in [148] and [149]

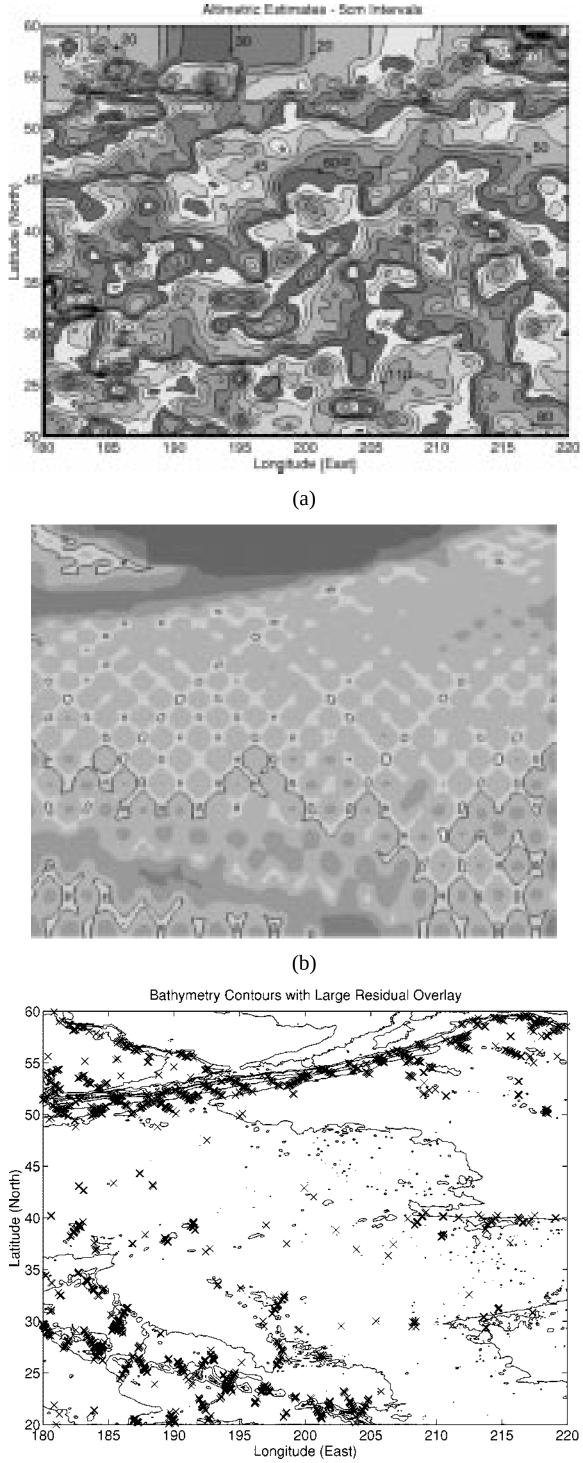
[22] For example, a significant complication arises due to the indirect “communication” that occurs when each agent’s control actions influences the subsequent measurements of other agents.

as *partially nested information patterns*, a construct that is directly related to the singly connected structure of our MR models and of MRFs on acyclic graphs more generally.

*Example 4:* As a first example, consider the optimal estimation of sea-level variations given satellite altimetry measurements such as in Fig. 2, as briefly described in Section II-A and in much more detail in [112] and [113]. As mentioned in Section II-A, sea-level variations have a fractal-like spectrum. Thus, the model (7), with variances of the noise process $w(s)$ that decrease geometrically at finer scales, represents a simple MR model that captures statistical behavior with this type of spectral fall-off. The finest scale in this representation corresponds to pixels of the same size as the resolution of the satellite data. These data, then, are simply modeled as measurements at those finest scale nodes corresponding to locations along the tracks in Fig. 2 (so that $C(s) = 0$ in (19) except for this irregular pattern of finest scale nodes). Fig. 9 shows the results of applying the MR estimation algorithm described in this section to data along the tracks shown in Fig. 2. Fig. 9(a) shows the optimal estimates of sea-level variations, while Fig. 9(b) shows the corresponding error variances over the entire field.

There are several clarifying remarks to make about this example. The first is simply the observation that the dimensionality of this example is nontrivial: we are attempting to estimate roughly 250 000 pixels from approximately 20 000 measurements and at the same time calculate the diagonal elements of the 250 000 × 250 000 error covariance matrix. Using the MR estimation algorithm and this simple model, this is a relatively modest computational task. A second point is that, as described in [112], the measurement noise model used in this case is highly nonstationary, due to the fact that errors in knowledge of the geoid were known to be much larger in regions in which there were significant gradients in the geoid, due, for example, to significant bathymetric features (i.e., variations in the sea floor, such as sea mounts, trenches, etc.). Indeed, one of the advantages of having the error variances computed by the MR estimation algorithm is that we can use these to detect statistically significant anomalies, that is, differences between measurements and estimates that are large compared to what the estimation algorithm would expect based on the computed error variances. Fig. 9(c) shows the locations of the set of the detected anomalies superimposed on a map of ocean bathymetry, showing substantial correlation with significant features of the ocean floor. This suggests, among other things, using these anomalies, together with maps of bathymetry, to provide localized corrections to the geoid.

A third point is that having error variances at multiple scales allows one to identify the optimal scale for reconstruction at different points in the field. In particular, it seems reasonable that one would have greater confidence in higher resolution reconstructions nearer to the regions covered by satellite measurements than in regions farther from any such satellite track. One method for quantifying this is to identify, at each finest scale pixel, its coarser resolution ancestor with the smallest error variance. Thus, at that location, estimation at finer resolutions leads to an increase in uncertainty. An example of this, applied to a different application (namely, the estimation of optical flow in image sequences), is given in [223].



(a)

(b)

(c)

**Fig. 9.** (a) Estimates of ocean height (relative to the geoid) based on a set of TOPEX/POSEIDON measurements along the tracks in Fig. 2. (b) Estimation error variances associated with the estimates in (a). Both of these maps were computed using an MR estimation algorithm. (c) An overlay of ocean bathymetry contours with the locations of statistically anomalous measurement residuals. (Reprinted from [112].)

It is also possible to use the MR algorithm together with more sophisticated models that attempt to capture what is known about the ocean surface more accurately. For example, as discussed in [112], knowledge of spatial inhomogeneities such as the Kuroshio current can be used to adapt the model locally, e.g., by increasing the variances of the $w(s)$ in particular spatial regions and at particular scales. In addition, it is also possible to use higher order models, such as those that have been developed for surface reconstruction problems (see Example 8 in Section VI-B1). Further, using the likelihood function computation methodology described in the next section, we can also tune our models by finding the maximum-likelihood (ML) estimates of the parameters in the model, e.g., the rate of geometric fall-off in noise variances in (7) (see [113]).

Finally, it is important to note that the results shown in Fig. 9 are not obtained quite as simply as the discussion to this point might imply. In particular, as we discuss in Section VI-B1 and as has been pointed out by other authors [67], [188], [261], [281], [322], MR models on trees and especially very simple models such as (7) can produce results that have significant artifacts across major tree boundaries (i.e., at points at the finest scale that are close together spatially but far apart as measured by the path from one to the other along the tree). There are several approaches to dealing with this, including those described in Section VI and also in Section VII. What was used to produce the results in Fig. 9 is the same simple method used by others [72], [270], and [298], namely, averaging the estimation results using several different tree models, each of which is shifted slightly with respect to the others, so that the overall average smoothes out these artifacts. We refer the reader to Section VI-B1 for further discussion of this important issue.[23]

While the preceding discussion is couched in the context of linear-Gaussian models, the same two-sweep structure for optimal estimation holds for *any* Markov model on an MR tree, although, instead of propagating means and covariances in the upward and downward sweeps, we now propagate probability distributions. For example, consider a finite-state MR process as discussed in Example 3, and suppose that we have observations $y(s)$ which we assume are conditionally independent measurements of the MR variables at individual nodes. That is

$$p\left(y(s),\, s \in \mathcal{V} | x(s),\, s \in \mathcal{V}\right) = \prod_{s \in \mathcal{V}} p\left(y(s) | x(s)\right). \tag{41}$$

The discrete case raises a number of issues not encountered in the linear-Gaussian case, one of which is the specific objective of processing. In particular, in the Gaussian case, the algorithm described previously in this section can be viewed as solving *several* problems simultaneously: it provides the overall joint conditional distribution for the entire MR process (implicitly specified as a tree model itself); it also yields the individual marginal conditional distribution for each individual node; *and* it yields estimates that not only are the least-squares optimal estimates but also are the individual node maximum *a posteriori* (MAP) estimates *and* the overall MAP estimate of the entire MR process. For discrete processes, however, computing the node-by-node MAP estimates or computing the overall MAP estimate for the entire process are quite different, and, depending on the application, one or the other of these may be preferable. For example, for image segmentation applications, strong arguments can be made [234] that the computation of individual node estimates rather than an overall MAP estimate is decidedly preferable as it reflects directly the objective of minimizing the number of misclassified pixels. Nevertheless, both of these criteria (as well as a third that we briefly discuss in Example 10) are of considerable interest, and, for graphical models on trees, algorithms for each have been studied and developed by many authors.

In particular, a variety of algorithms have been developed for the computation of the conditional marginals at individual nodes. One class, namely, the so-called "message passing" algorithms, are briefly described in Section IV-D. Another, which can be found explicitly or implicitly in several places (e.g., [7], [169], [199], [205], [267], [294], [295], and [332]) involves a structure exactly as that described previously for the linear-Gaussian case (see, in particular, [199] for a detailed development). As a preliminary step, we first perform a coarse-to-fine Chapman–Kolmogorov computation to compute the prior marginal distribution at each node. The algorithm then proceeds first with a fine-to-coarse step, analogous to the MR Kalman filter in (20)–(32), for the computation of the distribution at each node conditioned on all of the measurements in the subtree rooted at that node. Finally, there is a coarse-to-fine sweep, analogous to the RTS sweep in (33)–(37), which yields the marginals at each node conditioned on data throughout the entire tree. Choosing the mode of each of these marginals yields the so-called mode of the posterior marginals (MPM) estimate. Furthermore, as in the linear-Gaussian case [225], the distribution of the entire MR process conditioned on all of the data has the same tree structure, and the model parameters of this conditional model, i.e., the conditional distribution at the root node and the conditional parent–child transition distribution, are also immediately available as a result of the two-sweep estimation algorithm [332].

The computation of the MAP estimate for the entire process involves somewhat different computations but with very much the same structure and spirit, something that has been emphasized in several investigations [7], [169], [294], [295]. Computing the MAP estimate involves a generalization of the well-known Viterbi algorithm [118], one that can be traced at least back to the study of so-called "nonserial dynamic programming" [32] and to the work of others in artificial intelligence and graphical models [7], [89], [169],

[23] Note that, as discussed in [188], [261], [262], [281], and [322], obtaining shift-invariant algorithms (which are devoid of the artifacts noted in the text) requires, in principle, considering a full set of possible shifts of the MR tree. Such models can be thought of as mixtures of trees [237], [240], i.e., a probabilistic draw is made of one of these tree models. Note that, while straight averaging of the estimates from all of these trees does result in a shift-invariant algorithm, it is technically not the optimal Bayesian estimate. In particular, the optimal Bayesian estimate would require weighting each tree estimate by the conditional probability (based on the observed data) that particular tree was the one drawn from the mixture. While it is certainly possible to do this using the likelihood computation methods described in the next section, to our knowledge this has never been used. Further, the benefit of this additional complexity is, we believe, negligible.

[267], [294], [295]. A description of the algorithm that mirrors very closely the two-pass structure of the estimation algorithms we have described so far (and that also makes clear how this algorithm generalizes standard dynamic programming procedures) can be found in [89] and [199]. A first fine-to-coarse sweep is performed in which two functions are computed at each node. One of these specifies the optimal estimate at that node given the optimal estimate at its parent. The second is the optimal "cost-to-go," namely, the maximum value of the conditional distribution for the entire subtree rooted at that node given both the data in the subtree and the state value at the parent node. This latter quantity is passed back to the parent node for use in the computation of the analogous pair of quantities at that node. When the top of the tree is reached, the optimal estimate at that node is easily computed, initiating a coarse-to-fine recursion in which the estimate of each parent node, together with the function computed on the upward sweep, yield the optimal estimate at each child. As with the MPM algorithm and the computation of likelihoods described in the next section, the key to the existence of this very efficient structure is the fact that the conditional distribution of an MR process on a tree can be recursively factored.

### C. Likelihood Functions

In addition to the optimal estimation algorithms described in the preceding section, very efficient algorithms also exist for the computation of likelihood functions, quantities that are needed in the solution of problems such as hypothesis testing and parameter estimation. Specifically, by exploiting recursive factorizations of MR processes, one can develop an algorithm for computing the likelihood function $p(y(s),\, s \in \mathcal{V})$ that involves a single fine-to-coarse sweep through the data (see, e.g., [169], [267], [333], and [336]). Such an algorithm follows from the following equalities, displayed here for the discrete-state case:

$$p\left(y(t),\, t \in \mathcal{V}_s | x(s\overline{\gamma})\right) = \sum_{x(s)} p\left(y(t),\, t \in \mathcal{V}_s | x(s)\right) p\left(x(s) | x(s\overline{\gamma})\right) \qquad (42)$$

$$p\left(y(t), t \in \mathcal{V}_s | x(s)\right) = p\left(y(s)|x(s)\right) \prod_{i=1}^{K_s} p\left(y(t), t \in \mathcal{V}_{s\alpha_i} | x(s)\right). \qquad (43)$$

Note that, for $s = 0$, the root node $s\overline{\gamma}$ does not exist, and at this node the right-hand side of (42) is simply the overall likelihood function, while the transition probability $p(x(s)|x(s\overline{\gamma}))$ is simply the prior marginal for $x(0)$. Also, in the case of continuous variables, the summation in (42) becomes an integral, which reduces to simple matrix/vector equations in the linear-Gaussian case.

While the fine-to-coarse application of (42) and (43) is quite efficient and most likely the method of choice if the sole objective is the computation of likelihoods, there is an alternative two-pass method for the linear-Gaussian case, which also has total complexity that is linear in $N$, although with a slightly larger proportionality constant. Further, this alternate method computes quantities that can be of value for other problems (such as anomaly detection) and also brings out interesting and important similarities and differences with well-known ideas in state space estimation for temporal processes. In particular, one of the keys to the computation of likelihood functions for temporal state models—and, in fact, one of the key concepts more generally for temporal models of other forms [216], [217]—is the concept of *whitening* the measurements, i.e., of recursively producing predictions of each successive measurement (using a temporal Kalman filter), which, when subtracted from the actual measurement values, yield a sequence of independent random vectors, referred to as the *innovations*, whose covariance depends in a known way on the temporal state model. Since these whitened measurements are informationally equivalent to the original ones, the likelihood function can be written in terms of the joint distribution for the innovations, which is nothing more than the product of the marginals for the innovations at each successive time point.

The estimation algorithm described in the preceding section—and, in particular, the fine-to-coarse MR Kalman filtering sweep—does produce a set of measurement "prediction" errors, namely $\nu(s)$ in (21). However, because of the tree structure, this process is *not* white over the entire tree. In particular, thanks to the structure of the fine-to-coarse sweep [as depicted in Fig. 8(a)], each value of this process involves predicting $y(s)$ based only on the data in the subtree *below* node $s$. For that reason, it is not difficult to see that $\nu(s)$ and $\nu(t)$ are most certainly independent if $s$ and $t$ are on the same path from a leaf node to the root node (i.e., if one of these nodes is a direct descendent of the other), but $\nu(s)$ and $\nu(t)$ are generally not independent otherwise.[24]

As described in [225], to complete the whitening operation (so that the overall likelihood can be written as a product of distributions for the individual innovation values), we define a total ordering on $\mathcal{V}$, extending the partial order of the tree (in essence placing orders on cousins, $k$th cousins $n$ times removed, uncles, etc.) and then complete the whitening operation. By choosing this ordering in a systematic fashion, e.g., as illustrated in Fig. 8(c), we can accomplish this remaining whitening very efficiently, in fact using a somewhat different coarse-to-fine sweep complementing the Kalman filtering fine-to-coarse computation. The result, again, is an algorithm with total computational load that scales linearly with $N$.

*Example 5:* One application of the MR likelihood function computations is to the texture discrimination problem introduced in Section II-B and developed more thoroughly in [225]. As mentioned in Section II, many textures, such as those shown in Fig. 4, can be modeled effectively using MRF models. However, likelihood function computations for MRFs can be highly nontrivial, so that suboptimal methods (such as those in [50]) are often used.

Another approach starts with the simple and obvious statement that such MRF models do not represent "truth" but

[24] As discussed in [225], $\nu(.)$ represents a *martingale increment* process on the partially ordered set defined by the tree.

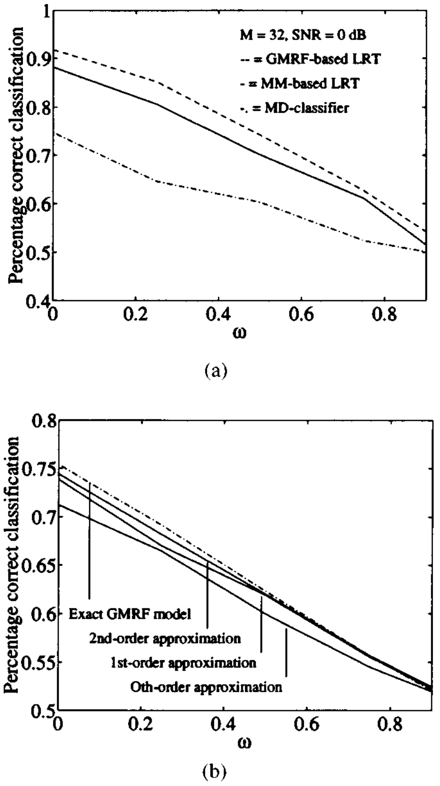


**Fig. 10.** (a) Comparison of the probabilities of correct classification (as a function of a parameter $\omega$). Here, the dashed line represents the optimal performance using the exact Gaussian Markov random field (GMRF) likelihood ratio test (LRT); the solid line corresponds to the performance using a MR model (MM)-based LRT [using what is referred to in [225] as a *zeroth-order model* (corresponding to keeping only a scalar state at each node in the quadtree model)], and the dashed–dotted line is the performance using the suboptimal minimum-distance (MD) classifier from [50]. The results in (a) are for a 32 × 32 image chip at an SNR of 0 dB. (b) Illustrating how performance approaches the optimal achievable as we increase the order of the approximate MR model (these results are for a 16 × 16 image chip at an SNR of 0 dB). (Reprinted from [225]).

rather is itself an idealization of real textures. As a result, it is reasonable to seek alternate models that lead to much simpler likelihood computations resulting in performance essentially as good as what one would be able to achieve using the original MRF models. Fig. 10 illustrates the results of such a texture discrimination system, that is, one based not on MRF models but rather on MR models constructed using reduced-order "cutset" models of the type described subsequently in Section VI-A1. This figure depicts the probability of error in texture discrimination between two models, where one of these models is the MRF model for the sand texture in Fig. 4, while the other model is parameterized by a scalar parameter $\omega$, where $\omega = 0$ corresponds to the model for the pigskin texture in Fig. 4, $\omega = 1$ corresponds to the sand texture, and intermediate values of $\omega$ correspond to MRFs with parameters that are a weighted average of the parameters for the pigskin and sand textures (so that the two textures being discriminated are the most different for $\omega = 0$ and become increasingly similar as $\omega$ increases toward 1, at which value they are identical).

Fig. 10(a) shows that discrimination performance using very simple approximate MR models for these textures is significantly better than the suboptimal method of [50] and nearly as good as the performance achieved if likelihoods using the exact MRF models are employed. Fig. 10(b) shows how the MR-based algorithm's performance varies as the order of the MR model is increased (see Section VI for a detailed discussion of modeling).[25] As this figure indicates, the use of low-order MR models results in performance essentially as good as that achieved using the exact MRF models for these textures.

### D. Further Ties to Graphical Models

To provide some additional perspective on the algorithms described in the preceding sections, we now take a brief glimpse at inference algorithms for models on more general graphs, a topic which has been and remains the subject of numerous investigations (see, e.g., [35], [36], [128], [132], [164], [169], [170], [200], [204], [208], [222], [267], [269], and [339]) and to which we return again in Section VII. To begin, consider the estimation of the state $x(s)$ (assumed to be zero-mean for simplicity) of a Gaussian graphical model on a possibly loopy graph $\mathcal{G}$, given a set of linear measurements, as in (19).[26] As we did in Section IV-B, if we collect all of the state values into a single vector $x$, with covariance $P_x$, and similarly collect all of the measurements into a vector $y$, so that the measurement equation is as in (38), then the optimal estimate $\hat{x}_s$ is once again the solution to (39), and the error covariance matrix is given by (40). Moreover, just as in the analysis in Section IV-B, we see that the block-diagonal structure of $C^T R^{-1} C$ implies that the structure of $P_e^{-1}$ is the same as that of $P_x^{-1}$, i.e., off-diagonal blocks are nonzero only for those blocks corresponding to edges in the graph $\mathcal{G}$. As a result, the estimation error also is Markov with respect to the same graph, a result that can be found in a number of places in the literature (e.g., [208]) and that generalizes the results for time series and MR trees discussed in Section IV-B. The same is also true for nonlinear and discrete-state models, namely that conditioning an MRF with independent measurements at individual nodes yields a conditional distribution that is also Markov over the same graph.

As the results in Section IV-B indicate, if the graph $\mathcal{G}$ is loop-free, there are efficient methods for computing estimates and error covariances in the linear-Gaussian case and marginal conditional distributions in the more general nonlinear case. The reason for this efficiency can be explained in terms of the existence of *elimination orders*, i.e., orders

[25] Here, the "order" of the MR model refers to the dimension $d$ of the state at each node.

[26] In the graphical model literature, there is often no distinction made between measurements and the variables defined at nodes on the graph; we simply have knowledge of some of these variables and wish to perform inference (estimation or likelihood calculation, for example) based on this information. This is a cosmetic rather than substantive difference, as we can easily add nodes to our graph corresponding to each nodal measurement $y(s)$ and a single edge for each such node connecting it to the corresponding original node $s$. We then wish to perform inference based on observation of the variables at these new nodes.

in which variables are eliminated in a first sweep and then added back in the second sweep, for which there is no fill.[27] From a random field/graph-theoretic perspective, the lack of fill for such elimination orders has a simple interpretation: if we *subsample* the random field or graphical model by eliminating a set of variables, this restricted model *remains* Markov with respect to a graph on the remaining nodes with the same neighborhood (and fill) structure as the original graph (see [164], [200] and, in particular, [269] for discussions of this issue for general graphical models).

Similarly, and as we saw in Section IV-C, the computation of likelihoods for loop-free models can also be performed efficiently. While this can be interpreted in terms of the existence of elimination orders without fill, it can also be directly tied to factorizations of the probability distribution over such graphs (e.g., in terms of a root node marginal and parent–child transition densities or as in (42) and (43) in Section IV-B). In particular, the existence of such factorizations implies that the partition function $Z$ in (11) is not a function of the values of the parameters of the model on a loop-free graph (e.g., the matrices $A(s)$ and $Q(s)$ in a linear-Gaussian MR model), a fact that greatly simplifies ML estimation of parameters for such models.

In addition to the MR models on which we focus here, there are other loop-free graphs and processes that have been examined in the literature, primarily in the context of image processing. One such class that received much early attention in the investigation of recursive estimation algorithms for random fields (see, e.g., [163], [343], [345] and the references therein) involves imposing a complete order on a regular 2-D grid of points—typically a “raster scan” order in which the “past” of each point in the lattice consists of all points in previous lines of the 2-D lattice plus the points on the same line that come earlier in the scan order. Another example is the class so-called “Markov mesh” models (for both Gaussian and discrete-state processes), which impose a partial order on pixels in 2-D [1], [78], [88], [98], [213]. However, in many problems, imposing such total or partial orders is clearly artificial. Moreover, often very high-order models of these types are needed to capture accurately the statistics of random fields of interest. As a result, there has been considerable work in developing stochastic models for images that do not impose orderings of pixel locations. For example, MRF models, such as so-called *first-order* MRFs on the nearest-neighbor graph associated with a regular 2-D lattice [132], [234], represent one widely studied class of this type. Consequently, the investigation of inference on loopy graphs, which are the rule rather than the exception in many other fields including artificial intelligence and turbo coding, is also of great interest in image processing.

As we have indicated previously, distributions for Markov models on loopy graphs do not admit elimination orders without fill or simple factorizations in terms of local marginal distributions. Furthermore, the partition function for such a graphical model is generally a complex function of the parameters of the graphical model, e.g., of the clique potentials in (11). Indeed, in the linear-Gaussian case, in which the partition function is proportional to the square root of the determinant of the process covariance, this has been known at least since the work of Whittle [340], [341] (see also [344]) on 2-D random fields, Thus, while the computation of likelihoods and optimal parameter estimates for models on loop-free graphs is computationally tractable and straightforward, the absence of simple factorizations and the dependence of the partition function on model parameters make optimal parameter estimation and hypothesis testing far more challenging computationally. Analogous challenges arise in solving estimation problems, i.e., computing conditional marginal distributions for general nonlinear and discrete-valued models or, in the linear-Gaussian case, solving (39) and determining at least the diagonal elements of the error covariance whose inverse is given in (40). In particular, in general, for such graphical models, successive elimination of variables in any order induces fill, implying both that the set of variables remaining after such an elimination step is Markov with respect to a graph with additional (and frequently *many* additional) edges (see [269]) and that subsequent stages of computation may be increasingly complex.

As a result of these complications, there has been considerable interest in developing either exact or approximate computationally feasible methods for such inference problems. For example, for simple graphs consisting of a single loop, very efficient noniterative algorithms exist that are closely related to methods for solving two-point boundary value problems. One first performs two-sweep computations analogous to RTS computations but ignoring the boundary conditions, e.g., the fact that the two endpoints of the sweep are linked. A subsequent correction step, taking these boundary conditions into account, then produces the correct estimates and covariances or the correct likelihood function.[28] Also, as described in Section VI-A, it is always possible in principle to construct exact noniterative algorithms for inference on loopy graphs, essentially by converting them to problems on loop-free graphs with nodes that correspond to groups of nodes in the original graph. Such methods, which correspond to eliminating groups of variables at once, also can be found in the linear algebra literature (e.g., so-called nested dissection methods [133]). However, such exact methods are not computationally feasible for many graphs, and, as a result, there has been considerable interest in developing approximate and/or iterative algorithms as well.

For the linear estimation problem in (39), a variety of methods exist, especially when the underlying graph, while loopy, is far from fully connected so that the matrix $P_e^{-1}$ on the left-hand side of (39) is sparse. For example, solving such an equation [208] for a first-order MRF essentially corresponds to solving (a discretized version of) an elliptic PDE [208] for which extremely fast algorithms (conjugate gradient, multipole, etc. [138], [256], [280]) exist. However,

[27] In particular, when such an order is used to solve (39) by Gaussian elimination and back-substitution, no new nonzero elements are introduced into the matrix on the left-hand side as variables are eliminated.

[28] See [4], [5], [141], [210], [258], and [260] for a methodology applied to time series, [101] for analogous results for MR trees, and [337] for results for discrete-state graphical models on single loops.

these methods do *not* compute the error covariances (i.e., either the diagonal blocks of $P_e$ or any of the off-diagonal blocks), a difficulty that we briefly discussed in Section IV-A and that also implies that such methods do not allow one to compute likelihood functions.

The graphical model literature also contains a variety of iterative methods that apply to both linear and nonlinear/discrete models and that directly yield approximations to the conditional distributions (i.e., estimates and covariances in the linear-Gaussian case) and in some cases to the computation of likelihoods. Among these are the methods based on generating samples from the conditional distribution for the entire process using techniques such as the Metropolis or Gibbs sampling algorithms [35], [132], [234], [339], which can be used either to estimate marginal distributions at individual nodes (from which approximate MPM estimates can be obtained) or as part of a simulated annealing procedure for the computation of the MAP estimate. Also, a variety of alternative deterministic methods exist. One is the method of iterated conditional modes (ICMs) [36] in which the value at each node is iteratively modified to maximize the conditional distribution for that node given the current iteration's values at its neighbors [a method that reduces to Gauss–Seidel iteration for the solution of (39)]. Others include so-called mean-field methods [359] and the rich class of variational methods [169], [170]. Another method that deserves mention is that developed in [351] using the so-called Bethe tree approximation. In this approximation, the computation of statistics at a specific individual node is approximated by replacing the original graph by a tree rooted at that node, where each path away from the node in the original graph is replaced by a path in the tree. If the path in the original graph contains a loop and thus goes through some node a second time, the corresponding path in the Bethe tree passes through a "distinct" node corresponding to a second "copy" of the actual node in the original graph. Tree algorithms can then be used to compute approximations to the desired statistics.

Several alternate methods for setting boundary conditions on such a tree are described in [351]. If one uses the particular approach of extending the tree without terminating it at some finite point, the concept explored in [351] intersects with a very important class of iterative methods that is widely known in the graphical model literature and that is the subject of considerable current interest, both as an object of analysis in itself and as a point of departure for developing and understanding other emerging methods. This is the class of so-called *Belief Propagation* (BP) [267] algorithms, originally developed in the context of discrete models. While there are a variety of forms for BP, especially for trees, one version that is introduced in [267] and that applies to loopy graphs as well is a "message passing" iterative algorithm, in which each node iteratively passes messages to and incorporates messages from its neighbors. After a single step of the iteration, each node has information from its nearest neighbors, while the next step includes information from nodes that are a distance two from the node in question, etc., providing an expanding "sphere of influence" for each node as the iteration proceeds.

The key to the algorithm is the method for incorporating successive messages from a set of neighbors and then generating the next messages to be passed back to the same set of neighbors. Intuitively, a BP message-passing algorithm incorporates each such message, assuming that the new information it contains is independent of previously incorporated information as well as the information provided by messages from other neighbors. This is the case as long as messages do not propagate around loops of the graph. For a tree, such an algorithm yields the optimal estimate after a number of iterations equal to the diameter of the tree (so that information propagates from each node to every other node). Thus, for a linear or nonlinear MR model on a tree, this local message-passing version of BP represents an alternative to the two-sweep algorithms described in Section IV-B, with a total computational load that is slightly greater than that of the two-sweep algorithm (roughly speaking, each of the $N$ nodes performs computations during each iteration, with the number of iterations corresponding to the diameter of the tree, which, for a balanced tree such as in Fig. 1, is on the order of $\log N$). Furthermore, as elucidated in [332], for a tree, BP also provides all of the parameters needed to construct factorizations of the distribution of the entire graphical process, providing the basis for both efficient likelihood function computation and the construction of equivalent directed models for any choice of root node.

When BP is applied to a graph with loops, the sets of information provided by different neighbors and on successive iterations are *not* independent, since, for example, a message sent by some node will be incorporated into a succession of messages that eventually make their way back to the originating node through a loop in the graph. Such dependencies are not accounted for in the BP inference computations, so that, roughly speaking, information is counted multiple times as it propagates around loops in the graph. As a result, BP may not converge, and if it does converge it will typically not converge to the correct statistical answers. Nevertheless, empirical success of this algorithm in a number of applications, including turbo-decoding [7], [236], [278] and artificial intelligence, as well as known failures of BP in other cases, has led to an intensification of efforts to analyze BP and to develop enhancements and variants of it [285], [332], [334], [337], [338], [357]. For example, as shown in [285] and [338], BP applied to the approximate solution of the linear-Gaussian estimation problem in (39) and to the simultaneous approximate computation of the diagonal error covariance blocks of $P_e$ may or may not converge. If it converges, the estimate $\hat{x}_s$ will, in fact, converge to the exact solution to (39). However, the computed approximations to the error covariances do *not* converge to the correct values (and typically underestimate the size of the estimation errors.)

As this discussion indicates, the complexity of inference on graphs with loops has sparked a considerable body of research and very active lines of inquiry, both to develop new algorithms and also to analyze the performance of existing procedures. In Section VII, we return briefly to this topic to describe several lines of current inquiry that exploit the efficiency of inference on trees in order to develop new algorithms for and insights into inference on more complex graphs.

## V. Some Orientation: Wavelets, Multigrid, and Inverse Problems

As we pointed out in Section I, MR methods span a very broad array of concepts and approaches, and in this section we examine several other components of this large field. Our objective in doing this is to help the reader both navigate through this larger domain and understand how these other lines of inquiry relate to the MR models and algorithms on which we focus.

### A. Wavelets

The use of wavelets [86], [228], [329] to analyze stochastic processes is, perhaps, the most familiar concept that comes to mind when the idea of MR analysis of stochastic processes is raised. Much of the reason for this stems either from analyses that demonstrate that wavelet transforms provide substantial decorrelation of important classes of processes such as fBm [69], [83], [100], [102], [114], [117], [137], [146], [154], [176], [191], [235], [273], [293], [320], [346], [360], or from constructions of processes using wavelet synthesis [62], [80], [83], [100], [114], [151], [275], [346]–[350], [358]. In this section we take a brief look at some of the relationships between wavelets and MR models on trees, a subject we have divided into two components. The first of these describes a set of important examples that are explicitly in the form of MR models on trees with variables at nodes corresponding to individual detail or scaling coefficients of a wavelet decomposition of a signal or image. The second focuses on the interpretation of wavelet synthesis as a coarse-to-fine, scale-dynamic system, a viewpoint that has also received significant attention in the literature but whose explicit connection to tree models (a topic we defer to Section VI-B) is not as obvious.

*1) Elementary Wavelet Models on Trees, HMMS, and Wavelet Cascades:* The fact that wavelet coefficients of many stochastic processes are nearly decorrelated leads directly to a first, elementary method for MR modeling in which we simply *assume* that these coefficients are completely decorrelated. Such an approach, for example, was proposed and developed in [346], [347], and [350] for the modeling of fractal Gaussian processes. Such models can be trivially identified with simple MR models on trees, in which individual detail coefficients serve as the variables that populate nodes at different levels of the tree. For example, for a 1-D signal and the dyadic tree of Fig. 1(a), each node $s$ corresponds to a particular scale $m$ and shift $n$, and the coefficient placed at that node would be the detail coefficient corresponding to that scale and shift. Identifying that variable with the state $x(s)$ of a linear MR model leads to a model as in (6) with $A(s) = 0$, capturing the fact that in this very simple model all coefficients are independent and Gaussian. In the 2-D case, using the quadtree in Fig. 1(b), each node corresponds again to a scale and a 2-D shift, and the variables resident at each node are the three detail coefficients associated with that scale and shift.[29] Once again, taking $A(s) = 0$ implies that the coefficients are independent from node to node. If we also take the covariance $Q(s)$ of $w(s)$ to be diagonal (as is often done in the literature), we then have that each of the coefficients at each node is independent of the others.

While the class of MR models described in the preceding paragraph have degenerate coarse-to-fine dynamics, it serves as a useful point of departure for the examination of other wavelet-based model constructs. A first of these involves an issue of great importance in image processing applications in particular, in which wavelet coefficients, while often nearly decorrelated, are most definitely neither Gaussian nor independent. In particular, as discussed by many authors (e.g., [46], [157], [218], [250], [299], and [300]), the distribution of wavelet coefficients tends to be highly kurtotic and have heavy tails—resulting from the fact that many coefficients come from relatively smooth regions of an image and thus are quite small, while others, corresponding to locations of edges, can be very large. Furthermore, as is also discussed in the literature (e.g., [46], [80], [296], [299], [300], and [333]), large wavelet coefficients generally form *cascades* that are localized in space and propagate across scale, reflecting the presence of edges. Indeed, so-called embedded zerotree approaches to image coding [296] take explicit advantage of these properties of wavelet coefficients.

A class of MR models that captures the non-Gaussianity of wavelet coefficients but not their dependence involves a simple modification to the model described previously. In particular, we still use (6) with $A(s) = 0$, but we use a non-Gaussian distribution for $w(s)$. One possibility is a distribution from the class of so-called generalized Gaussian distributions [41], [250] of the form $K \exp\{-|x|^a\}$, with $0 < a < 2$. An alternative is the class of *mixture distributions* [80], [218]. For example, one of the simpler models introduced in [80] consists of modeling wavelet coefficients as independent with distributions consisting of finite weighted Gaussian mixtures. Such a model, while not truly heavy-tailed (since the Gaussian fall-off is still present as $|x| \to \infty$), can capture a substantial range of non-Gaussian behavior. For example, a very simple two-component mixture (consisting of low-variance and high-variance Gaussians) is used in [80]. Note that, in this case, we can think of each node of the tree as having a hidden variable in 1-D or a set of three hidden variables in 2-D, corresponding to the random choices of which mixture component is used for each wavelet coefficient. Alternatively, as discussed in [333] (see also [46], [300], and [308]), one can obtain a rich variety of truly heavy-tailed distributions as so-called "scale mixtures," i.e., by multiplying a unit mean Gaussian random variable by a positive random variable, generated as a nonlinear function of a second independently drawn Gaussian random variable. Both of these Gaussians are, in essence, hidden variables, as it is only the heavy-tailed product that is observed. Optimal estimation for

[29] As is standard in wavelet analysis of images, these three coefficients correspond to: 1) high-frequency detail in both dimensions; 2) low frequency in the horizontal and high frequency in the vertical; and 3) high frequency in the horizontal and low in the vertical.

any of the choices of distribution mentioned in this paragraph corresponds to performing nonlinear operations on individual wavelet coefficients. Among the algorithms that result from such models are so-called wavelet shrinkage algorithms [2], [49], [57], [68], [104], [192], [193], [251], [301], [330].

As described in [80], such independent mixture models result in very simple nonlinear operations on individual wavelet coefficients for optimal estimation. Moreover, both the discrete-state hidden models as well as the continuous ones in [333] open the door to building MR models on trees that not only capture the non-Gaussianity of wavelet coefficients but also the dependencies displayed by these coefficients in real imagery. A number of approaches to capturing such dependency have been developed (see, e.g., several contributions in [251]). One method is described in [155], where wavelet shrinkage is first used to obtain a coarse estimate of denoised wavelet coefficients, and once these denoised coefficients are subtracted from the observed data a linear MR model is used, together with the two-sweep algorithm in Section IV, to estimate smaller-scale fluctuations. Other approaches attempt to capture directly cascade behavior in which the occurrence of a large wavelet coefficient at one scale implies that nearby coefficients at other scales are also likely to be large. The following example illustrates one important method for accomplishing this.

*Example 6:* In a series of papers [59], [80], [261], [281], [282], a methodology is developed for signal and image processing applications based on so called hidden Markov trees. The basic idea behind these models builds on the Gaussian mixture model just described. In particular, the MR model in this case is a finite-state model as described in Example 3. In the basic version of this model, each state $x(s)$ consists of either a single binary random variable (for 1-D signals) or a set of three binary random variables (for 2-D), which correspond to choices of low- or high-variance Gaussians for the corresponding wavelet coefficients. In order to capture cascade effects of large wavelet coefficients, one can choose the parent-to-child transition probabilities so that the choice of “high” at a parent node makes it likely that “high” will be chosen at a child node. The wavelet coefficients are then modeled as being conditionally independent given the value of the corresponding discrete state $x(s)$. Given such a model, the discrete-state version of the two-sweep estimation algorithm described in Section IV can be directly applied in order to perform image denoising. Fig. 11 depicts the result of applying this methodology to the noisy image depicted in Fig. 3(b) using a set of complex-valued wavelets [186] chosen to minimize the problem of nonshift-invariance that arises with the use of decimated wavelet decompositions [282]. Comparing Fig. 11 with the Wiener filtering results in Fig. 3(c) and (d), we see visually that the nonlinear processing inherent to the MR hidden Markov tree model leads to excellent noise rejection without the blurring evident in the linear Wiener filter.

We also refer the reader to [333], in which the cascade and heavy-tailed behavior of image wavelet coefficients are captured through the use of a linear MR Gaussian model as in (6) for the hidden state, where the measurement model involves multiplying a nonlinear function of this hidden state with a white noise sequence to produce a model for the actual wavelet coefficients. These coefficients, then, are uncorrelated but not independent and, in fact, display the same type of cascade behavior seen in real imagery. Furthermore, as mentioned previously, such a model can capture truly heavy-tailed distributions. Estimation for this nonlinear model, however, requires iterative relinearization. Thanks to the linear MR model for the hidden variables, each iteration can be performed efficiently using the two-sweep algorithm described in Section IV.

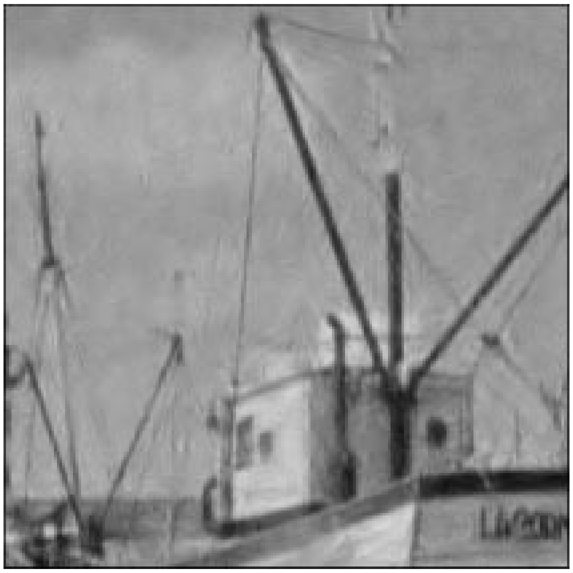


**Fig. 11.** Denoised version of the noisy image shown in Fig. 3(b) using a hidden Markov MR tree model and complex wavelet decomposition. (Reprinted from [282].)

The idea of using multiplicative models to capture cascades in MR decompositions actually has its own rich literature in areas ranging from mathematical physics and the study of random cascades and multifractals [15], [94], [136] to multiscale models for counting processes [188], [261], [263], [322] to multifractal cascade models for communication network traffic [279]. One example of such a model (in which the state variables at each node are scaling rather than wavelet coefficients) is described in Example 7 in Section V-C.

The lack of shift-invariance associated with decimated wavelet representations has led a variety of authors [72], [203], [270], [282], [298], [333] to consider alternatives generally involving the use of overcomplete, undecimated wavelet representations in 1-D and overcomplete “steerable” pyramids in 2-D that allow one to avoid artifacts due to a lack of rotational invariance as well. The use of such overcomplete representations, however, implies that any faithful statistical model must capture the fact that there are constraints among the coefficients in this representation. Since including such constraints greatly complicates any statistical model, it is common in practice to ignore them and thus to use a model that produces estimates of sets of variables that are inconsistent in that they do not correspond to the coefficients in the overcomplete representation of *any* signal. In practice, a variety of methods for projecting such estimates onto a consistent set—or, equivalently, for directly reconstructing a signal given these estimates—have been developed, and we refer the reader to the references cited previously in this paragraph. Note also that this inconsistency is closely related to the concept of internal models first introduced in Example 1 and discussed further later in this section and in Section VI-B.

Finally, there is also a substantial body of work on so-called adaptive representations (e.g., [52], [73], [192], [193], [228], and [229]) using entire families or "dictionaries" of bases, which taken together generally form vastly overcomplete sets. The objective in each of these methods is to select one basis from this collection that leads to the "best" representation—in terms of maximal decorrelation among and/or sparsity in the resulting expansion coefficients for the signal in question. The argument can be made that such an approach produces a tree-structured signal decomposition in which the correlation or dependence among coefficients in the tree is minimized. As a result, using such an optimized representation can improve the accuracy of a resulting MR model for the signal—e.g., the trivial linear model (with $A(s) = 0$) or a more complex model such as one using hidden Markov trees.

*2) Scale-Recursive Models Based on Wavelet Synthesis:* We now turn to an alternate approach to using wavelets in scale-dynamic models. A first method, described by several authors [62], [100], [137], is based on the wavelet synthesis equation. Specifically, let $x_m$ denote the vector of all scaling coefficients at the $m$th scale in an orthogonal wavelet decomposition, and let $w_m$ denote the corresponding vector of wavelet coefficients at the same scale. Then, assuming that we use a compactly supported wavelet, corresponding to a particular pair of wavelet and scaling filters [86], [228], [329], the wavelet synthesis equation can be written as a scale-to-scale recursion

$$x_{m+1} = H_m x_m + G_m w_m \tag{44}$$

where $H_m$ and $G_m$ are matrices corresponding to the conjugate filter pair for the particular wavelet decomposition chosen, and where the first term on the right-hand side corresponds to the coarse-to-fine interpolation of scaling coefficients and the second to the insertion of the additional detail at that next finer scale.[30]

Equation (44) is the starting point for several important observations and investigations. The first involves its direct use in both multiscale modeling and estimation when the wavelet coefficients are modeled as being white noise. In this case, estimation based on such a model given noisy measurements of the wavelet coefficients corresponds to a standard scale-recursive Kalman filter with state at scale $m$ corresponding to the entire vector of scaling coefficients $x_m$. Note that this model allows the direct fusion of nonlocal measurements as long as they correspond to observations of individual wavelet or scaling coefficients [62], [100], [151].

Second, one can do better than this in both modeling and estimation by taking any residual correlation into account. Indeed, several authors have considered methods for doing this [85], [146], [275], [358], and (44) suggests a very simple method of this type, similar to an approach described in Section VI-B. In particular, suppose that the objective is to construct a model as in (44) so that the finest scale process has covariance that closely approximates a given covariance (e.g., of fBm). Since the coefficients $x_m$ and $w_m$ are the scaling and wavelet coefficients of the finest scale process $x_M$, we can directly use the specified, desired statistics of this fine-scale process to determine the statistics of the coefficients at coarser scales. For example, such a computation can be used to determine the variances of the wavelet coefficients $w_m$. If we then ignore any correlations, i.e., if we use these computed variances but otherwise assume that the $w_m$ are white, we obtain an approximate model of the type we have already described. However, it is easy to obtain a more accurate approximation analogous to (44) without any increase in dimensionality. Specifically, suppose that, rather than assuming that the $w_m$ are white, we make the weaker assumption that $x_m$ forms a Markov sequence in scale. In this case, we can capture any correlation between $w_m$ and $x_m$ by writing

$$w_m = L_m x_m + \mu_m \tag{45}$$

where $L_m x_m$ is the best estimate of $w_m$ based on $x_m$ and where $\mu_m$ is uncorrelated with $x_m$ and, in fact, is a white sequence in scale under the assumption that $x_m$ is Markov. Further, both $L_m$ and the covariance of $\mu_m$ are directly computable in terms of the target fine-scale statistics. Substituting (45) into (44), we obtain the following dynamic model:

$$x_{m+1} = (H_m + G_m L_m) x_m + G_m \mu_m \tag{46}$$

which represents a more accurate approximation as it captures some of the residual correlation among wavelet and scaling coefficients. Of course, one can consider higher order approximations (e.g., modeling $x_m$ as a higher order Markov process in scale), although such approximations will require defining a "state" for recursive estimation that consists of several successive resolutions of scaling coefficients. Similarly, by attaching a hidden Markov tree to $\mu_m$, we can, in principle, cpmbine the approach described here with that introduced in Example 6, a possibility that to date has not been examined in the literature.

In addition, other nonlinear variations of this type of structure can be found in [124] and [125] in which the estimation of $w_m$ based on $x_m$ takes on a form that is in some sense both more general and more restrictive than (46). In particular, in the models in these references, the individual detail coefficients comprising $w_m$ are assumed to be conditionally independent when each is conditioned on a specified window of neighboring scaling coefficients (i.e., a cor-

[30] Note that, in the representation of 1-D signals, $x_{m+1}$ has twice the dimension as $x_m$, each corresponding to the full set of variables at the corresponding level of a pyramidal representation. In 2-D, the number of coefficients increases by a factor of four as we move from one scale to the next finer scale. For this reason, the linear operators $H_m$ and $G_m$ are rectangular and also have dimensions that vary with scale.

responding subvector of $x_m$). Both the dependence of each such coefficient on only a local window of scaling coefficients and the resulting conditional independence of the entire vector of coefficients represent what in principle are restrictions compared to the general form of (46), although such restrictions are generally required in order to obtain computationally tractable models (e.g., we will see something similar in Section VI-B). What is more general in [124] and [125] is that the conditional distribution for each component of $w_m$ is modeled as Gaussian, with mean and variance that are *nonlinear* functions of the window of neighboring scaling coefficients (see Section VI-C3 for a further brief discussion).

Note that both of the models (44) and (46) represent consistent models in that the values of the vectors $x_m$ and $w_m$ are, with probability 1, the scaling and wavelet coefficients of the finest scale process $x_M$. However, as discussed in [62], it is also possible to specify models that do not have this consistency but which still exploit wavelet structure. In particular, a model studied in [62] is the following variant of (44):

$$x_{m+1} = H_m x_m + w_m \tag{47}$$

where $w_m$ is white noise. We note here only that the additional DOFs in models such as (47) provide additional flexibility—and, hence, the potential for greater accuracy—in approximating the statistics of any process. However, while examples exist demonstrating the potential of such models (see, e.g., [111], [223], and the discussion in Section VI-B), exploiting this flexibility in a systematic fashion remains an open problem. What is true, however, is that estimation for such models admits efficient solution in exactly the same cases as for (44) and (46).

In particular, note that the dimensions of the state variables in (44) and (45) are substantial, corresponding to *all* of the scaling coefficients at a single scale. As a result, direct implementation of Kalman filtering equations is prohibitively complex for signals or images of even modest size. However, as discussed in [62], if the data that are available are either independent-noise-corrupted measurements of the wavelet or scaling coefficients or can be transformed into this form, the Kalman filter can be implemented in decoupled, diagonalized form using the wavelet transform. One case in which this occurs is if there are dense measurements of the process at one or more scales, corrupted by independent additive noise of variance that can vary from scale to scale but is constant at each scale. In this case, application of the wavelet transform to these data transform them into independent measurements of individual wavelet coefficients.[31] On the other hand, if the available data are sparse or irregularly sampled *or* simply have varying noise variances within any scale, the wavelet transform does not yield such a simplification, and the complexity of the Kalman filter associated with these models becomes prohibitive. One of the reasons for this apparent complexity is that the models (44) and (46) do not correspond to MR models on trees, while (47) does only for the case of the Haar wavelet. Fortunately, however, a more careful, component-by-component look at the structure of the wavelet synthesis equation—and a very different definition of the MR state—does indeed allow us to construct MR models on trees such as in Fig. 1. The construction of such wavelet-based models is described in Section VI-B3.

### *B. Multigrid and Coarse-to-Fine Algorithms*

In this section, we discuss several classes of algorithms using multigrid and coarse-to-fine algorithmic structures. With the exception of the last few approaches we describe, these methods are fundamentally different from those on which we focus in this paper. For that reason, we will be rather brief in our descriptions and not as exhaustive in our review of the literature.

Most of the algorithms that fall within the category on which we focus here correspond to problems that can originally be described at a single, finest resolution—i.e., the data and desired estimates are only available or required at that single resolution. For a variety of reasons, however, direct solution of that single-resolution problem is either too complex to consider directly or is subject to large numbers of local minima, many of which are far from the optimal solution. The general idea of coarse-to-fine methods for such problems is to construct approximate, coarser versions of the problem (typically at multiple resolutions) and to use the solution of the coarser, and hopefully simpler, problems to guide solutions at finer (and eventually the finest) scales. Perhaps the simplest examples of such approaches are to problems in which the spatial phenomenon to be estimated is not modeled as a random field but rather is simply viewed as an unknown, which is represented in an MR fashion to allow coarse-to-fine algorithms for ML estimation. One example of such an approach is given in [63] in the context of an inverse conductivity estimation problem, in which the unknown conductivity field is modeled as piecewise constant at a sequence of resolutions from coarse-to-fine (corresponding to Haar wavelet approximations), and the solution of the problem at one resolution is used as an initial condition for the solution at the next finer resolution. Similarly, coarse-to-fine approaches for the detection and localization of significant anomalies (modeled as unknowns) in a background field (modeled as a random field) [120], [245] have been developed to allow efficient zooming in on features of interest.

There is also extensive literature on the use of MRF models[32] together with either full multigrid computational algorithms or purely coarse-to-fine algorithmic structures. Examples of the former can be found in [70], [109], [319], and [356], where the treatment in [319] represents what to the author's knowledge is the first thorough examination of the application of multigrid methods to image pro-

[31] We refer the reader to [110] for a significant extension of these ideas to a construction using wavelet packets [73] rather than wavelets and in which taking the wavelet packet transform of dense measurements at one or more scales transforms the problem into a set of almost-decoupled MR tree estimation problems, coupled only through a common root node.

[32] In some of these treatments, e.g., [319], MRF models are not explicitly discussed. However, the regularization formulations used, which involve variational penalties on the reconstructed field (e.g., on smoothness), have direct interpretations as MRF models. See Section VI-B1 for more on this.

cessing/computer vision problems. Full multigrid methods, such as those used in [319] and discussed in much more depth in references devoted to the subject such as [44] and [45], involve both coarse-to-fine *and* fine-to-coarse operations in an iterative algorithmic structure. The idea in the coarse-to-fine step is essentially the same as for the methods described in the preceding paragraph: we interpolate a coarser approximation of the estimate to the next finer scale to provide a starting point for the optimization at that scale. Various types of interpolation can be used. For example, there is the general class of wavelet-based interpolation schemes, in which the interpolation from one scale to the next involves simply propagating a scaling coefficient from one scale to the next finer one with the corresponding detail coefficient set to zero (e.g., (44) with $w_m$ set to zero). In other interpolation schemes, the coarser variables might simply represent subsampled versions of the field, with sparser subsamplings at coarser scales. In this case, the coarse-to-fine interpolation might correspond to replication or to something slightly more complicated such as bilinear interpolation.

As discussed in [44] and [45], the fine-to-coarse operations in multigrid reflect the fact that, in most applications of multigrid, the problem that is actually solved at a coarser scale represents an *approximation* to the original problem. For example, in approximating the solution to PDEs (as often arise in continuous-space MRF estimation problems [208]), derivatives at each scale are replaced by differences. Such approximations suffer from aliasing errors which can, in principle, be reduced once we have estimates at the next finer scale. Indeed, as discussed perhaps for the first time in [209], a general interpretation of multigrid algorithms as applied to random field estimation is that, at each resolution, such an algorithm computes the optimal estimate *assuming that there is no finer scale detail in the field*, i.e., that the coarse-to-fine interpolation process is exact so that any finer scale detail coefficients are zero. For example, if we were to assume that an image is constant over a $2 \times 2$ block of pixels, then noisy measurements of those four pixels would simply be averaged in computing the corresponding optimal estimate at that resolution. However, if we subsequently have available finer scale estimates, which in general will vary over these coarser $2 \times 2$ blocks, this fine-scale detail could then be used to correct for the erroneous averaging at the coarser scale, allowing new, coarser scale estimates to be computed. This is exactly what the fine-to-coarse multigrid correction step does.

In a number of other MRF estimation algorithms [40], [126], [135], [140], [144], [145], [195], [257] the full multigrid structure is not used, and only coarse-to-fine operations are performed. In some of these (e.g, see [40], [140], [144], and [145]), the problems that are solved at coarser scales correspond *exactly* to the original problems but with a constrained set of allowed reconstructions (e.g., finding the optimal estimate among all fields that are piecewise constant at each resolution), while in others this is not the case.

While it is certainly possible to view such coarse-to-fine algorithms as purely computationally motivated constructs, [209] makes clear that there are statistical interpretations of at least some of the computations and representations embedded in such algorithms. As a result, a number of authors [164], [144], [200], [269] have looked in more detail at the following question. Suppose we begin with an MRF model at the finest resolution; what is the corresponding statistical structure of a coarsened version of the field (e.g., corresponding to a coarse wavelet approximation or to a subsampled version of the field)? As discussed in [140], [144], and [145] (and as can be inferred from the discussion in [40]), if one begins with a nearest-neighbor MRF in 2-D and then uses Haar-based coarsening, i.e., block-averaging at a set of increasingly coarse scales, then each of these coarser fields is also a nearest-neighbor MRF.[33]

However, as discussed in detail in [164], [200], and [269], in all but a special set of circumstances, coarser scale fields resulting from other coarsening procedures such as subsampling [195], [200] and so-called renormalization group methods [99], [131], [135], [257] do not have such simple exact descriptions—and indeed may correspond to graphical models with fully connected graphs. In such cases, what are generally used at coarser scales are *approximations* to the exact statistics of the coarsened representation that are much more tractable computationally. For example, methods for constructing MRF models that represent optimized approximations for such coarsened fields are described in [195] and [200], while [135] and other renormalization approaches use statistical methods as a means of averaging over finer scale fluctuations in order to construct approximate coarser scale MRF models. Using such approximate models implies that the problems being solved at coarser scales are only approximations to (rather than constrained versions of) the finest scale problem. Of course, this is completely consistent with the philosophy of multigrid, in which coarse scale computations are of no intrinsic interest in themselves but rather serve the purpose of helping to guide the finer scale computations.

As the preceding discussion makes clear, in all of the multigrid/coarse-to-fine image processing and random field estimation algorithms described so far (and in the vast majority of such methods in the literature), coarse-scale representations of the phenomenon of interest are introduced primarily for computational purposes, and any explicit or implied statistical structure is isolated from scale to scale—i.e., there is no single, consistent statistical model across the scale. As a result, these methods are rather different from those on which this paper focuses. However, there are several investigations that both fall into the general category of multigrid or coarse-to-fine procedures and *do* involve random quantities at multiple resolutions that are explicitly linked statistically across scale through a graphical model. One such example using MR models on trees for coarse-to-fine SAR segmentation can be found in [119]. Another, which involves graphs other than trees for image segmentation, is that originally developed in

[33] This is closely related to the fact, mentioned in the preceding paragraph, that these methods produce exact, solutions to the original estimation problem at each resolution under the constraints that there is no finer scale detail present in the random field.

[42] (see also [53], [183], and [323]), which we describe and illustrate in Section VII. In addition, there are several other such modeling frameworks described in the literature, including those in [179] and [180], in which the basic graphical structure superficially resembles the quadtree of Fig. 1(b), except that there are also edges within each resolution—so that each scale by itself has connectivity exactly as with an MRF (e.g., nearest-neighbor edges for first-order models or larger neighborhoods for higher order dependencies at each resolution). Such a graph has complexity[34] that is considerably greater than a single-scale first-order MRF, and methods such as simulated annealing need to be applied in order to obtain solutions. In fact, the authors of [179] and [180] develop multitemperature, MR annealing algorithms and demonstrate that there are potentially some advantages to this approach that result from the usual multigrid/coarse-to-fine philosophy of using coarser grids to guide solutions at finer ones. We also refer the reader to other MR models [74], [125], [127], [212], [213], [289], that involve structures other than trees, together with algorithmic structures that are reminiscent of multigrid and coarse-to-fine procedures.

One final point to make about the methods mentioned in the preceding paragraph concerns the modeling of the measurements. One of the critical properties of the graphical estimation problems and algorithms discussed in the preceding sections is the assumption that each observation consists of a measurement of the state at a single node on the graph corrupted by independent noise. If this is not the case, then the actual graphical structure of the estimation problem must reflect the additional dependencies introduced by the measurements. In some methods, such as that in [42], the assumption of conditional independence is clearly satisfied, as all of the measurements are of individual finest scale pixels. However, in others, such as [74], [179], [180], and [213], the original finest scale measurements are transformed—e.g., simply by replicating the same measurements at different resolutions, by transforming the raw data using wavelet transforms, or by extracting features at multiple resolutions. Furthermore, the algorithms in these papers implicitly assume that these transformed data are conditionally independent, a condition that in some cases (e.g., measurement replication) is clearly not true and in others represents an implicit assumption of whitening resulting from wavelet transformation or feature extraction.

### C. MR Algorithms for Inverse Problems

In this section, we take a look at MR algorithms for inverse problems, i.e., problems in which the measurements are nonlocal functions of the random field to be reconstructed. Inverse problems span a broad array of applications and mathematical formulations, and the literature investigating their properties and developing methods for their solution is equally vast (see, for example, [33]). Many of the problems and methods in this field have no or at most tangential connection with the MR statistical models and processing algorithms with which we are primarily concerned. Consequently, our discussion here is brief, focusing exclusively on aspects of work in this area that intersect with the main themes of this paper.

As a first comment, we note that MR methods for inverse problems have significant overlaps with the topics of both of the preceding sections. For example, the use of multigrid or coarse-to-fine algorithms for inverse problems in order to combat problems of computational complexity and local minima is well documented (with [63], [125], [126], [215], [289], [290], and [356] representing examples). As for the use of wavelets, the motivations include not only the fact that wavelets decorrelate many stochastic processes but also that in many cases they lead to significant sparsification of the nonlocal operator relating the measurements and the underlying random field which we wish to reconstruct. For example, a number of authors [38], [95], [264], [271], [276], [286], [287], [361] have exploited this fact in the specific context of tomographic reconstruction in order to develop deterministic inversion algorithms that are either very efficient or that allow fast high-resolution reconstructions of localized regions. Also, a number of authors [39], [103], [107], [246]–[248], [305], [335], [362] have used both the decorrelation and sparsification properties of orthogonal wavelet decompositions in order to develop statistical reconstruction algorithms for tomography, deblurring, and other inverse problems. For the most part, such methods involve the use of the simplest class of statistical models described in Section V-A, namely, those resulting from assuming that the wavelet coefficients are uncorrelated random variables. If one were then also to assume that the wavelet transform truly diagonalized the measurement operator, the associated measurement model would have the form of (19), corresponding to uncorrelated measurements of individual wavelet coefficients.

Given the discussion in Section V-A, an obvious variation of such methods is one in which we use wavelet shrinkage techniques—e.g., corresponding to modeling the wavelet coefficients as independent but non-Gaussian random variables. Several such methods have been developed (see, for example, [244]), but the predominant use of wavelet shrinkage for inverse problems, introduced and popularized by Donoho [105], uses a variation, known as the wavelet–vaguelette decomposition (WVD). As shown in [105] and in other references in this area (see, e.g., [3], [187], and [206]), WVDs can be designed for important classes of inverse problems (including tomography). Such decompositions correspond to using an orthogonal wavelet decomposition for the random field to be reconstructed and a *biorthogonal* basis for the measurement domain that together exactly diagonalize the measurement operator. Shrinkage in this transformed domain then corresponds to projecting the measurements onto the nonorthogonal measurement basis, shrinking component by component, and then reconstructing using the orthogonal image domain basis. As with other shrinkage applications, these algorithms have important asymptotic optimality properties, many

[34] Where "complexity" can be assessed in terms of the loop structure of the graph or, more precisely, in terms of the complexity of the associated junction tree (see Section VI-A).

of which are related either explicitly or implicitly to the heavy-tailed nature of wavelet coefficients of imagery.[35] We refer the reader to the references for details.

While some of the approaches described in the preceding paragraphs can be interpreted as employing degenerate MR models (i.e., in which $A(s) = 0$ in (6) and where $w(s)$ is modeled as either Gaussian or heavy-tailed), there are other statistical approaches to inverse problems that involve more complex MR models. One that suggests itself based on the development in Section V-A, but that to our knowledge has not been explored, is to combine models that capture cascade behavior in wavelet coefficients (e.g., as in the hidden Markov trees described in Example 6) with the WVD decompositions described in the preceding paragraph. A second that has been studied in detail is that in [124] and [125] which uses nonlinear models (as briefly introduced in Section V-A and again in Section VI-B) that relate wavelet and scaling coefficients across scale in a pyramidal graphical structure that is generally more complex than a tree, leading the authors to employ an iterative estimation algorithm for the tomographic reconstruction problem on which they focus. Also, while the random field model that is used in that work is an MR model in the wavelet domain, the reconstruction for the tomographic reconstruction problem is performed directly in the image domain (to ensure positivity of the reconstruction), using the full nonlocal tomographic measurement operator so that the measurements are not local with respect to the nodes of this graph (which implies that the estimation structure is quite different than that described in Section III).[36]

While the examples described so far in this section are only tangentially related to MR models on trees or can be interpreted as involving degenerate MR models, there are also investigations that make much more explicit and substantive contact with the primary focus of this paper. One such example involving Poisson data and multiplicative scale-to-scale dynamics is illustrated in the following example.

*Example 7:* In this example, we take a brief look at MR models introduced and studied in [188], [261], and [322] in the context of Poisson measurements and in [279] for the modeling of traffic in communication networks. The idea of modeling counting processes at multiple resolutions has a substantial history, including a number of examinations of models for self-similar counting processes for a variety of applications [165], [201], [202], [221], [231], [311], [318].

The models introduced in [188] and [322] make use of very simple properties of Poisson counting processes, namely the facts that: 1) the numbers of counts of such a process over nonoverlapping intervals are independent and 2) the sum of independent Poisson random variables is also Poisson. For a 1-D Poisson process over a time interval $I$, these properties suggest a simple “dyadic partitioning” [188] of the interval in which we construct an MR tree exactly as in Fig. 1(a), in which each node is identified with a corresponding dyadic subinterval of $I$ (so that, for example, the nodes $s\alpha_1$ and $s\alpha_2$ in Fig. 1(a) correspond to the two halves of the interval corresponding to their parent $s$). The variable $x(s)$ placed at each node is then simply the number of counts over that subinterval. These variables have an obvious fine-to-coarse relationship, i.e., $x(s)$ is the sum of the two child values, $x(s\alpha_1)$ and $x(s\alpha_2)$, so that this model is what we have termed internal since the value at each node is a deterministic function of its descendants. This leads directly to a coarse-to-fine model, reflecting the fact that $x(s\alpha_1)$ and $x(s\alpha_2)$ are complementary fractions of $x(s)$

$$\begin{aligned} x(s\alpha_1) &= \delta(s)x(s) \\ x(s\alpha_2) &= [1-\delta(s)]\,x(s) \end{aligned} \tag{48}$$

where $\delta(s)$ is a random variable, independent from node to node, taking on values between zero and one. As discussed in [188], [263], and [322], the conditional distribution for each child conditioned on its parent is a binomial distribution with a parameter corresponding to the “fractional rate” of counts in one half interval versus the other. That is, if we let $\lambda(s)$ denote the mean of the number of counts over the interval corresponding to node $s$, then

$$p\left(x(s\alpha_1)|x(s),\,\lambda(s),\,\lambda(s\alpha_1)\right) = \mathcal{B}\left(x(s\alpha_1)|x(s),\,\frac{\lambda(s\alpha_1)}{\lambda(s)}\right) \tag{49}$$

where $\mathcal{B}(m|n,\,\rho)$ is the binomial distribution

$$\mathcal{B}\left(m|n,\,\rho\right) = \binom{n}{m}\rho^m(1-\rho)^{n-m}. \tag{50}$$

Thus, if the rate function of the Poisson counting process were known, (48) would be a simple MR tree model for $x(s)$.[37] However, of central interest for Poisson estimation and imaging applications is the case in which the rate function $\lambda(.)$ *itself* is random and is to be estimated from the observations of the count process $x(s)$. Since the mean of the sum of Poisson random variables is the sum of their individual means, the count function $\lambda(s)$ also has the same additivity property as the count process. Thus, we have the following parent–child relationships for the rate as well:

$$\begin{aligned} \lambda(s\alpha_1) &= \rho(s)\lambda(s) \\ \lambda(s\alpha_2) &= [1-\rho(s)]\,\lambda(s) \end{aligned} \tag{51}$$

[35] Note that shrinking individual coefficients of WVD decompositions implicitly corresponds to assuming that these transformed measurement coefficients are independent (see, e.g., [206] for an explicit discussion of this and related issues). While this is a perfectly valid probabilistic model—and one that leads to impressive results for many applications—it is worth noting that, if the measurements are corrupted by white noise, the coefficients of the measurement expansion are not exactly uncorrelated since the basis used is not orthogonal (but rather is half of a biorthogonal pair).

[36] In [289] and [290], some of these same authors develop an alternate MR modeling framework that works directly with MR pixels as state variables rather than wavelet and scaling coefficients. The approach to tomographic reconstruction in these references involves a coarse-to-fine/multigrid algorithm in which the data are used at each resolution to perform estimation assuming that the finer scale detail is absent, and a simple scale-to-scale pyramidal MR model. This graphical structure and the algorithms used are closely related to the one introduced in [42], which is illustrated in Example 10 in Section VII.

[37] Actually, (48) is not quite a tree model as $x(s\alpha_1)$ and $x(s\alpha_2)$ are not independent conditioned on $x(s)$, since their sum must equal $x(s)$. This is the same issue as mentioned in Section V-A concerning the relationship of (44) to tree models and which is examined and solved in Section VI-B by very simple state augmentation (in particular, in this example, $x(s)$ and $\delta(s)$ together become the state in the MR tree model).

where $\rho(s)$, which takes on values between zero and one, is precisely the fractional rate $\lambda(s\alpha_1)/\lambda(s)$ appearing as the binomial parameter in (49).

To complete the formulation of the problem of estimating $\lambda(s)$ given observation of $x(s)$, we need to specify the distributions for the independent random variables $\rho(s)$, as well as the random variable $\lambda(0)$ corresponding to the total mean count over the entire interval. As discussed in [188] [263], and [322], a particularly judicious set of choices are to model $\lambda(0)$ with a gamma distribution and each $\rho(s)$ with a beta distribution. In particular, this choice is *conjugate* to the choice of a Poisson distribution for the total number of counts $x(0)$ and binomial distributions in (49). With these choices, the conditional distribution for the rates and fractional rates have the same forms as their priors and are simply obtained by updating the parameters of these distributions using the observed count values $x(s)$. The estimation problem can then be put exactly in the form considered in Section III, leading to a two-sweep algorithm. Thanks to the internality and special structure of this model, the upward, fine-to-coarse sweep consists simply of summing counts over intervals, i.e., computing the unnormalized Haar scaling coefficients of the numbers of counts over dyadic intervals, followed by a nonlinear coarse-to-fine sweep to specify the parameters of the conditional distributions and hence the optimal estimates at each node. We refer the reader to [188], [263], and [322] for details.

It is important to make several comments at this point. First, the use of Haar scaling coefficients in this case is critical, as the structure of this model depends crucially on the summing of independent Poisson random variables. Hence, the same ideas do not work using other wavelets. Second, as mentioned in Section V-A, the multiplicative form of the models in (48) and (51) makes contact with other research on wavelet cascades and multifractals (see, e.g., [15], [94], [136], and [279]). Furthermore, by carefully choosing the parameters of the beta distributions for $\rho(s)$ or by using mixture models instead, one can obtain a wide variety of count models, including ones with self-similar or with bursty behavior, as well as estimation algorithms that have shrinkage-like characteristics. We refer the reader to [188] and [322] for details. Also, the extension of the models to 2-D deserves some comment. In particular, the approach described in [188], [263], and [322] uses an asymmetric model in which a coarse-level square at one resolution is first split horizontally and then each half is split vertically. Each of these two steps corresponds to splitting a region into two halves, resulting in models analogous to (48) and (51). As pointed out in [261] and [263], one could directly consider a quadtree structure in which counts over a square are split into four separate child counts over each of the quadrants. In this case, the conditional distributions for each child given the parent becomes multinomial rather than binomial, and the corresponding conjugate distributions for the fractional rates then become Dirichlet rather than beta. In either case, the resulting two-sweep estimation algorithm has exactly the same structure as in the 1-D case.

Finally, the estimation problem we have considered so far in this example is one in which we directly observe the counts as a function of time or space, i.e., we have direct measurements of counts for each time interval or each pixel. However, in many applications such as emission computed tomography, the estimation problem is a true inverse problem in that the total count recorded by each measurement detector is the sum of counts corresponding to photons from each pixel in the field to be imaged that impinge on the detector. At first glance, this problem seems to require very different solution techniques from the ones that we described previously in this example, since we do not directly observe the total number of photons emitted at each location in the image. However, in [263], the iterative expectation–maximization (EM) algorithm[38] is employed, each step of which involves MR algorithms of the type just described and which is guaranteed to converge to the optimal estimate. The key to this method is a clever choice of the so-called "complete information" for the EM algorithm, namely the number of photons impinging on each measurement detector from each individual image pixel. An example of the application of this procedure to simulated emission computed tomographic (ECT) data is illustrated in Fig. 12.

As a final comment, we note that most of the methods for inverse problems described in this section—and certainly those that involve taking wavelet transforms of the observed data—assume the availability of a regular set of nonlocal measurements. For example, the near-diagonalization of many such measurement operators using wavelet transforms rely on this regularity to transform the observed data into what can be modeled as measurements of individual wavelet coefficients. Of interest as well are problems in which we may have only a relatively few and irregularly spaced nonlocal measurements, e.g., as arises in data fusion problems such as described in Section II-F. In such a situation, wavelet-based methods do not apply (except in the very special and unlikely case that the nonlocal measurements correspond exactly or approximately to measurements of individual wavelet coefficients), and instead other methods are required to invert and fuse these data together with other measurements. In Section VI-B2 we describe a way in which this can be accomplished using MR models on trees.

## VI. MR Model Construction and Identification

As the development in Section IV makes clear, MR models offer the possibility of very efficient and scalable algorithms. A caveat to this, of course, is the requirement that the phenomenon and data of interest be well modeled within the MR framework, where "well modeled" refers not only to capturing the desired statistics to the required level of fidelity but also to the parsimony of the resulting model. In particular, while the complexity of various MR tree-based computations scales linearly with problem size as measured by the number of nodes in the tree, estimation algorithms for linear-Gaussian problems scale quadratically or cubically in the dimension of the state of the model at each node, while inference algorithms for discrete-state processes are polynomial in the cardinality of the state set at each node. Thus,

[38] See Section VI-C1 for a brief description of the EM algorithm.

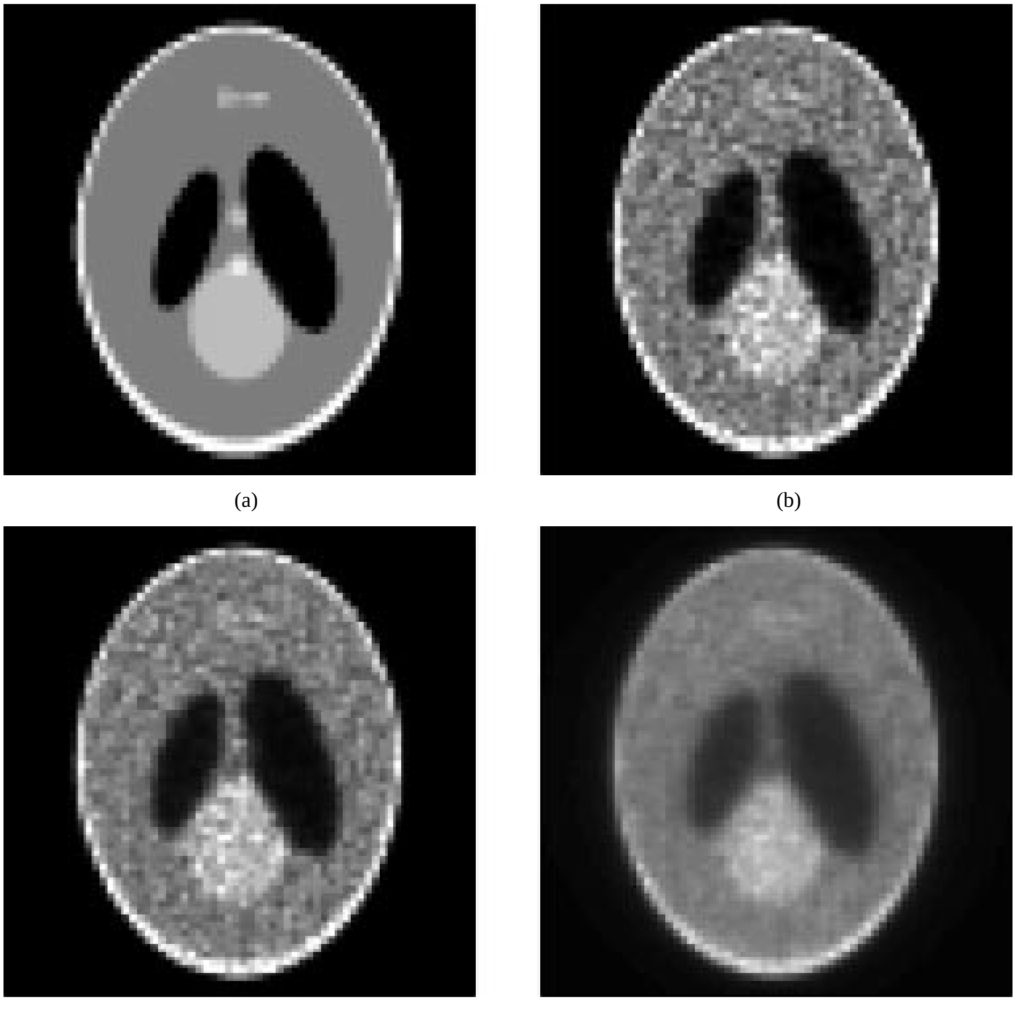


**Fig. 12.** (a) Noise-free image to be reconstructed from ECT measurements. (b)–(d) Three different reconstructions from ECT measurements using three different MR models corresponding to different parameter values and corresponding "strengths" of the prior models, ranging from weakest in (b) to strongest in (d) (for details see [263] from which this figure is reprinted).

a requirement for a *useful* MR model is that its state dimension or cardinality be comparatively small or increase at most modestly with problem size.

The examples in the preceding sections make clear that there is a substantial body of inference problems that can, indeed, be well modeled and solved using MR models on trees. In this section, we step back from specific examples and, for the most part, from predetermined notions of what variables (e.g., wavelet coefficients) populate the nodes of our model and take a more careful look at the topic of MR modeling from several different vantage and starting points. Modeling and model estimation are, of course, vast topics taking on a variety of guises in different fields, and we will make contact with several of these. While we certainly cannot explore all of these connections in real depth, we hope that the following serves not only to show the unique blend of ideas that enter into MR modeling but also to provide points of entry for further exploration.

The starting point for MR modeling has two components: 1) the available data or statistical information whose characteristics we wish to capture in our model and 2) the model class that is available to us. For the former, we have several possibilities, namely, sets of measurements of some or all of the variables that we wish to model; an explicit and complete probabilistic specification of the variables whose statistical variability we wish to capture in our model; or an implicit specification of that probabilistic specification in terms of a graphical model. (e.g., an MRF model). As for the specification of the class of models available to us, there are also several possibilities: models in which the nature of all of the variables at every node in the MR tree are already specified so that what is required for MR modeling is the identification of the coarse-to-fine probabilistic dynamics; models in which the tree structure is specified but only some of the variables are already specified (so that we must determine both the nature of the "hidden" variables and then the MR tree dynamics); and models in which not even the tree structure is specified so even that needs to be identified or learned. Interestingly, some of these possibilities are standard in some fields (e.g., signal processing, systems and control, or graphical models) and comparatively foreign to others, a point we highlight on several occasions in what follows.

### A. From Graphical Model to Tree Model

We begin with some ideas that are well known in the graphical models field but not nearly as common in the fields of signals and systems, as not only do they involve one idea that *is* known in those fields (namely the aggregation of variables or state augmentation) but one that is not usually considered, namely, a complete redefinition of the index set of the graph. For the purposes of discussions to follow, it is useful to describe two alternate viewpoints for such graphical constructions, namely, one based explicitly on separators or *cutsets* and one on the construct of a *junction tree* [108], [128], [143], [168], [169], [204], [207], [291], [314], [339].

*1) Cutset Models:* Consider a graphical model $x(.)$ on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, and suppose that $\mathcal{A}$ is a cutset of the graph (see Section III-C), so that $\mathcal{V}/\mathcal{A}$ is partitioned into two disconnected subsets $\mathcal{U}$ and $\mathcal{W}$ separated by $\mathcal{A}$. Each of the subsets $\mathcal{U}$ and $\mathcal{W}$ has its own induced graph (e.g., $\mathcal{G}_1 = (\mathcal{U}, \mathcal{E}_1)$, where $\mathcal{E}_1$ is the subset of $\mathcal{E}$ of all edges between elements of $\mathcal{U}$), and thus we can repeat this process (i.e., finding cutsets that partition graphs into disconnected subsets) on each of these two subgraphs and continue until we reach sufficiently small subsets of nodes (i.e., until an ultimate "finest" scale of singleton nodes).

Equation (10) then allows us to construct a tree model. For example, since $x_{\mathcal{U}}$ and $x_{\mathcal{W}}$ are conditionally independent given $x_{\mathcal{A}}$, we place the vector $x_{\mathcal{A}}$ at the root node of our tree and then continue the process using finer and finer cutsets and partitions. An example of this is depicted in Fig. 13. As shown in Fig. 13(b), the set $\mathcal{A} = \{3, 5, 7\}$ is a cutset, with $\mathcal{U} = \{1, 2, 6\}$ and $\mathcal{W} = \{4, 8, 9\}$, and in Fig. 13(c) we have indicated that $x_{\mathcal{A}} = \{x(3), x(5), x(7)\}$ plays the role of the state at the root node of the tree. Further, the singleton set $\mathcal{B} = \{1\}$ is a cutset for the subgraph on $\mathcal{U}$, which might lead one to take $x(1)$ as the state at the corresponding second-level node. However, $x(2)$ and $x(6)$ are not independent when conditioned on $x(1)$, since, in the full graph, there are other connections between nodes 2 and 6. However, if we *augment* the value of the state at this second-level cutset with the values at its parent, i.e., at the first level, then, conditioned on all of these values, $x(2)$ and $x(6)$ *are* independent. The result is a tree model as shown in Fig. 13(c) in solid lines.

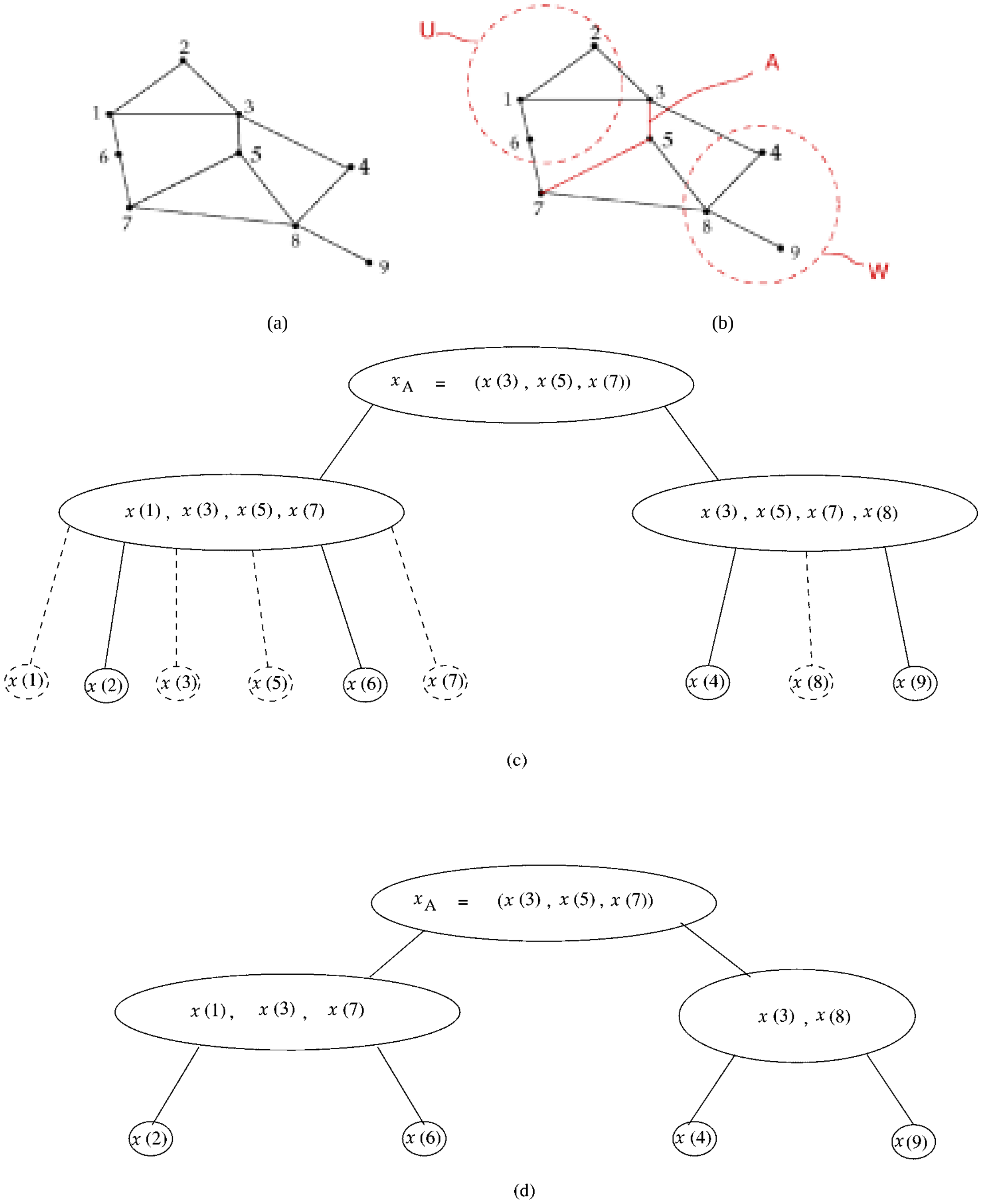
There are several points worth noting about this construction. The first is that the individual values of the graphical process in the model in solid lines show up at nodes at different levels in the tree. It is straightforward, however, to extend the model, as shown in Fig. 13(c) with the additional dashed lines, to a model in which all of the individual values of the process do indeed appear at the finest scale. This corresponds simply to a deterministic copying of some of the variables from the parent node, much as the root node value is copied to each of its children. Second, note that, when we incorporate the dashed part of the model in this figure, the resulting model is almost internal in the sense introduced in Section II. In particular, each of the states on the left branch of the tree (from root to leaf) is a deterministic function of its finest scale descendants. The same is not true on the right side, but this could easily be accomplished by a redundant copying of the values of $x(3)$, $x(5)$, and $x(7)$ to the finest scale.

Obviously, these models (corresponding to the solid lines in Fig. 13(c) or the solid and dashed lines together) have considerable redundancy, which may or may not be of value in a representation but certainly do increase dimensionality and thus raw computational complexity for algorithms using these models. Indeed, the simple fact of augmenting the state at a given node with the full state at its parent is generally unnecessarily complex. For example, if we examine the graph in Fig. 13(a) or (b), we see that nodes 2 and 6 are independent conditioned on less information than that contained in the $\mathcal{U}$-cutset $\{1\}$ augmented with the full cutset $\mathcal{A}$. Indeed, we only need to add the *boundary values* $\{x(3), x(7)\}$ [and not the internal value $x(5)$] to accomplish this. Similarly, we only need augment the $\mathcal{W}$-cutset $\{8\}$ with a single boundary value $x(3)$ in order to achieve the desired conditional independence. Thus, a considerably more parsimonious tree model is that shown in Fig. 13(d). We refer the reader to [156] for further discussion and examples of the construction of such reduced-dimension tree models.

Note that the model for Brownian motion described in Example 2 is a cutset model, and we now see that that style construction could be equally well used to construct an MR model for *any* temporal Markov chain or process. In this case, the dimensions of the states do not depend on the level in the MR representation nor on the extent of the time interval of interest essentially because of the very simple structure of minimal cutsets of the linear graph of a Markov chain. For more complex graphs, however, state dimensions can vary and, in particular, grow with the size of the original graph. For example, consider the regular 2-D nearest-neighbor graph shown in Fig. 14. In this case, an obvious cutset is the red line of nodes in the center of this grid: conditioned on the values of a graphical process on this cutset, the sets of values in the two remaining halves of the graph are independent.[39] Note, however, that in this case the cardinality of the cutset equals the linear dimension of the grid which (for square grids) is the square root of the total number of nodes, implying that the dimension of the state of a linear-Gaussian MR model using such a cutset is $O(\sqrt{N})$, while for a discrete-state model the cardinality of the state set corresponding to this cutset is exponentially large. Consequently, the resulting MR representations, which are closely related to so-called *nested dissection* methods in scientific computing and numerical linear algebra [133], lead to inference algorithms with complexities that do not scale linearly with problem size.

Finally, as we develop in more detail in Section VI-B2, it is often possible to construct reduced dimension approximate models in which we keep only a lower dimensional projection of the vector of cuset values at each node on the tree,

[39] Note also that the choice of a somewhat larger cutset, e.g., taking both the red and blue lines in Fig. 14, leads to partitioning of the graph into *four* disjoint components, which in turn leads to a quadtree MR structure. Obviously one can find other sets of cutsets that lead to more general MR trees (e.g., unbalanced trees, trees with differing numbers of children at each node, etc.).



(a)

(b)

(c)

(d)

**Fig. 13.** Illustrating the construction of cutset tree models. (a) A graph over which a graphical model $x(.)$ is defined. (b) Illustrating a particular choice of root cutset, $\mathcal{A}$, along with the disjoint subsets $\mathcal{U}$ and $\mathcal{W}$ separated by $\mathcal{A}$. (c) The cutset tree model resulting from this choice of root cutset; here the dashed lines indicate redundant values that can be added to the finest scale so that the entire process resides at the finest scale. (d) A lower dimensional model in which unnecessarily repeated variables are removed from various nodes.

neglecting the residual dependency that remains among descendent nodes when we condition only on this projection. A simple version of this approach was used in the texture discrimination problem described in Section II (Example 5) and [225]. In particular, rather than assuming that the four quadrants in Fig. 14 are independent given the entire set of values of the field along the red and blue boundaries, we assume that they are independent given only a coarser version of the values along these boundaries (essentially coarser 1-D wavelet approximations of these sets of boundary values). As shown in [224], such models can have artifacts across tree boundaries (which can be reduced somewhat by more judicious choice of lower dimensional projections, as described in Section VI-B2), but they are more than adequate for texture discrimination problems, as the results in Fig. 10 indicate.

*2) Junction Tree Models:* A second viewpoint on constructing a tree model from a given graphical model involves

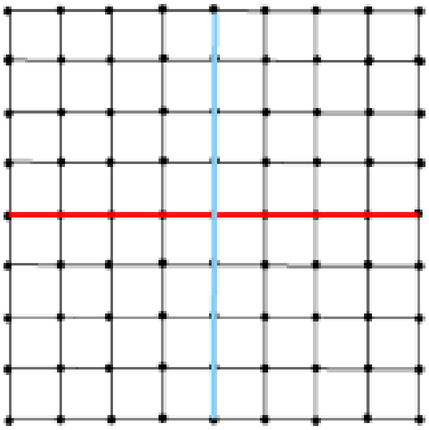


**Fig. 14.** A regular 2-D nearest-neighbor lattice, with a horizontal cutset (in red) containing a number of nodes equal to the linear dimension of the lattice. Taking this cutset together with the vertical one (in blue) leads to a quadtree MR model structure.

the concept of a *junction tree*. The first step in the construction of a junction tree representation is the addition of edges in order to *triangulate* the graph, resulting in what is also referred to as a *chordal* graph. In particular, such a graph is defined by the property that every cycle through the graph passing through four or more nodes contains a chord (i.e., an edge in the graph that connects two nonconsecutive nodes on the cycle). For example, in Fig. 15(a), we have depicted one triangulation of the graph from Fig. 13(a).

The second step in the construction is to identify a set of maximal cliques[40] of the triangulated graph that form a junction tree.[41] Such a tree is shown in Fig. 15(b). Note that each node of this graph corresponds to a maximal clique of the graph in Fig. 15(a). Furthermore, the corresponding vectors of values of the original graphical process that now populate this junction tree do indeed form a Markov model on the tree. For example, it is straightforward to see from the original graph that, conditioned on $\{x(3), x(5), x(6), x(8)\}$, the vectors $\{x(5), x(6), x(7), x(8)\}$, $\{x(1), x(2), x(3)\}$, and $\{x(3), x(4), x(8)\}$ are mutually independent. Further, the clique tree in the figure has the property required for it to be a junction tree: if a node of the original graph appears in cliques corresponding to two of the nodes in the junction tree, then it must also appear in the cliques of every node on the path between those two junction tree nodes. For example, node 8 of the original graph appears in both the root node and the bottom-right node $\{8, 9\}$, and it also appears in the intervening node $\{3, 4, 8\}$. The reason for this is very simple: the statistical model we wish to construct on this tree should be equivalent to the statistical model of the variables on the original graph and thus must provide unambiguous values for each of the variables in the original model. Consequently, variables such as $x(8)$ must, with probability 1, take on the same value in each of the vectors of clique variables involving node 8. This can be ensured if the junction property is satisfied, e.g., the MR dynamics for Fig. 15(b) simply copy the value of $x(8)$ from the root node to its two children that include node 8, and that value is then copied again to the bottom-right node. This notion of consistency in an MR model is very closely related to the idea of an internal MR model that we have introduced previously and which we explore in more depth in Section VI-B2.

Note that the construction of a cutset model implicitly provides a triangulation of the graph (in particular the nodes in each separator set can be viewed as cliques in an augmented graph). Typically, there are many different ways in which to triangulate a graph, and finding a triangulation that leads to a junction tree with small maximal cliques can be a challenging graph-theoretic problem [169], [204], [207], [314], [315].[42] Moreover, for some graphs, all triangulations have maximal cliques that are quite large. For example, the triangulation of the regular 2-D graph in Fig. 14 is nontrivial, requiring much more than simply adding diagonal connections across each of the 2 × 2 blocks of nodes [e.g., there are cycles that require triangulation much as the one that required the inclusion of the dashed edge in Fig. 15(a)]. On the other hand, for some graphs, the construction of low-dimensional MR cutset or junction tree models is straightforward. One such example is the MR model considered in [213] based on a quadtree structure as in Fig. 1(b) but with edges connecting each of the four children of each of the parents. In this case, each set of four children of a single parent form an elementary cutset, resulting in an MR model in which each such set of four variables is aggregated into a single node, resulting in an MR quadtree model.

### B. Methods for Constructing Linear MR Models

In this section, we describe several general approaches to linear MR model construction that make contact with the examples and discussions in previous sections and that also introduce several other important concepts. The perspective pursued in this section is motivated by image and signal processing applications in which the ultimate objectives, namely, performing estimation or hypothesis testing, are quite different than the objective of exact matching of a statistical model. Indeed, in many such applications, the implied or imposed prior model on the quantity to be estimated or analyzed is either only known approximately or represents statistical regularization rather than "truth"

[40] A maximal clique is a clique that is not contained as a proper subset of any other clique.

[41] In some treatments of graphical models, the junction tree explicitly includes intermediate nodes, representing *separator sets*, that is, sets of nodes that are common to the two cliques to which the separator is connected. For example, in Fig. 15, this would lead to including a node labeled $\{3, 8\}$ between the root node $\{3, 5, 6, 8\}$ and the node $\{3, 4, 8\}$. Including such nodes is often convenient, as it facilitates checking the junction tree property and identifying the explicit form of the resulting factorization of the overall distribution. However, such nodes are unnecessary probabilistically, since the coarse-to-fine dynamics can just as easily be defined on the model in Fig. 15(b).

[42] Note that cutset models do not generally lead to the smallest maximal clique sizes. For example, the MR model for samples of Brownian motion depicted in Fig. 6(b) corresponds to creating cliques consisting of subinterval endpoints and midpoints. In this case, the original graph, in which each time point is connected only to the points immediately before and after it, is already a junction tree. Thus, there is a modest increase in dimensionality using the MR cutset model. In other cases, such as cutting the grid in Fig. 14 with vertical and horizontal lines, the resulting clique sizes are of the same order as in optimal triangulations.

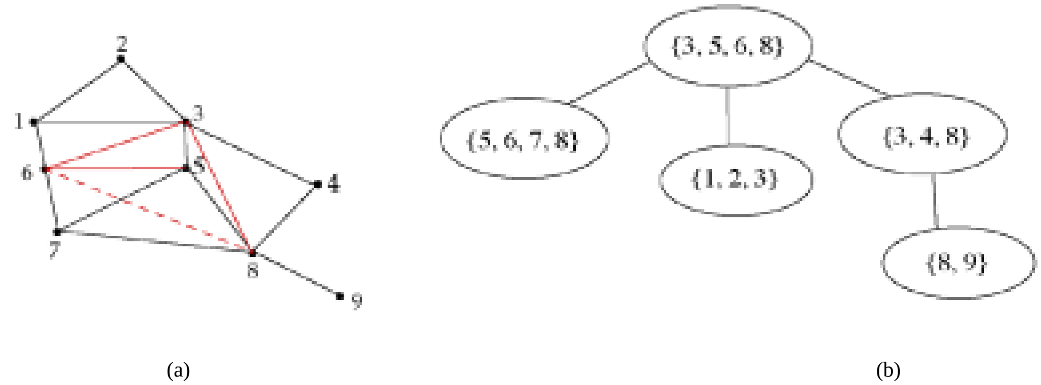


**Fig. 15.** (a) A triangulation of the graph in Fig. 13. (b) The junction tree for this triangulation.

about the phenomenon of interest. Such contexts suggest the idea of approximating or replacing such a prior model with an MR model of (hopefully) modest dimension that serves the desired purposes just as well in terms of capturing the expected behavior of the phenomenon and much better in terms of admitting efficient and scalable algorithms. This line of inquiry has been the subject of considerable attention, and in this section we describe four lines of thought that have been pursued for linear models.

*1) MR Variants of "Smoothness Priors":* Some of the earliest work on exploiting MR models resulted from a simple observation concerning a class of image processing algorithms based on variational formulations with what are often termed "smoothness priors." One example of such smoothness priors is that introduced in Section II-D (and revisited in Example 8 to follow) in the context of the problem of surface reconstruction. Another simpler example is the problem of image denoising. In particular, suppose that we observe $y(\boldsymbol{r})$, a noise-corrupted version of an underlying 2-D image $f(\boldsymbol{r})$, both defined over the 2-D image domain $\boldsymbol{r} \in \boldsymbol{I}$. One approach to denoising involves choosing as our estimate the function $f(\boldsymbol{r})$ that minimizes the following functional:

$$\alpha \int_{\boldsymbol{I}} [y(\boldsymbol{r}) - f(\boldsymbol{r})]^2 \, d\boldsymbol{r} + \beta \int_{\boldsymbol{I}} |\nabla f(\boldsymbol{r})|^2 \, d\boldsymbol{r}. \qquad (52)$$

The first term in (52) is a data fidelity term, while the second is a penalty on the size of the gradient of $f(\boldsymbol{r})$ penalizing the roughness of the reconstruction.

The first of these terms has a simple statistical interpretation, namely that $y(\boldsymbol{r})$ consists of measurements of $f(\boldsymbol{r})$ over $\boldsymbol{I}$ corrupted by 2-D white Gaussian noise of intensity $1/\alpha$. The precise statistical interpretation of the second term requires some care,[43] but, if we take the perspective that all we are attempting to do is to capture the *intent* of this regularization penalty, we can use a simple observation to replace it with a very simple MR quadtree model. In particular, as pointed out in [223] and [312], the second term in (52) can be thought of as a (Gaussian) "fractal prior." For example, in 1-D, each of the functions in Fig. 16 yields identical values for the 1-D version of the second term in (52) and, thus, are "equally likely" under this implied prior. Alternatively, as argued in [111] and [223], this penalty term corresponds to a Gaussian process with a fractal $1/f^2$ spectrum.

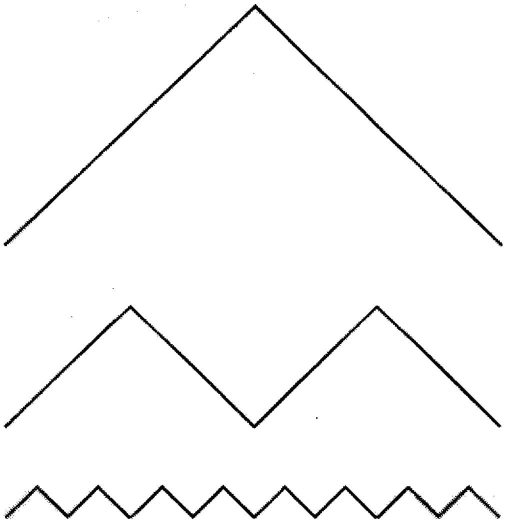


**Fig. 16.** A set of 1-D signals, each of which yields the same value for the 1-D version of the smoothness penalty in (52). Taken from [223].

As is well known in the literature (see, e.g., [116], [117], [313], and [349]) and as we have discussed in Examples 1 and 2 and already exploited in Example 4, the self-similar scaling behavior exhibited by such fractal processes can also be captured through simple MR models, such as the scalar model (7), with the variance of $w(s)$ taken to decrease geometrically as we move to finer scales. While definitely not identical to the prior model implied by the smoothness penalty in (52), this model has similar qualitatitve characteristics, and its use instead of the smoothness prior in (52) allows us to use the efficient two-sweep estimation algorithm of Section IV to compute both estimates and associated error variances. This should be contrasted with the computation of the optimal estimate corresponding to minimizing (52) which requires the solution of an elliptic PDE for the computation of the optimal estimates and essentially the inversion of the elliptic differential operator to calculate the corresponding error variances.

This same concept has been applied to somewhat more complicated image processing and computer vision problems (optical flow [223] and surface reconstruction [111]), and, in Example 8 to follow, we provide one such illustra-

[43] See, for example, [284].

tion, including showing how the error statistics computed by the two-sweep estimation algorithm can be used to localize significant anomalies, i.e., spatial locations at which the smoothness penalty should be relaxed (e.g., where there is an edge or abrupt change in the image or field being imaged).

The use of this simple MR model does, however, raise an issue that we have encountered previously (e.g., in Example 4) and revisit in more detail here, namely the possibility of artifacts in the estimates produced using MR tree models. In particular, consider the tree in Fig. 1(b), where the finest scale of this tree represents the actual image field to be reconstructed and on which we wish to impose a smoothness or fractal penalty. The simple MR model (7) certainly accomplishes the latter but does a spatially variable job of the former. In particular, consider the two finest scale shaded nodes near the center of the image in Fig. 1(b) but on opposite main branches of the MR tree. In this case, the tree distance between these nodes is far greater than the spatial distance and, as a result, reconstructions based on this MR model can lead to noticeable discontinuities across such major tree boundaries. As argued in [223], whether this is of any significance or not depends on the application. However, if it is of significance, this simple MR model is inadequate for the desired purpose.

In this case, there are four alternatives that one can consider. The first is relaxing the requirement of using a tree by allowing an edge between nodes across such tree boundaries. This, of course, leads to the requirement to solve estimation problems on loopy graphs, which, as we have discussed, can be complex. However, we return to this possibility in Section VII. A second possibility is to increase state dimension in order to capture the correlation across this major boundary more accurately. In particular, note that, conditioned on the root node of the tree in Fig. 1(b), the two shaded nodes are independent. Since smoothness dictates that these two pixel values are strongly *dependent*, this suggests that the state at the root node must capture all or at least a significant portion of this dependence.[44] This is the basic idea behind approaches described in the next topic. A third possibility is the one used in Example 4 and also in many of the investigations of other authors (see, e.g., [188], [261], [263], [281], and [322]), namely, to construct several MR models using trees that are spatially offset with respect to each other (so that their major boundaries do not coincide) and then to combine the results of estimation based on each of these trees.

Using multiple shifted trees is one approach to overcoming the problem of placing some nodes on one side or another of a major tree boundary when we actually would like these nodes to be on *all* sides of such boundaries. Another approach aimed at this same objective but employing only a single MR tree involves the use of so-called "overlapping trees" [159]. In standard MR tree models, such as that shown in Fig. 1(b), the nodes at each level of the tree correspond to nonoverlapping portions of the 2-D field being represented. For example, the standard quadtree in this figure has nodes corresponding to square regions of the field, each of which is subdivided into four nonoverlapping subregions at the next finer scale. In the overlapping tree framework, each region is subdivided into overlapping subregions at the next scale, so that each of these subregions has linear dimensions somewhat larger than half of the dimensions of its parent.

There are several consequences of this construction. The first is that an overlapping tree will always have more resolutions than a standard one, since the linear dimensions decrease more slowly from level to level. As a consequence, the total number of nodes in an overlapped tree model is larger than that for a nonoverlapped tree model and, in fact, increases geometrically with the number of additional scales that are added, implying that the number of scales that can be added (and hence the degree of overlap achieved) needs to be carefully controlled in order to maintain computational feasibility. Second, each finest scale pixel in the image domain actually corresponds to *several* finest scale nodes in the overlapped tree. Indeed, this is precisely the intention of the construct, as the set of fine-scale nodes that correspond to a single real pixel are guaranteed to include elements that reside on different sides of major tree boundaries. Of course, this then creates the question of how to *use* such a tree when the real data and desired estimates both reside in real image space. As described in [159], there is a straightforward way in which to "lift" the image domain estimation problem to the redundant overlapped domain. This construct involves two linear operators, one of which takes an image domain pixel value and replicates it at each of the redundant nodes corresponding to that pixel. The second operator (which is a left inverse of the first) collapses the values at each set of redundant tree nodes into a single weighted average, which is then mapped to the corresponding image pixel. An estimator using an overlapped tree, then, consists of the application of the first operator to lift the image domain measurements to the tree domain, followed by the application of the tree estimation procedure described in Section IV-B, and then by the application of the second operator to project the tree estimates back to the image domain.[45]

*Example 8:* One example of a successful application of this overlapping technique is to the surface reconstruction problem introduced in Section II-D. The full development in [111] combines several important features. The first is the construction of MR models that accommodate *both* of the smoothness terms (thin membrane and thin plate penalties) in (2) simultaneously. The second is overcoming the need to deal explicitly with the integrability condition mentioned in Section II-D and captured in the consistency relation (3). The method used to deal with these in [111] is to

[44] For 1-D signals, it is always possible to maintain continuity of the signal modeled by an MR model by using an approach as in Example 2 in which endpoints of intervals are included in the states at each node in the tree. In 2-D, the problem is much more complicated, motivating the development in Section VI-B2.

[45] Note that it appears that this approach "overcounts" each real measurement by replicating it at several tree nodes. As shown in [159], this is avoided by modeling each of the resulting tree measurements as having measurement noise with covariance that is a multiple of the covariance of the real measurement, where the multiple used is the cardinality of the set of tree nodes that correspond to the real image pixel being measured.

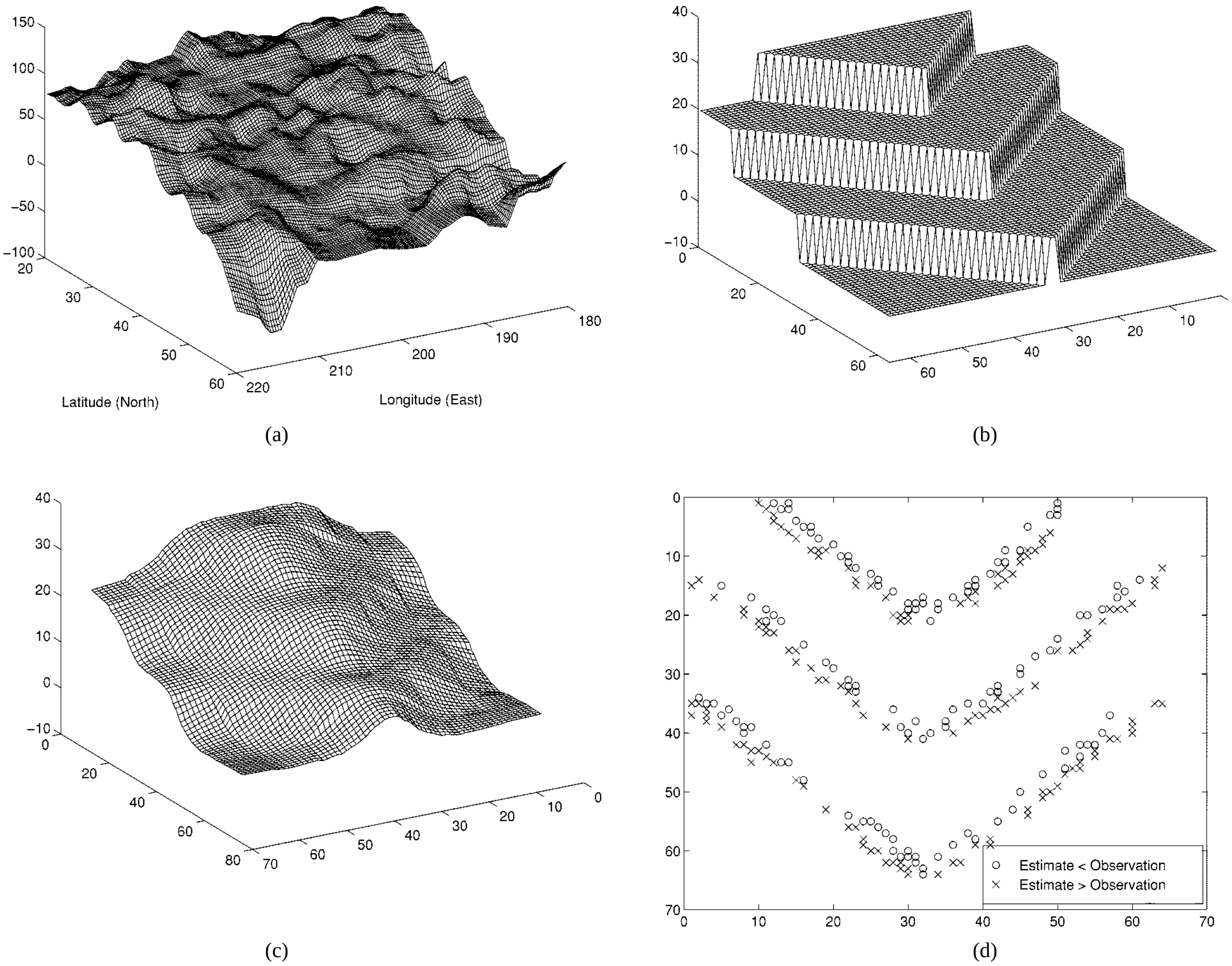


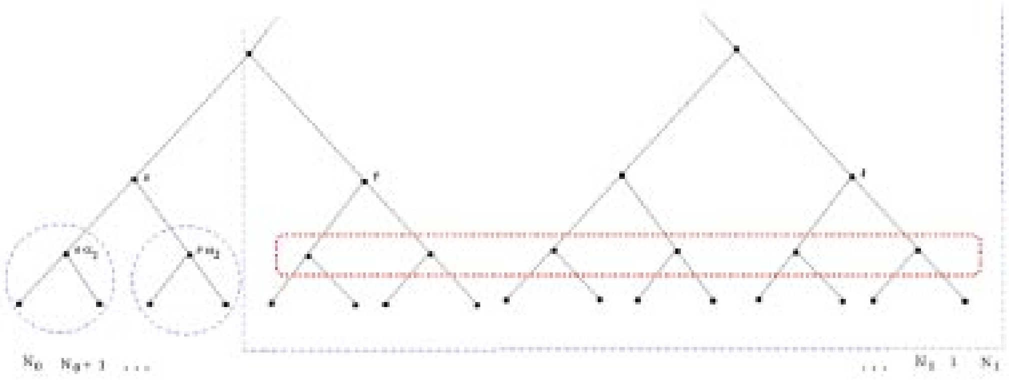
**Fig. 17.** (a) Ocean surface reconstruction from TOPEX/POSEIDON data such as in Fig. 2. (b) Discontinuous surface to be reconstructed. (c) Reconstruction of the surface given noisy measurements and using the MR estimation algorithm described in the paper employing thin-plate and membrane priors which lead to smooth reconstructions across surface discontinuities. (d) Distribution of locations and signs of statistically significant measurement residuals, providing clear statistical evidence for the detection, localization, and estimation of the surface discontinuities. (Reprinted from [120].)

construct a MR model with a 3-D state that approximates both of the smoothness penalties as well as the consistency condition (3). The three state variables at each node correspond to surface height $z$ and surface gradient $(p, q)$ at each scale and spatial location. In addition, we replace the hard constraint (3) with a softer one, consisting of an additional "measurement" at each node, corresponding to requiring that the difference between the gradient[46] of $z$ and the value of the vector $(p, q)$ at each node is a zero-mean white noise process on the tree, with small variance. Finally, the third component of the model construction is the use of overlapping, which has the effect of modifying the geometric decay in the covariance of the process noise $w(s)$ as one moves to finer resolutions.

Fig. 17 shows an illustration of this method to such surface reconstruction problems. Fig. 17(a) illustrates the effectiveness of this method for the reconstruction of a relatively smooth surface [namely ocean surface height as discussed in Example 4 but here using a more sophisticated MR model based on (2)], while Fig. 17(b) and (c) shows the well-known problem that this as well as other smoothness-based reconstruction methods have when there are discontinuities in the actual field being reconstructed: the smoothness penalty leads to blurring or smoothing of the surface across the discontinuity. However, because of the ready availability of error variances when using MR methods, we can employ these statistics to identify regions in which the difference between the raw surface measurements and the reconstructed surface are statistically anomalous. Fig. 17(d) illustrates the result in this case, in which we have also indicated the sign of these anomalies, which provide clear indication of the nature of the discontinuous jumps in surface height in this example.

The fact that MR error statistics allow us to localize discontinuities such as in Fig. 17(d) suggests that it should be possible to develop an adaptive algorithm that performs optimal estimation without blurring across high-contrast edges or discontinuities. One such method is developed in [292] by considering a more complex MR model, namely, one which leads to estimates not only of surface height but also of an auxiliary field [roughly corresponding to $\alpha_3(\boldsymbol{s})$ and $\alpha_4(\boldsymbol{s})$ in (2)] that controls the spatially varying smoothness of the

[46] Actually a discrete, multiscale approximation to the gradient (see [111]).



(a) (b)

**Fig. 18.** Illustrating the role of the state in linear-Gaussian MR models. (a) In general, conditioned on the value of the state $x(s)$, the three sets of variables indicated in dashed lines must be uncorrelated with each other. (b) For internal MR models, the state $x(s)$ need only decorrelate $x(s\alpha_1)$, $x(s\alpha_2)$, and the set of values in the single set of nodes encircled by the dotted red line—the remaining decorrelation required as in (a) is automatically satisfied thanks to internality.

field.[47] An alternate approach, developed in [324], addresses this and other related problems using an expectation–maximization formalism.

*2) Internal Models and Approximate Stochastic Realization:* In this section, we describe a more general and formal construction of linear MR models [82], [85], [122], [161]. The approach makes use of concepts adapted from state space theory [8], [9], [16], [214]; however, the adaptation to trees uncovers some important differences with the temporal case. First, in contrast to the usual temporal state space framework—and, for that matter, to the framework implicitly used in most graph-theoretic studies—we consider problems in which the random process or field whose statistical behavior we wish to realize corresponds to variables on only a fraction of the nodes of the MR tree. In particular, we focus initially (and primarily) on problems in which the process to be realized, whose covariance or power spectrum is assumed to be given (and which, for simplicity, we assume is zero-mean) resides only at the finest scale of the tree, i.e., at the leaf nodes. In standard time series analysis, in which the index set is a completely ordered interval of integers, the set of "leaf" nodes is a singleton consisting of the single point at one end of the interval. However, for dyadic trees or quadtrees, as in Fig. 1, the finest scale can accommodate entire 1-D or 2-D processes.

Another issue that arises in constructing MR models is specifying how to map the 1-D or 2-D process of interest to the finest scale of the MR tree, a specification that then also determines which finest scale values have common parents, common grandparents, etc. The problem of estimating or identifying this graph structure does not arise in standard time series analysis but is an important problem in its own right both for trees and for more general graphs. We return to this problem in Section VI-C3, but, for this discussion, we assume that this structure is fixed and given. A typical example,[48] to which we refer on occasion, is that shown in Fig. 18. Here we begin with a zero-mean Gaussian process $z[n]$, $N_0 \leq n \leq N_1$, whose second-order statistics are given and which we wish to realize, either exactly or approximately, at the finest scale of the MR tree shown in the figure, in which we have indicated that the finest scale nodes are mapped to consecutive integers over the interval in question. Note that, with this ordering, each node at coarser scales corresponds to an interval of the process (e.g., the node labeled s in the figure corresponds to the interval $[N_0, N_0+3]$), which is exactly the type of correspondence we have seen before, e.g., for Brownian motion in Example 2.

With the graph structure fixed the problem of MR modeling bears a number of similarities—and some significant differences—with state space modeling for time series. For example, we refer the reader to [29] for the development of a state space theory for deterministic MR dynamics on trees and to [24], [25], [65], and [66] in which MR counterparts to autoregressive modeling and efficient algorithms analogous to Levinson's algorithm for time series are developed for the class of isotropic processes on trees (i.e., processes in which the covariance between variables at different nodes depends only on the distance between the nodes). In this section, we focus on problems of approximate stochastic realization, where, as in standard state space realization theory we must deal with two basic issues: 1) we need to define the "state" of the process at each of the unspecified, coarser-scale nodes of the tree and 2) we must then define the coarse-to-fine dynamics among these state variables, resulting in a model that is Markov with respect to the MR tree. Because of the special property of trees, the first problem bears a resemblance to the standard temporal problem for Gaussian processes [8], [9], [214] in which the role of the state at any point in time is to decorrelate two sets of variables, namely, the values of the process in the past and the values in the future. Referring to the tree in Fig. 18 (and the blue dashed lines therein), the role, then, of the state $x(s)$ in this example

[47] The formulation in [292] represents a relaxed version of the widely-studied Mumford–Shah functional for image denoising and segmentation [11], [252].

[48] For simplicity we illustrate these concepts using dyadic trees and 1-D examples; however, the same ideas work for quadtrees and 2-D fields.

is to decorrelate *three* sets of variables, namely those in the subtree rooted at node $s\alpha_1$, those in the subtree rooted at $s\alpha_2$, and the large set of variables over the entire tree except for those in the subtree rooted at node $s$.

Note that the definition of the state in this case is given in a manner that couples the definitions at distinct nodes, e.g., $x(s)$ in Fig. 18 must decorrelate $x(t)$ from $x(s\alpha_1)$, a fact that makes the general analysis of realization more complex than its temporal counterpart, and a complete treatment still remains to be developed. However, a considerable amount can be said if we borrow another concept from temporal realization theory, and one whose MR counterpart we have encountered informally on several occasions already, namely, that of an *internal model*. For temporal systems, an internal model is one in which the state at each point in time is a deterministic function of the process being realized. In particular, the state of an internal model is typically taken to be a function of the past of the process to be modeled, capturing the memory in the process required to decorrelate past and future.[49] Thus, an internal state does not introduce any additional randomness not present in the process to be realized. A very important result in standard state space realization theory [8], [9], [214] is that it is always possible to find internal state space realizations that are *minimal*, i.e., that have the smallest state dimension possible, so that considering noninternal realizations does not buy us anything in terms of model dimensionality.

By analogy, for our problem, we define the concept of an internal model as one in which the state of the model $x(\sigma)$, at any node $\sigma$, is a deterministic, linear function of the values of the process at the finest scale in the subtree descending from $\sigma$. For example, the state of an internal model at node $s$ in Fig. 18 must be a linear function of the values of $z[n]$ for $n = N_0$, $N_0 + 1$, $N_0 + 2$, and $N_0 + 3$. As shown in [122] and [161], there are several very significant implications of restricting attention to internal models. First, of course, these models do not introduce additional randomness at coarser scales (and thus coarse-scale nodes, while technically not observed, are not really "hidden"). Second, there is a potential price to be paid by excluding noninternal models, as it is sometimes possible to find noninternal models of smaller state dimension. However, third, and very importantly, internality considerably reduces the complexity of constructing state variables. In particular, as shown in [122], in this case, the states can be defined resolution by resolution, as it is sufficient to design the state at each node simply by looking at the nodes one scale finer. For example, the state $x(s)$ in Fig. 18 must be a linear function of the states at its children $x(s\alpha_1)$ and $x(s\alpha_2)$, and it is sufficient to choose that linear function so that $x(s)$ decorrelates its two descendent values from each other and from the remaining nodes at that next finer scale (the red dotted region in the figure). Note that, since each state is a linear functional of its children, we maintain consistency along paths from coarser to finer scales. In particular, the fact that $x(s)$ consists of linear functionals of its two children implies that the coarse-to-fine dynamics (6) must deterministically "copy" the information from each parent to its children. This consistency, while different in detail, is identical in purpose to the consistency requirements for a junction tree as described in Section VI-A2.

Note also that, once we have performed this fine-to-coarse process of defining the state at every node, we can immediately determine the coarse-to-fine dynamics, i.e., the matrices $A(s)$ in (6) and $Q(s)$, the covariance of $w(s)$ in (6). In particular, $A(s)x(s\overline{\gamma})$ is the best estimate of $x(s)$ given $x(s\overline{\gamma})$, and $w(s)$ is simply the error in this estimate. Thus, computing $A(s)$ and $Q(s)$ requires only the joint statistics of $x(s\overline{\gamma})$ and $x(s)$. Since both of these states are linear functions of the finest scale process (which, in our 1-D example in Fig. 18, is the process $z[n]$ whose second-order statistics have been specified), the joint statistics of $x(s)$ and $x(s\overline{\gamma})$ can be computed in terms of the statistics of the finest scale.

Let us return to the issue of constructing the states at individual nodes. There are two points that make this problem challenging even in this form. The first is that the required state dimensionality to achieve complete decorrelation may be prohibitively high. For example, note that the state at the root of any of our trees must, in principle, completely decorrelate the disjoint regions associated with each of the root node's children, a task that may require very high state dimension (e.g., as we saw in Section VI-A1 for the realization of 2-D MRFs on regular nearest-neighbor lattices). Large state dimensionality, of course, is also a problem in standard state space realization, and once again we adopt ideas from that context. In particular, we may wish to construct *reduced-order models* that yield realizations at the finest scale that only approximate the desired statistics of the process we are attempting to model. In fact, suppose that we follow such a procedure (yet to be defined) to define reduced-order states in a fine-to-coarse manner, so that each state is still a linear function of its children. In this case, we can still define the coarse-to-fine MR dynamic matrices $A(s)$ and $Q(s)$ using the same procedure as previously described, and we can also be assured that the "noise" $w(s)$ is uncorrelated with $x(s\overline{\gamma})$, since $w(s)$ is simply the error in an estimate based on $x(s\overline{\gamma})$. However, since we have used reduced-order states, so that in particular $x(s\overline{\gamma})$ may not completely decorrelate its two children, it will generally not be true that the values of $w(s)$ are white over the entire tree.[50] This is also the case for standard state space models, where the idea is to neglect these residual correlations; that is, we use the model dynamics $A(s)$ and $Q(s)$ we have constructed and simply assume that $w(s)$ is white. The resulting model then yields a process that has statistics at the finest scale that do not completely match those of the target process $z[n]$. Specifically, the variance of individual samples of $z[n]$ will be captured exactly by this model, but the cross-covariances at different points in time may not be realized exactly.

[49] It is equally possible, however, to define the state of an internal model as a function of both the past and future. See, for example, [214].

[50] The noises $w(.)$ *are* guaranteed to be white on coarse-to-fine paths but not necessarily uncorrelated between nodes (such as $s\alpha_1$ and $s\alpha_2$) that are not on the same path from the finest scale to the root node. As a result, when we construct our reduced-order model by neglecting this residual correlation, we are guaranteed that *some* of the desired statistics are preserved in this approximate model, e.g., the statistics along any coarse-to-fine path and, in particular, the covariances of the individual finest scale nodes, but not the complete statistical description of the finest scale process.

It remains to specify precisely how state-order reduction is to be carried out. For standard state space models, this is a well-developed field, particularly for time-invariant systems [8], [9], [214] in which state definitions are identical across all points in time, so that the impact of the decision on how to define the state on global measures of fidelity can be computed. In contrast, except in very special situations (e.g., in Example 2 or in the more general MR modeling of self-similar Gaussian processes in [83]), the states at different resolutions in MR models represent very different quantities and may, in fact, have varying dimension. Moreover, one of the major domains of application of these methods involves highly nonstationary phenomena. As a result, the use of standard global measures of model accuracy (e.g., such as the Kullback–Leibler divergence) have not yet led to computationally tractable algorithms for model construction, and thus the methods that have been developed (and borrowed from standard time series contexts) focus on local criteria for defining states at individual nodes. In particular, the fundamental objective in defining the state $x(s)$ in Fig. 18 is to have it decorrelate $x(s\alpha_1)$, $x(s\alpha_2)$, and the set of variables inside the dashed region in the figure. Consequently, it is natural to consider choosing approximations in terms of how well they perform this decorrelation.

In the context of standard state space systems, this idea has led to the use of the statistical notion of *canonical correlations* [9]. Roughly speaking, canonical correlation analysis takes two random vectors $z_1$ and $z_2$ and identifies ranked-ordered pairs of linear functionals of each: (${c_1}^T z_1$, ${d_1}^T z_2$), ($c_2^T z_1$, $d_2^T z_2$), where each of these individual variables has unit variance, the functionals of $z_1$ are all uncorrelated with each other, as are the functionals of $z_2$, and the pairs of functionals ($c_i^T z_1$, $d_i^T z_2$) have correlation $\lambda_i \geq 0$, with $\lambda_1 \geq \lambda_2 \geq \cdots$ (so that ($c_1^T z_1$, $d_1^T z_2$) represents the most highly correlated functionals of the two vectors $z_1$ and $z_2$, ($c_2^T z_1$, $d_2^T z_2$), the next most correlated). In the context of time series, $z_1$ represents the past of the process and $z_2$ the future, and the functionals $c_i^T z_1$ form a rank-ordered set of natural state variables, providing a quantitative basis for choosing the state variables to keep in a reduced-order model.[51]

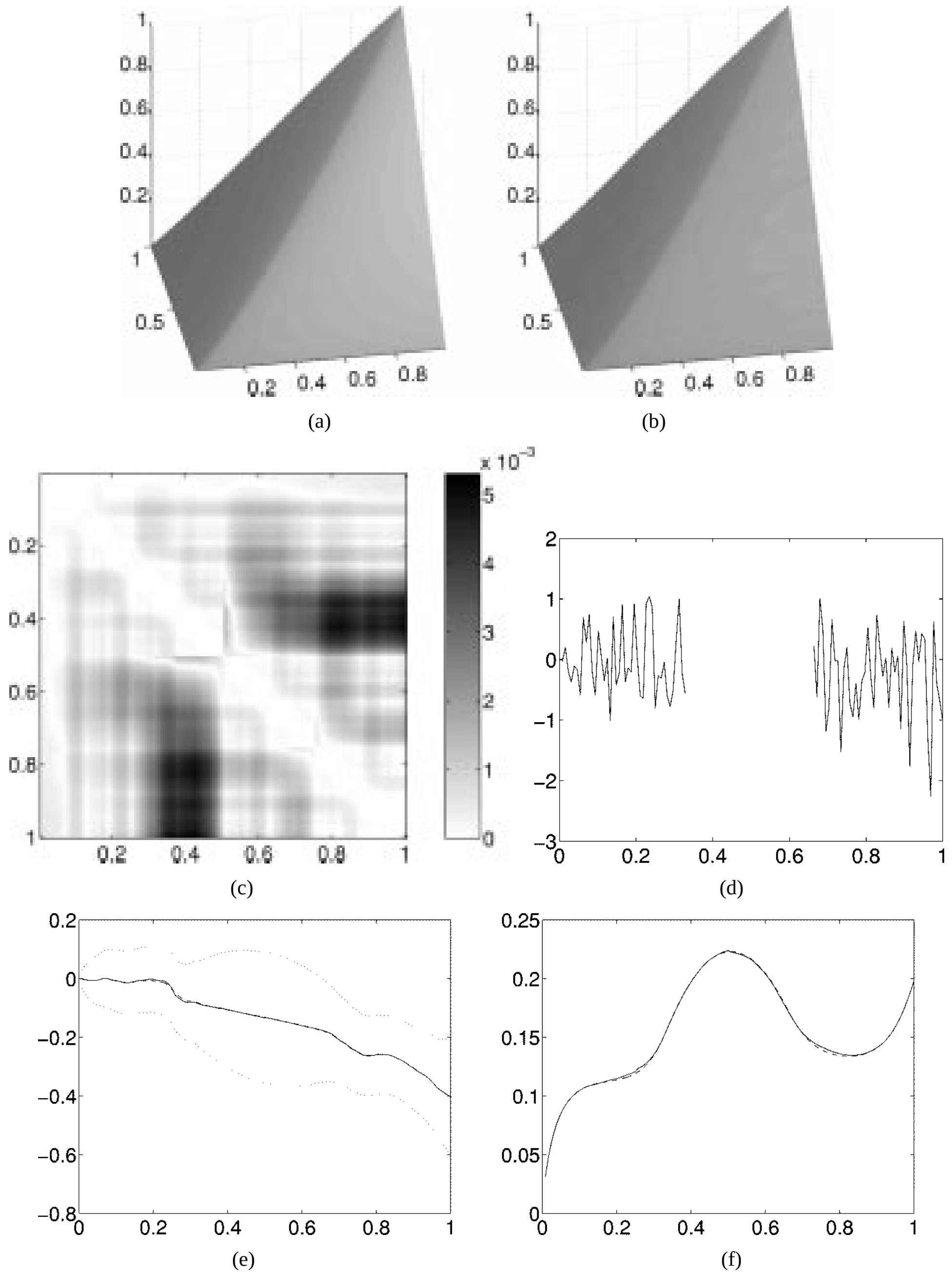
As discussed in [16] and [122], one of the potential drawbacks of the metric corresponding to canonical correlations is that the rank-ordering is based on correlation coefficients (since the linear functionals are normalized to have unit variance). As a result, components of the past and future that contribute very little to the total variance of either of these may rank higher than components that contain much more of the process variance. This observation suggests an alternate criterion referred to as *predictive efficiency*, which, as opposed to canonical correlations, treats the variables involved asymmetrically: for the two random vectors $z_1$ and $z_2$, the objective is to produce a set of scalar linear functionals $c_i^T z_1$ which are uncorrelated with respect to each other and are rank-ordered in terms of the amount of variance reduction each of these functionals provides toward the estimation of $z_2$. If $z_1$ represents the past of a time series and $z_2$ the future, this provides an alternative criterion for ranking possible state variables in a time series model.

The computation of functionals for either the canonical correlations or predictive efficiency criteria involves the singular value decomposition (SVD) of a matrix derived from the covariance matrix of the two sets of variables. For time series, both of these are typically of the same dimension, so that there is not a computational advantage to either approach. However, for MR tree models, this is generally not the case, and, in addition there are other important differences with the time series case. For example, consider the node $s$ in Fig. 18. Note that in this case we wish to design $x(s)$ to decorrelate not two but *three* random vectors, namely $x(s\alpha_1)$, $x(s\alpha_2)$, and the very large vector (call it $\chi$) of values of $x(.)$ at the remaining nodes at the same scale as $s\alpha_1$ and $s\alpha_2$. The first complication is that we have three rather than two variables, making the problem of defining what we mean by the "best" variables to include in $x(s)$ more complex. Second, the dimensions of the three vectors are quite different, and this asymmetry greatly favors using the concept of predictive efficiency, with $\chi$ always playing the role of all or part of $z_2$. As described in [122], this then leads to a procedure in which we sequentially add variables to form $x(s)$, first including some linear functionals of $x(s\alpha_1)$ of most value in estimating $x(s\alpha_2)$ and $\chi$, then adding functionals of $x(s\alpha_2)$ that are of most *additional* value (i.e., taking into account the functionals already constructed), and then possibly alternately adding functionals of $x(s\alpha_1)$ and $x(s\alpha_2)$ to enhance the fidelity of the state design.[52] The resulting approach to constructing an MR model has $O(N^2)$ complexity. This complexity may not be prohibitive for some applications, since it need only be performed once to build the model. On the other hand, in many cases, the complexity is still too large and, more to the point, is often wasteful. For example, in many time series problems, the primary correlation between the past and future relative to some time $t$ is captured in a comparatively small interval around that time. This suggests, for example, rather than using all of the $O(N)$-dimensional vector $\chi$ in Fig. 18 in order to define $x(s)$, we might safely replace $\chi$ with an $O(1)$-dimensional vector of the values closest to the nodes $s$, $s\alpha_1$, and $s\alpha_2$. Using this so-called *boundary approximation* yields an algorithm of complexity $O(N)$.

A number of examples illustrating the use of this method are given in [122]. For example, if applied to a Markov process, such as Brownian motion, the algorithm does indeed identify that the correct linear functionals to keep at each node consist of the boundary points of the time interval corresponding to that node, as in Example 2. Similarly, for 2-D nearest-neighbor MRF s, in which each of the tree nodes

[51] Typically this is done in one of two ways: 1) we fix the state dimension, say at a value $d$, and thus choose the functionals corresponding to the $d$ largest of the $\lambda_i$ or 2) we set a threshold for residual correlation and keep as many functionals as needed so that the sum of the remaining $\lambda_i$ falls at or below the threshold.

[52] Note that for a quadtree there are five sets of variables that need to be decorrelated by the state at any node, namely the variables at each of the four children and the vector of all other tree variables at the same scale as those children.



(a) (b) (c) (d) (e) (f)

**Fig. 19.** Illustrating the result of applying the scale-recursive method for constructing internal approximate MR stochastic realizations, in this case of fBm (with Hurst parameter $H = 0.7$). (a) Plot of the exact covariance matrix for a window of fBm. (b) Plot of the covariance achieved using an MR model with state dimension four. (c) The difference between the covariances in (a) and (b), plotted as an image. (d) A set of noisy measurements of this fBm process over the two ends of the interval of interest. (e) The estimates using the MR algorithm and the four-dimensional state model (solid line), the optimal estimates using the exact fBm statistics (dashed line almost completely obscured by the solid line), and plus/minus one standard deviation error bars (dotted line). (f) Error standard deviations given by the MR estimator (solid line) and based on the exact fBm statistics (dashed line, again almost completely obscured by the solid line). (Reprinted from [122].)

corresponds to a square region of the 2-D image domain, the full state required to decorrelate one such region from the rest of the image consists of the values around the boundary of the region—exactly the type of construction we saw in the cutset models in Section VI-A1. However, the framework we have described here allows us to consider reducing the dimensionality of the full cutset state by keeping only those functionals of the process around the boundary that rank highest in terms of predictive efficiency in estimating the rest of the domain.[53]

Fig. 19 illustrates another example, in this case to the approximate modeling of fBm, a process that is not Markov but

[53] An alternate approach to reducing complexty of approximate cutset models involves not reducing the dimension of the state corresponding to a boundary in a 2-D MRF but rather to reducing the complexity of the *model* for that state. We briefly discuss this idea in Section VII.

that does possess fractal, self-similar scaling properties that generalize those of Brownian motion.[54] As shown in [83], the optimal choices of linear functionals to keep in the state at each node also have approximate self-similarity in scale. In this figure, we show both the approximation errors in the realized process covariance and also what is really the most important result, namely that the differences in estimation accuracy using the exact fBm statistics or the approximate ones captured in the MR model are statistically insignicant (especially given that fBm itself represents a mathematical idealization of real processes).

Finally, it is worth noting that the scale-recursive procedure for state construction that we have described has an important extension [82] to problems in which we have specific functionals of the finest scale process that we would like to include as state variables at coarser scales in the tree. As a simple example, consider a fine-scale process $z[n]$, $N_0 \leq n \leq N_1$, and suppose that we require that a particular linear functional of the finest scale, e.g., a linear combination of $z(N_0)$, $z(N_0+1)$, $z(N_0+2)$, and $z(N_0+3)$, be available as a component of the state at some node. In this case, the first common ancestor of the points $N_0, \ldots, N_0+3$ is the node $s$ in Fig. 18. However, as discussed in [82], [122], and [161], placing the desired linear functional directly at this node is *not* generally advisable, as conditioning on this linear functional can actually *increase* the correlation between the variables at the children of node $s$. In particular, the correlation between two random vectors is always reduced if we condition on a linear functional of one or the other alone, but may, in fact, increase when conditioned on a linear combination involving *both* vectors (e.g., two uncorrelated random variables are no longer uncorrelated when conditioned on their sum). As a result, we actually place *two* nonlocal functionals at node $s$, namely, the separate functionals of $z(N_0)$, $z(N_0+1)$ and of $z(N_0+2)$, $z(N_0+3)$ whose sum is the required linear functional. To maintain internality, then, we need to ensure that these individual linear functionals can, in fact, be expressed as linear functionals of $x(s\alpha_1)$ and $x(s\alpha_2)$, respectively, and this, in general, will also specify several components of each of these states as well.

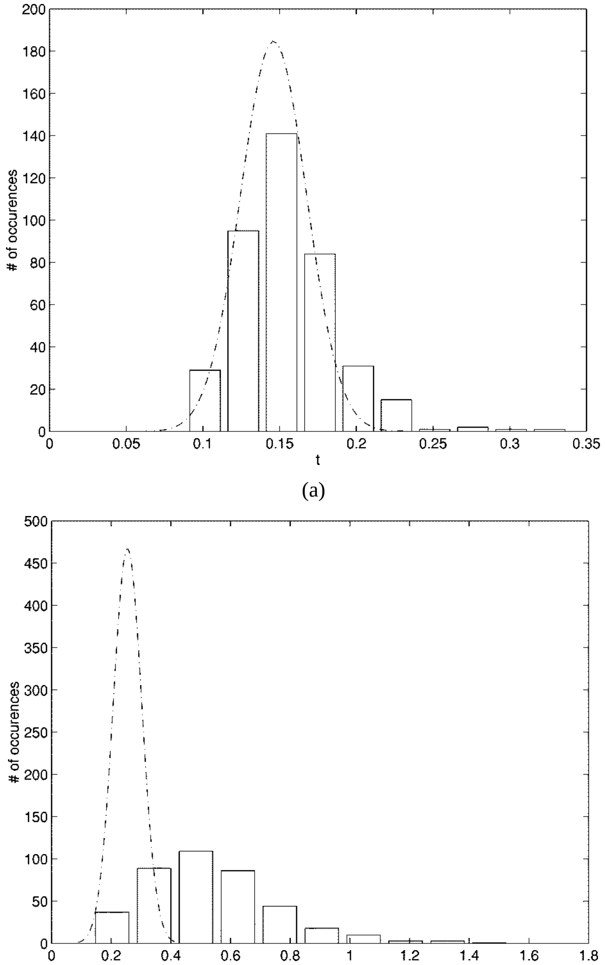
In general, this purely algebraic process involves successive examination of each of the descendants of nodes at which nonlocal functionals have been placed. At each such node, internality first requires that specific functionals be available, while the requirement of reducing rather than increasing correlation among the children of *that* node typically leads to each of these functionals be broken into several separate functionals. Once we have completed this process, we can then begin the fine-to-coarse construction of the full state at each node, except that now, when we come to design the state at each node, we may already have several components of that state prespecified. In this case, the only change to the procedure that we have outlined is that the choice of additional linear functionals is based on the measure of predictive efficiency taking into account the information already provided by the prespecified functionals. As the next example illustrates, the ability to incorporate particular functionals of the field in question as states of our MR model allows us to fuse MR measurements and estimate specific coarse-scale functionals using the $O(N)$ estimation algorithm described in Section IV-B.

*Example 9:* An example of the construction and exploitation of MR models in which specific nonlocal variables are included in the states at coarser scale nodes is the groundwater hydrology problem examined in [84] and introduced in Section II-F. In particular, in this problem, the random field modeled by the MR tree represents the log-hydraulic conductivity field of a region of interest. The measurements of log-conductivity represent point, i.e., finest scale, measurements of this process at a scattered set of fine-scale nodes corresponding to the locations of the well measurements. However, as we discussed in Section II-F, hydraulic head, which is also measured at each well location, is a strongly nonlocal and also nonlinear function of log conductivity. The approach taken in [84] is to linearize this nonlinear relationship about a known background log-conductivity value over the region of interest, resulting in a linearized model for hydraulic head at each well location, namely as a weighted integral of log conductivity. Using the method developed in [82] and just described, an MR model was constructed in which each of these nonlocal functionals was included in the state at individual nodes of the tree.

As discussed in Section II-F, the real objective of the application considered in [84] is the estimation of the travel time for solute particles to migrate from one spatial point to another. This quantity is also a complex, nonlinear, and nonlocal functional of log conductivity, and in [84] two alternate methods are described for its estimation. The first is analogous to the method used to incorporate head measurements: we linearize the relationship between travel time and log-conductivity, producing a model for travel time (or, more precisely, the perturbation in travel time from its nominal value as computed using the assumed background conductivity field) as a weighted linear functional of log-conductivity. This linear functional can then be augmented to the MR model in the same manner as the linearized head measurements, so that the resulting MR estimation algorithm automatically estimates this travel time perturbation as well.

As discussed in [84] and as illustrated in Fig. 20, this approach works as long as the perturbations from the background conductivity are not too large. If this is not the case, an alternate method can be used that emphasizes another feature of the MR formalism. In particular, suppose we do not augment the MR model with a linearized model of travel time but simply use the available log-conductivity and head measurements to estimate the log-conductivity field. As we discussed in Section IV-B, the result of this estimation process is not just a best estimate of that log-conductivity field but also an MR model for the errors in that estimate. This model can then be used to perform conditional simulations, which, as pointed out in Section II-F, is a well-known concept in geo-

[54] Fractional Brownian motion processes are characterized by the so-called Hurst parameter $H$, which controls the rate of spectral fall off. In particular, while fBm is nonstationary, its power spectral density is well defined over any band of frequencies [346] and falls off as $l/f^{2H+1}$.



(a)

(b)

**Fig. 20.** Illustration of the effectiveness of the MR algorithm in estimating travel time perturbations. In each figure, the dashed line represents the distribution predicted by the direct estimation of travel time perturbation modeled as a linearized functional of log conductivity, which was directly incorporated into the MR model as a state at a coarse-scale node, i.e., these dashed Gaussians have means and variances corresponding to the resulting estimate and error variance computed by the MR estimation algorithm. The histogram in each figure depicts the result of conditional simulation, in which we use the estimates of fine-scale conductivity and the MR model of the errors in these estimates to draw sample conductivity fields which are then used to drive the hydrologic equations, yielding sample values for travel time. The figure in (a) corresponds to the case in which the log-conductivity perturbation from the background value is comparatively small, while it is an order of magnitude larger in (b). (Reprinted from [84].)

physics [171]–[173].[55] In particular, we can generate samples from this MR error model, add them to the best estimate, and use the resulting log-conductivity field to solve the hydrology equation (4), which in turn yields the velocity field (5), which can then be used to compute the travel time for this particular log-conductivity field. By repeating this conditional simulation process many times, we can estimate the probability distribution for travel times. The key here is that, thanks to the tree structure of our MR models, we can construct such conditional realizations of the log-conductivity field very efficiently.

*3) Linear MR Models and Wavelet Representations:* We now return to the topic of MR models and wavelets. As mentioned in Section V-A, while wavelet synthesis can readily be viewed as a dynamic recursion in scale (44), more is needed to build MR models on trees based on wavelets. Consider first the simple case of the Haar transform and suppose that we wish to represent a stochastic process $z[n]$, $N_0 \geq n \geq N_1$, at the finest scale of a dyadic tree as in Fig. 18. In this case, the use of the Haar transform might first suggest that the states at each coarser scale node should be the corresponding (normalized or unnormalized) Haar scaling coefficient, e.g., referring to Fig. 18, we might consider taking

$$\begin{aligned} x(s\alpha_1) &= \frac{1}{2}\left(z[N_0] + z[N_0+1]\right) \\ x(s\alpha_2) &= \frac{1}{2}\left(z[N_0+2] + z[N_0+3]\right) \\ x(s) &= \frac{1}{4}\left(z[N_0] + z[N_0+1] + z[N_0+2] + z[N_0+3]\right). \end{aligned} \tag{53}$$

However, note in this case that the dynamics of the Haar wavelet synthesis imply that

$$\begin{aligned} x(s\alpha_1) &= x(s) + w(s\alpha_1) \\ x(s\alpha_2) &= x(s) + w(s\alpha_2) \end{aligned} \tag{54}$$

where

$$\begin{aligned} w(s\alpha_1) = -w(s\alpha_2) = \tfrac{1}{4}\,(z[N_0] + z[N_0+1] \\ -z[N_0+2] - z[N_0+3]). \end{aligned} \tag{55}$$

Thus, not only are $w(s\alpha_1)$ and $w(s\alpha_2)$ not independent, they are in fact deterministically related.

One solution to this would simply be to *assume* that $w(sa_1)$ and $w(sa_2)$ are independent, corresponding to the noninternal model in (7) and (47), in which case the variables in the MR model no longer correspond to the scaling coefficients of the finest scale process (e.g., $x(s)$ will *not* be the deterministic average of its children). Alternatively, if we do wish to maintain internality *and* the interpretation of states as components of the wavelet representation of the finest scale process, detail coefficients, such as in (55), must also be included as components of the state at each node (except for the finest scale at which we only need the individual signal value). For example, the 2-D state at node $s$ in Fig. 18 consists of the scaling coefficient $x(s)$ in (54) and the wavelet coefficient in (55). Note that, with this definition of the state, the coarse-to-fine dynamics of our MR model is partially deterministic. For example, as (54) and (55) show, the first components (i.e., the scaling coefficients) of the states at the two children of node $s$ are deterministic functions (a simple sum and a simple difference) of the two components of the state at node $s$. The other components of the states at these two nodes are then the new wavelet coefficients that will then be used in the next step of the coarse-to-fine synthesis. If the Haar transform did indeed exactly whiten the process $z[n]$, the coarse-to-fine dynamics would simply insert a white noise value for this detail coefficient. However, the structure of our MR model allows us to do better than this if the Haar transform does not perfectly whiten the process. In particular, using the same procedure for constructing the

[55] We refer the reader to [198] for another example of the use of MR models for conditional simulation in geophysics.

$A(s)$ and $Q(s)$ matrices as in Section VI-B2, we can define dynamics that take advantage of the residual correlation in the wavelet coefficients to do the best job of predicting the finer scale wavelet coefficient from its parent scaling and wavelet coefficients. This idea has been used, for example, in [83] to obtain better approximations to fBm than methods that completely neglect this residual correlation.[56]

The Haar case is a particularly simple and obvious one in its connection to MR tree models, thanks to the nonoverlapping support of the shifted and scaled Haar scaling functions and wavelets that comprise a dyadically scaled orthogonal basis. For example, the scaling coefficients in (54) involve only values of the fine-scale process $z[n]$ at points in the subtree below the corresponding nodes. However, if we consider more complex orthogonal or biorthogonal wavelets, e.g., ones with additional vanishing moments [86], [228], [329] and thus with higher degrees of smoothness, the situation appears to be much more complicated. In particular, in this case, the synthesis of each signal value, e.g., $z[N_0+1]$ in Fig. 18, requires contributions from *all* of the wavelet and scaling functions whose support includes the point $N_0 + 1$. Thus, achieving a form for the coarse-to-fine wavelet synthesis "dynamics" that has the structure of (6) requires that *all* of these required scaling and wavelet coefficients be part of the state $s\alpha_1$.

As discussed in [85], this state augmentation can indeed be done, but it is only half the story, since if only this condition were used to define the state at each node the resulting model would not be internal. The implications of this lack of internality in this case are severe. In particular, because of the need to augment the state at each node, individual scaling and detail coefficients appear at multiple nodes. Suppose, for example, that a particular coefficient appears in the state at two nodes $s$ and $t$. The junction tree constraint (or, more precisely, its counterpart for linear models) would then require that this coefficient also appears in (or, more precisely, be a deterministic linear function of) the state at each of the nodes on the path between $s$ and $t$. However, the augmentation done simply to make sure that everything needed for wavelet synthesis at a particular node is included in the state at that node does not satisfy this condition. As a result of this violation of the junction tree constraint, the multiple replicas of what is supposed to be a single coefficient need not (and typically with probability 1 *will* not) be equal.

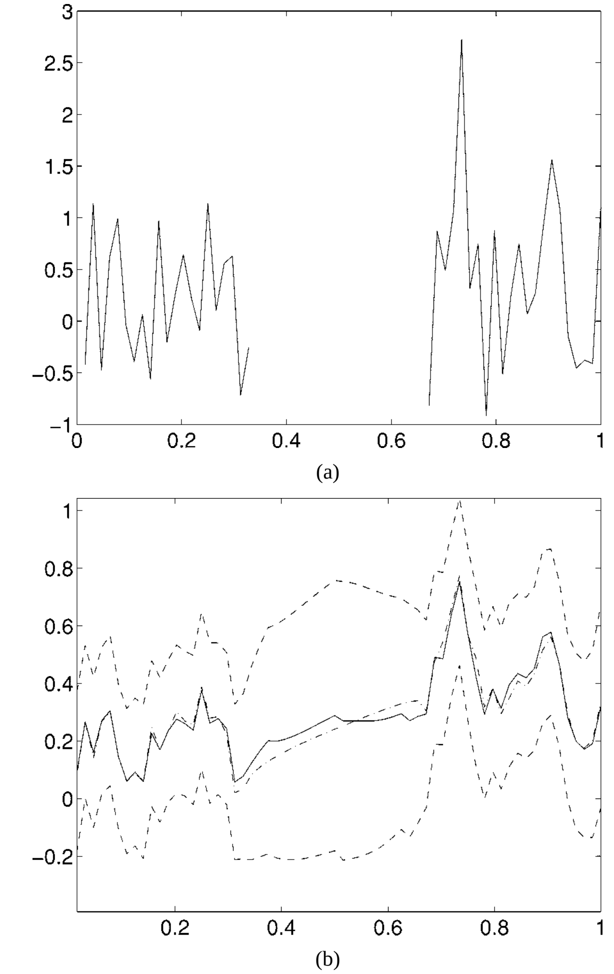
The key to see how to overcome this problem and recover internality is the examination of the fine-to-coarse wavelet *analysis* dynamics, in which wavelet and scaling coefficients at successively coarser resolutions are constructed as linear combinations of scaling coefficients at the previous finer resolution. Again, because of the overlapping supports of the wavelets and scaling coefficients at each scale, each of these coarser scaling coefficients, say one located at node $s$ in Fig. 18, is a linear combination of a number of the finer scale scaling coefficients. If the model we construct is to be internal, we must then have that all of these required coefficients must be resident at the two children of node $s$. Accomplishing this requires some additional state augmentation, which at first blush might lead one to believe that it would then be necessary to revisit the synthesis side of the problem in order to guarantee that we have everything we need at each node for consistent coarse-to-fine dynamics as in (6), and then to revisit the analysis side, etc. However, as shown in [85], this is not necessary, as the combination of a preliminary state definition for the synthesis dynamics followed by a second augmentation to ensure internality leads to a well-defined internal, linear, MR model in which the state at each node consists, with probability 1, of a vector of scaling and detail coefficients of the finest scale process.

Not surprisingly, the dimension of the state of the resulting model grows linearly with the support of the wavelet (or wavelets in the biorthogonal case), as typically does the degree to which such a wavelet transform whitens a given process such as fBm. In pure wavelet analysis, this often suggests the use of fairly high-order wavelets so that the residual correlation can be safely neglected. However, using our coarse-to-fine dynamics, we do not *need* to neglect the residual correlation and can in fact *exploit* it to enhance the fidelity of the resulting model. The result of this is that high-fidelity approximations of processes such as fBm can be created using MR representations based on wavelets of much smaller support than are typically used otherwise.

While the benefit described in the preceding paragraph is interesting and while the rapprochement with wavelets is intellectually satisfying, neither of these by themselves make a compelling case for why one would want to use such a representation. However, one good reason given in [85] is that, once we have this model, we can consider estimation of a process based on sparse, irregular, and even multiresolution measurements, i.e., to problems in which the data themselves are so erratically distributed that direct application of wavelet analysis is not possible. Fig. 21 illustrates the use of an internal MR tree model based on the Daubechies six-tap orthogonal wavelet [86] to estimate an fBm process given "gappy" measurements of differing quality over the two ends of the interval over which the process is to be estimated. Note that the resulting estimate is nearly identical to the one based on the exact fBm statistics, with deviations only a tiny fraction of one standard deviation of the errors in the optimal estimates.

*4) Covariance Extensions, Maximum Entropy, and MR Models:* We now examine another important topic in statistical signal processing with strong ties to graph theory and to MR models. The problem is that of covariance extension. Specifically, in many applications, it is unreasonable to expect to be provided with the complete covariance of a random process or field—or to have data available from which such a complete specification could be estimated. For example, in dealing with large-scale remote sensing problems, the random fields of interest can have dimensionality in the millions, making not only the availability but even the

[56] Note also that exploiting this residual correlation in order to do the best job possible of predicting finer scale wavelet detail coefficients is similar in spirit to discussion in Section V-A and in particular to the model (46). The key difference is that the model in (46) allows the use of the entire vector of scaling coefficients at the preceding scale to be used to estimate each detail coefficient at the next scale. In the tree-based approach described here for the Haar transform, each detail coefficient is predicted based only on the state (single wavelet and detail coefficient) at its parent node.



(a)

(b)

**Fig. 21.** Illustrating the use of an internal MR tree model for fBm (with Hurst parameter $H = 0.3$) using the Daubechies six-tap orthogonal wavelet. (a) We consider the problem of estimating a sample path of fBm given noisy measurements of the process over subintervals at the two extreme ends of the overall interval of interest. (b) The estimation results using both this MR model (solid line) and also using the exact fBm statistics (dashed–dotted line), as well as ± one standard deviation bars (dashed line). (Reprinted from [85].)

storage of a full covariance matrix prohibitive. In such problems, what is more likely to be the case is that only a comparatively small part of the covariance matrix is specified, and what we then seek is a model that is consistent with that partial specification.

This idea of modeling from partial specifications is well known in the signal processing field. In particular, consider the construction of a stochastic model for a stationary time series that matches a partially specified correlation function consisting of the first few values of that correlation function. If this partial specification is valid (i.e., if it does indeed correspond to the first few values of a completely specified correlation function), then it is indeed possible to find models that match that partial specification, one of which, namely the *maximum entropy* extension, corresponds to an AR signal model whose coefficients can be efficiently calculated from the specified portion of the correlation function, using, for example, the celebrated Levinson recursions (see, for example, [307]). Further, although not typically emphasized, the resulting AR model can then be used for the efficient, recursive computation of the covariance values not specified originally.

Interestingly, that maximum entropy extension also has important implications for MR modeling. In particular, the resulting AR model is a $k$th-order Markov process, where $k$ is the number of covariance values that were originally specified. As a result, using a construction analogous to that for Brownian motion in Example 2, we can, in principle, construct an MR model for this process analogous to that shown in Fig. 6, except that the number of boundary points kept at each end of each subinterval would be $k$ rather than one. That construction, however, requires knowledge of what appears to be a considerable number of covariance values other than those that were originally specified. For example, referring to Fig. 6, for a first-order Markov process, specifying the coarse-to-fine dynamics requires knowledge of the covariance between points (namely end- and mid-points) that are not near each other. In principle, these can be computed from the AR model as mentioned in the preceding paragraph, but that recursive method also calculates many elements of the covariance sequence that are not needed to construct the MR model. This raises the question, then, of whether the needed elements can be computed much more directly and efficiently, and that, in turn, leads to some important ties to graph theory.

In particular, suppose that $P$ is a *partially specified covariance matrix*, i.e., only certain elements of $P$, always including its diagonal, are specified. Furthermore, we must also have that $P$ is *valid*, namely, that any completely filled principal submatrix (i.e., any submatrix consisting of the same choices of rows and columns of $P$ and for which all of the elements are specified) is positive definite. An *extension* of $P$, then, corresponds to filling in some of the unspecified values in $P$ while still maintaining validity, while a *completion* is an extension in which every element is specified, yielding a full, positive-definite covariance matrix.

Two important questions are as follows.

1) Does a given partially specified covariance matrix have extensions and completions?
2) If so, what is the maximum entropy extension?

Answers to both of these questions have important graph-theoretic interpretations. In particular, if $P$ is an $N \times N$ matrix, consider the undirected graph with nodes labeled $1, 2, \ldots, N$, where we include the edge $(i, j)$ between distinct nodes $i$ and $j$ if the $ij$th element of $P$ has been specified. Then, the following results hold [21], [75], [142].

1) Given a particular graph of this type, extensions and completions exist for any valid partially specified covariance with this graph structure if and only if the graph is chordal.[57]
2) If a completion exists for a given partially specified covariance, then the maximum entropy extension is Markov with respect to the graph determined by $P$.

One typical example in which 1) holds is if a consecutive set of diagonal bands of $P$ are specified—this is simply the generalization of the usual AR modeling framework to allow

[57] If the graph is not chordal, the existence of extensions and completions depends on the specific numerical values of the specified elements of $P$.

for the time series to be nonstationary (so that each of the diagonals within the band need not have constant values). In this case, the resulting maximum entropy model is, as we have said, $k$th-order Markov, which is identical to the process being Markov with respect to the graph determined by $P$.

Suppose that the graph $\mathcal{G}$ of $P$ is chordal. The question of then calculating particular elements of $P$ or, more precisely, the possible recursive orders in which these elements can be calculated also has a graph-theoretic interpretation, as shown in [121]. Specifically, let $P_e$ be an extension of $P$ (so that $P_e$ agrees with $P$ wherever $P$ is defined), and let $\mathcal{G}_e$ be the graph corresponding to $P_e$ (so that $\mathcal{G} \subset \mathcal{G}_e$). Then, if $\mathcal{G}_e$ is also chordal, we can calculate the additional elements of $P_e$ without having to calculate any other elements beyond those in $P_e$. Moreover, this can be accomplished recursively by constructing a *chordal sequence*

$$\mathcal{G} \subset \mathcal{G}_1 \subset \mathcal{G}_2 \subset \cdots \subset \mathcal{G}_e \tag{56}$$

where each step in this sequence corresponds to adding a single edge to the preceding graph. This sequence then provides a recursive ordering for the computation of the required elements of $P_e$.

In [121], it is also shown that each step of this recursion defines a range of values for the new element of the extension, providing, in essence, a complete characterization of all possible extensions much as reflection coefficients do for standard time series models [307]. In particular, it is shown in [121] that the required computations for each of the steps in the sequence involves a submatrix corresponding to the new maximal clique formed by the addition of the new edge. This submatrix has a single new element to be computed. Choosing any value that makes the submatrix positive definite is valid, and choosing the particular value that maximizes the determinant of this submatrix corresponds to the maximum entropy extension. In some cases, the maximal clique size can grow as the recursion progresses, apparently implying that the required computations also grow. However, we refer the reader to [121] for an additional set of graph-theoretic conditions on the chordal sequence in (56) that in essence guarantee that most of the computations required at each stage of the recursion have already been performed at previous stages. This result provides a nontrivial extension of Levinson-like recursions.

Finally, let us examine the specific extension required to form the MR tree model when we begin with a partial covariance consisting of $k$ diagonal bands on either side of the main diagonal. In this case, as described previously, the additional elements of the maximum entropy extension that must be computed are unusually distributed (and, in fact have a fractal pattern—see [121]). Moreover, they are an extremely sparse subset of the elements of $P$ (having $O(N)$ elements). Surprisingly, however, the graph corresponding to this extension is chordal. Moreover, when the corresponding chordal sequence is constructed to compute these needed elements, we find that the resulting sequence of new maximal cliques remains bounded in size, so that the total computational load to construct the resulting MR model is also $O(N)$. We refer the reader to [121] for details.

### *C. Estimation of Model Parameters and Learning of MR Models*

In this section, we take a brief look at the problem of estimating or learning MR models from data. There are three separate classes of problems we describe, as detailed in the following sections.

*1) Estimation of MR Model Parameters:* The first class of problems involves the estimation of parameters of MR models of fixed and known structure.[58] As discussed in Section IV-C, the computation of likelihood functions can be performed efficiently for MR models, implying, for example, that one can use these computations as the basis for ML parameter estimation. Examples of this for linear models can be found in [113] and [114] for both the estimation of the Hurst parameter of fBm and for the estimation of noise correlation structure and parameters for models used in oceanographic remote sensing.

For discrete or hybrid MR models, e.g., as in the MR segmentation model in [42] or the hidden Markov tree models illustrated in Example 6 and developed in detail in [59], [80], [261], [281], and [282], a very effective approach to parameter estimation involves the use of the EM algorithm [97]. The employment of EM requires the specification of the so-called *complete data*, which includes not only the actual measured data but also some additional hidden variables, which, if available, make the computation of parameter estimates much easier. For example, as described in [80], for hidden Markov tree models, the clear choice for the complete data consists not only of the actually observed wavelet coefficients but also the hidden discrete states at each node, where for the following discussion we let $\mathbf{w}$ and $\mathbf{s}$ denote the vectors of all of the wavelet coefficients and discrete state variables, respectively. We also denote by $\theta$ the vector of parameters to be estimated, consisting of the probabilities defining the discrete-state hidden Markov model and the means and variances for each wavelet coefficient conditioned on each of the possible values of the corresponding discrete state. The E or expectation step then consists of computing the conditional expectation of the log-likelihood function $\log[p(\mathbf{w}, \mathbf{s}|\theta)]$ conditioned on both $\mathbf{w}$ and on the previous iteration's estimate of $\theta$. This corresponds to averaging over the possible values of $\mathbf{s}$ given the conditioning information. The maximization (or M step) then involves maximizing the function computed during the E step in order to compute the next iteration's estimate of $\theta$. Note that the vector $\mathbf{s}$, consisting of choices for all of the discrete states at every node in the tree, has a number of possible values that is exponential in the number of nodes in the tree. Thus, as discussed in [222] and also in Section VI-A, the computation of such expectations for general graphical models can be prohibitively complex, while for trees such computations can be performed extremely efficiently. As a result, the implementation of EM-based algorithms for parameter estimation for MR models on trees

[58] In many cases, the quantities to be estimated are often referred to as *hyperparameters*, typically a small number of unknowns, where the actual parameters of the model, e.g., the elements of the matrices $A(s)$ and $Q(s)$ in linear MR models, are (typically nonlinear) functions of these unknowns.

are computationally attractive. We refer the reader to [42] and [80] for further discussion and examples.

*2) Learning MR Models:* An alternative to the parametric estimation methods discussed in the preceding paragraph is the use of a nonparametric or machine learning philosophy in building such models. We briefly describe two lines of investigation, in each of which wavelet transforms are used to populate the variables in such an MR model and then MR models are learned from training data.

The first of these, described in [124] and [125], involves the construction of nonlinear coarse-to-fine statistical models for wavelet transforms, i.e., models much as in (45) and (46) except that the estimate of each wavelet coefficient is allowed to be a nonlinear function of a window of nearby scaling coefficients. Specifically, the wavelet coefficients at each scale are modeled as being independent when each is conditioned on its own local window of scaling coefficients, and the conditional distribution for each coefficient is assumed to be Gaussian. However, the mean (variance) of that conditional distribution is modeled as a piecewise affine (constant) function of the window of scaling coefficients. For any such "piece," we have a linear parametric model as in (45). However, determining how many such linear pieces there should be and specifying the region over which each linear function should be applied require nonparametric estimation techniques. We refer the reader to [124] and [125] for details and also for the application of these models to problems of tomographic reconstruction. Finally, we note that, since each wavelet coefficient depends on several scaling coefficients, the question arises as to whether the resulting MR model forms a tree or, more precisely, if state augmentation methods such as described in Section VI-B3 can be applied to transform this model into a tree model. In some cases, this will certainly be the case. However, in others, the result will be a more complex graph that does not yield a junction tree or cutset tree model with acceptably small state dimension.

Another very interesting approach to MR modeling for image processing is that developed in [90]–[92]. The basic idea behind this approach is quite simple. Given a sample image, we form an MR pyramid by performing an MR decomposition of the image—the specific decomposition used in [91] is an overcomplete steerable pyramid [298]. Thus, at each node $s$, on a quadtree, we have a vector of coefficients, denoted by $z(s)$, sensitive to variations in different directions at the location and scale corresponding to node $s$. From this one image sample—or perhaps from a small set of images [20]—we wish to learn non-Gaussian, nonlinear, coarse-to-fine statistical dynamics. In particular, what is done in [20] and [90]–[92] is to use nonparametric density estimation methods to estimate both the distribution of the vector of values at the root node and the conditional distributions for every other node given all of its direct ancestors. That is, for each node $s$ other than the root node 0, we estimate the density

$$p\left(z(s) \,|\, z(s\bar{\gamma}),\, z\left(s\bar{\gamma}^2\right),\, \ldots,\, z(0)\right). \qquad (57)$$

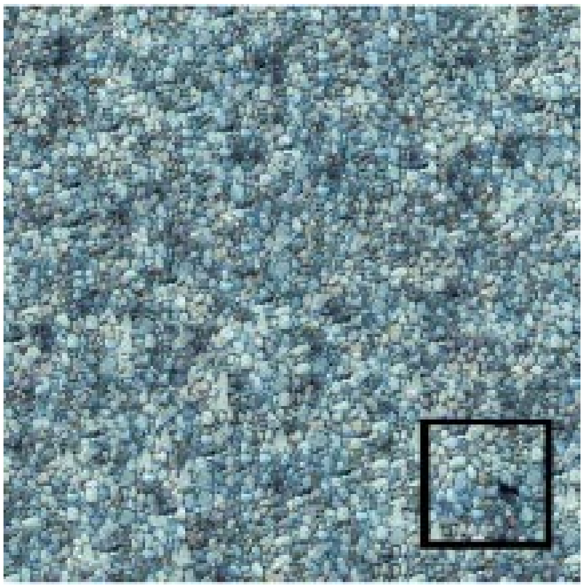


**Fig. 22.** An example of the method developed in [91] for nonparametric estimation of MR models of steerable wavelet pyramids of a sample image. The small region included within the black border in the image on the right is a real image, from which an MR model was learned. The entire image shown is entirely synthetic, using this learned model, except for this one small region in which the real image is located. (Based on [90] and [91].)

The way in which this is done is to assume an interesting variation of stationarity, namely, that each node $s$ at a given scale has the same conditional distribution (57) as all other nodes at that scale. What this implies is that, even with a single image from which to learn the distribution, we will have a significant number of samples from which to estimate these densities at finer scales (at which there are many nodes), although the resulting learned densities at coarser scales will be less certain. Note also that the distributions in (57) do not correspond to a Markov model on the tree for $z(s)$, since, if that were the case, conditioning $z(s)$ on $z(s\bar{\gamma})$ would make further conditioning on more distant ancestors of no informational value. However, just as in time series analysis, it is simple to turn such a higher order Markov description into a first-order representation by state augmentation, i.e., by defining the state at each node to consist of a vector of the values of $z(.)$ at the ancestors of that node. As with hybrid and nonlinear wavelet-based methods, such as in [59], [80], [261], [281], [282], and [333], the motivation for this modeling methodology is to capture both the non-Gaussian nature of wavelet statistics and the cascade behavior characteristic of wavelet decompositions of natural images. Fig. 22 illustrates one example, suggesting the promise of this approach for the modeling of natural imagery.

*3) Learning MR Tree Structure:* Throughout the discussions in this entire section, we have focused on aspects of the MR modeling problem other than identifying or learning the structure of the MR tree; that is, we have focused on identifying the variables to place at particular nodes on a prespecified tree and/or the problem of determining the parameters of a model once those variables have been specified. A strong argument can be made that this is reasonable for signal and image processing applications, since the nodes and variables at these MR nodes have at least rough intuition associated with them related to the representation of phenomena at different scales and spatial locations. Nevertheless, it is worth noting that the topic of identifying the structure of the tree has received some attention [64], [177], [230], [238], [239], [303], mostly in fields other than signal and image processing.

Perhaps the best known work in this area is that of Chow and Liu [64]. The idea in this work is that we are given an index set $\mathcal{V}$, a set of random variables $\{z(s)|s \in \mathcal{V}\}$, and a number of independent realizations of this set of variables. We assume that the joint distribution of these variables is given by a tree distribution, i.e., that the $z(s)$'s form a graphical model with respect to a tree with node set $\mathcal{V}$. However, we neither know which of the many trees with this index set is the correct one nor the distribution for these variables. The objective, then, is to use the available measurements to determine the ML estimate of both the tree and the distribution with respect to that tree. Since for any a specific choice of tree structure the ML estimate of the distribution over that tree is simply the empirical factored distribution based on the observed data, the central problem reduces to identifying the best choice of tree structure. As shown in [64], this problem can be solved very efficiently.

The special nature of trees is reinforced by the observation that the solution to this problem in which we allow graphs other than trees is much more difficult and, in fact, is NP-Hard [303]. An important recent advance in this area is the work reported in [177] and [303] which focuses on chordal graphs with bounded tree width (so that maximal clique sizes are bounded). Even for this limited set of graphs optimal identification is prohibitively complex, but the results in [177] and [303] show that it is possible to develop computationally feasible algorithms that have ranks (with respect to maximizing the likelihood) that are provably bounded relative to the optimal. The significance of these results for MR modeling have yet to be developed, but some additional motivation for considering graphs that are in some sense "close" to trees is given in the next section.

## VII. Moving Beyond Trees

As the preceding sections make clear, MR models on trees have many attractive properties that lead to powerful and efficient signal and image processing algorithms that have extensive domains of application. The fact that these models are Markov on *trees*, i.e., graphs without loops, leads both to the power of these algorithms and also to the apparent limitations on their applicability. In particular, as we discussed in Section VI, while it is always the case that any process (e.g., any MRF or graphical process on a loopy graph) can be modeled (exactly or approximately) using an MR model on a tree, in many cases the resulting dimensionality or cardinality of the state of that tree model is large. Since the complexity of the algorithms we have described grows polynomially with state dimension or cardinality, we have three alternatives: 1) reduce the dimension or cardinality of the state; 2) develop alternative algorithms and approximations that reduce complexity but allow us to keep higher dimensional or higher cardinality states; and 3) consider MR models on graphs with loops.

In the preceding sections, we described a variety of approaches to the first of these alternatives. In some cases (e.g., with the smoothness-based models described in Section VI-B1), reduction of state size can be accomplished by replacing one model with another that serves essentially the same purpose. In others (e.g., with the approximate stochastic realization methods in Section VI-B2) the method used is to reduce state dimensionality by keeping a limited-dimensional projection of the full state at each node, chosen to minimize the residual correlation between variables that the full state completely decorrelated. Alternatively, as discussed at the end of Section VI-B1, one could use the method based on overlapping trees to overcome the severe burden, especially at coarser scales, that is placed on the state of a tree model, namely providing complete conditional mutual independence to the sets of variables in the separate subtrees descendent from that node. However, as we also indicated, one of the prices for using an overlapping tree is that the total number of tree nodes $N$ increases by a factor of two for a dyadic tree and by a factor of four for a quadtree with every additional scale. Thus, the amount of overlap that can be accommodated while maintaining computational efficiency of the resulting inference algorithms also has limits.

As a result, there is considerable motivation to consider the other two alternatives mentioned previously. In the next section, we take a brief look at an approach that keeps state size large but reduces the computational burden of the resulting inference algorithms, and in this context we also make contact with the very important field of space–time processes and algorithms and its counterparts, e.g., dynamic Bayes nets (DBNs), in the graphical model literature. In Section VII-B, we then take a brief look at several recently introduced methods that have been developed to deal with estimation on graphs with loops. While these methods were originally motivated by inference problems for MR pyramidal structures, they can be applied to arbitrary graphical models and thus are of independent interest to the graphical model community.

### *A. Reduced-Complexity Cutset Models and Time-Recursive Approximate Modeling*

In this section, we return to the cutset models discussed in Section VI-A1, and while the ideas we describe can (and have) been applied to more general graphs and to nonlinear models, for illustrative purposes we frame our discussion in terms of a linear-Gaussian, nearest-neighbor MRF on a regular 2-D lattice as in Fig. 14. As we discussed in Section VI-A1, an exact MR model can be formed by taking as the state at the root node either the full set of variables along either the red row or blue column of the grid in Fig. 14 (leading to a dyadic tree structure in which, for example, we alternatively bisect regions horizontally and vertically) or where we take the root node state to be the set of variables along *both* the red row and blue column in the figure (leading to a quadtree model). In either case, the dimension of the state at the root node is $O(N^{1/2})$, where $N$ is the number of nodes in the 2-D grid. Note that the dimensions of nodes at successively finer scales are only half the dimension of their parents, so that the overall complexity of inference algorithms is not as bad as if every state had dimension $O(N^{1/2})$, but it is still the case that large state dimension at even a small number of nodes leads to problems for procedures such as the estimation algorithm described in Section IV-B.

In particular, the dynamic model (6) and the estimation algorithm given by (20)–(37) are written in a form that requires the explicit representation, computation, and storage of various estimates and error covariances of the state at each node in the tree [e.g., see (20), (24)–(26), (35), and (36)], and in general each of these covariance matrices will be full. The key to the method described in this section is an alternative form for estimation equations known as the *information filter*, in which the quantities that are stored and computed directly are *information matrices*, $P^{-1}$, i.e., inverses of covariances, and *information states*, i.e., $P^{-1}\hat{x}$, where the covariances and estimates here could be any of the pairs appearing in the algorithm in (20)–(37). As is well known for time series [174], the information filter algorithm has a form analogous to (20)–(37), in which it is these information quantities that are recursively computed.

At first glance, this approach seems to have bought us nothing, since, in general, there is no guarantee that the computations involved in this alternate form will be any less demanding than (20)–(37), and we now have additional computations to perform, namely, to recover $\hat{x}$ from $P^{-1}\hat{x}$. However, as pointed out in Section III-C, the inverse of a covariance matrix can be interpreted as specifying a model for a random vector or process, and this is the key to an alternate form of approximation for cutset models to that considered in Section VI-B2. In particular, consider the set of variables corresponding to the values of a first-order Gaussian MRF along the center, red row of Fig. 14. What we would like to do is to think of this set of values as a 1-D signal. However, while the inverse of the entire covariance of the full 2-D process is sparse (thanks to the graphical structure of the model), the inverse of the covariance matrix of this center row alone is generally full, implying that the graphical model associated with this 1-D signal is fully connected. On the other hand, as discussed and illustrated in [82], [166], and [316], in many cases, e.g., in particular for many first-order MRFs, these information matrices are nearly banded, i.e., the dominant nonzero values are in a relatively narrow diagonal band around the main diagonal of the matrix. Consequently, if we were to approximate such matrices by setting to zero the values that are deemed to be small,[59] we obtain an approximation for the inverse covariance of this 1-D signal that is banded. This corresponds to pruning edges from the fully connected graphical model to produce an approximate 1-D Markov model of order equal to the width of the nonzero diagonal band.

This suggests the structure of an MR modeling algorithm—and of the corresponding estimation algorithm as well—in which we recursively compute banded approximations to information matrices and information states for each of the cutset states in a tree model on the lattice of Fig. 14. Since the approximate information matrices have small numbers of nonzero elements, the computations involved in each step are much simpler (e.g., only linear or at worst quadratic in state dimension). As a result, this approximate algorithm has total complexity that is at worst $O(N)$ [166]. Moreover, the recovery at each node of $\hat{x}$ from $P^{-1}\hat{x}$ is also straightforward—and, in fact, corresponds precisely to a 1-D Kalman filter/Rauch–Tung–Striebel smoother—thanks to the banded structure of our approximation to $P^{-1}$. We refer the reader to [82], [166], and [316] for examples and details. In addition, these same principles can be applied to discrete-state and nonlinear graphical models, although the criterion for pruning edges must obviously take a form other than examination of elements of the inverse of a covariance matrix (see, e.g., the references in the following discussion of DBNs).

The method just described has close relationships both to well-known methods for the numerical solution of PDEs [106], [138] and to algorithms and ideas for space–time processes and DBNs. In particular, as described in [81] (see also [181]), suppose that, instead of beginning with the middle red row in Fig. 14, we begin either with the top row or the leftmost column and then "march" either downward row by row or from left to right column by column, where in either case we propagate an approximate version of the inverse covariance, i.e., an approximate 1-D Markov model for the values of the field along each successive row or column. Note that such an approach effectively treats one of the two spatial dimensions as a time-like variable for the row-to-row or column-to-column recursion, and this in turn provides a direct connection to space–time processes, e.g., processes on grids such as in Fig. 14 in which one of the two independent variables *is* time.

The idea of propagating approximate graphical models in time is a topic of significant current interest [43], [134], [254], and we refer the readers to these references for details. We note in particular that in [43] the authors confront a problem of considerable concern not only for DBNs but for the approximate MR modeling methods described in this and previous sections, namely the issue of how approximation errors propagate and accumulate over time. In particular, these authors obtain results that show that, as long as the temporal dynamics of the process of interest have sufficient "mixing," the Kullback–Leibler (K–L) divergence between exact and approximate models decreases with temporal propagation, which implies that, if comparable approximation errors are made at each time step, the accumulation of these errors over time (as measured by K–L divergence) remains bounded. Developing comparable results (and possibly stronger ones in

[59] In particular, as discussed in [143] and [204], it is relatively easy, with local computations, to compute what is known as the partial correlation coefficient corresponding to a particular edge, namely the conditional correlation between the nodes connected by that edge conditioned on all the rest of the variables in the entire graph. Removing edges corresponding to small partial correlations is one approach taken in [166].

the linear case) for MR tree models remains an open topic whose resolution would provide a way in which to relate local approximations at each node in a tree with the impact on global model accuracy.

We also refer the reader to [54]–[56] for related research in space–time estimation in which the spatial phenomenon can be 2-D or 3-D, and for which the objective is to propagate an MRF model (e.g., a first-order or higher order MRF on a grid such as Fig. 14), and to [147] and [342] for approaches to direct temporal propagation of an MR tree model for a space–time process. An important issue that [147] begins to address is that of the temporal "mixing" of spatial scales, i.e., features at one scale at one point in time can interact through the temporal dynamics to produce features at different scales as time evolves. This characteristic implies that the statistical relationships, e.g., between a parent and child node, at one point in time depend on the relationships at the previous time among nodes that may be at several different scales. The resulting structure consists of a temporal sequence of MR models on trees with directed edges between nodes in the tree at one time to nodes in the tree at the next time in order to capture temporal dynamics and mixing.

### B. MR and Tree-Based Algorithms for Graphs with Loops

In this section, we describe approaches that relax the requirement that the MR model live on a loop-free graph, thereby reducing the decorrelation burden on (and hence the dimension of) coarse-scale nodes by allowing additional paths between finer scale nodes. For example, rather than using MR models on trees such as in Fig. 1(a), one might consider MR models on pyramidal structures such as in Fig. 23, in which there are edges between pairs of nodes at various scales, in essence providing a short circuit between nodes that would otherwise be far apart, as measured by distance along the MR tree.

Other examples of MR graphs with loops, e.g., ones in which each node is connected to several parent nodes (e.g., see [42], [110], and [156]) have been mentioned previously. With the use of any such graph, however, one must confront the problem of inference, which, as we discussed in Section IV-D, is a challenging problem that is the subject of considerable current interest. Indeed, the graphical model and turbocoding literature contain important results on the behavior of belief propagation algorithms (e.g., [123], [278], [332], and [337]), including results for linear-Gaussian models [285], [332], [338], as well as the introduction of new classes of iterative algorithms (e.g., [332], [334], and [357]). It is not our intention to describe or review this vast and active area of research. Rather, we refer the interested reader to the references just given and the others on graphical models cited previously and limit ourselves here to brief descriptions of several investigations that have been directly motivated by MR models and tree algorithms, beginning with the following example.

*Example 10:* In this example, we return to the problem of image segmentation introduced in Section II-E and, in particular, to an approach introduced in [42] and subsequently employed in a variety of other contexts (see, for example, [53], [183], [289], [290], and [323]). For simplicity, we describe the idea in the context developed in [42], which employs an MR discrete-state Potts model, as introduced in Example 3 and (8), to describe the coarse-to-fine dynamics of a hidden Markov tree representing segmentation labels at a sequence of resolutions. The observed data in [42] consists of image measurement at the finest scale which are assumed to be conditionally independent when each measurement pixel is conditioned on the value of the hidden discrete label at that pixel (where the form of the conditional distribution is taken to be Gaussian or a Gaussian mixture in the examples in [42] or a generalized Gaussian in other references (e.g., [289] and [290]).

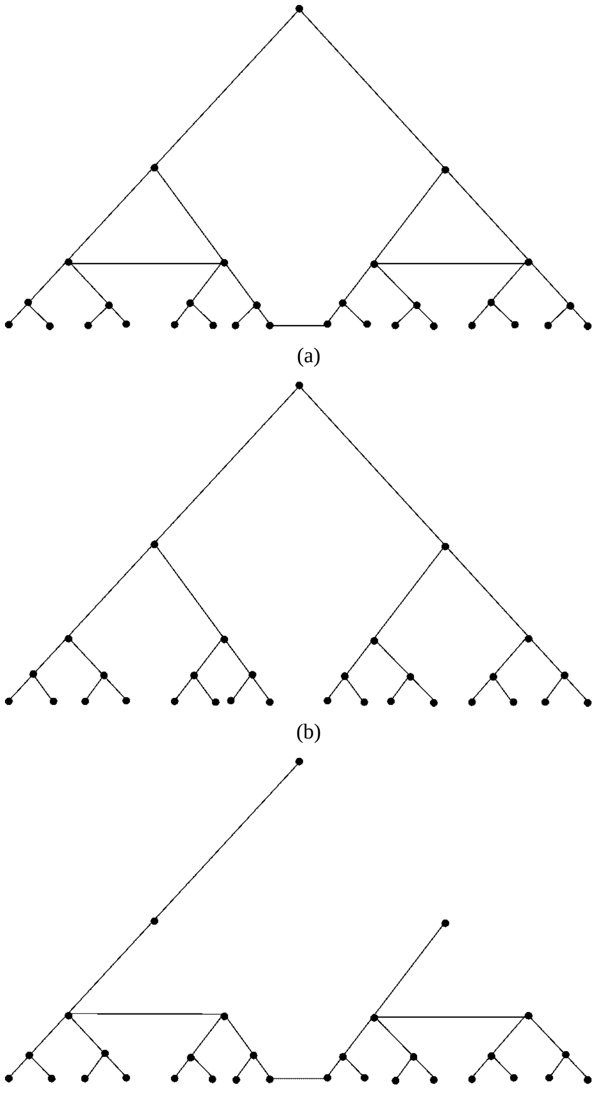


**Fig. 23.** (a) An example of an MR (loopy) graphical structure, including several direct connections across what are major boundaries of the underlying MR tree. (b) and (c) Two spanning trees of this graph are shown.

While the model described in the previous paragraph is essentially identical in structure to others we have described (e.g., see Example 6), there are two significant additional components of the complete approach developed in [42] that distinguish it and lead to inference algorithms with a very different structure. First, rather than using either an MAP or MPM criterion for optimal estimation of the discrete label set, the authors suggest an alternative measure aimed in part at overcoming the problematic use of MAP estimation for segmentation [234] and also at exploiting the structure of MR models. In particular, using the argument that errors at coarser scales are geometrically more expensive (since they correspond to misclassifications over geometrically larger regions), the authors derive a criterion that puts exponentially larger weights on errors that occur at coarser scales (see [42] for the precise formulation). The result is a criterion whose precise optimization is rather complicated. However, the authors demonstrate that a very good approximation to the criterion to be optimized at each scale results in a very simple and intuitively appealing structure: a first fine-to-coarse sweep, much as in the two-sweep algorithms we have described previously, is performed to compute at each node the conditional log-likelihood for the data in the subtree below that node conditioned on the discrete-state value at that node. That is, if we let $x(s)$ denote the discrete state at node $s$ and $Y_s$ denote the data at the finest scale nodes descendent from $s$, the coarse-to-fine sweep computes

$$\ell_s(k) = \log p(Y_s | x(s) = k) \tag{58}$$

at each node $s$ and for each value $k$ that $x(s)$ can take on. At the root node, we can then compute the optimal estimate as

$$\hat{x}(0) = \arg\max_k \ell_0(k). \tag{59}$$

The approximate estimation algorithm then proceeds in a coarse-to-fine manner where at each stage we essentially assume that the estimate at the parent node is correct. That is, the approximate coarse-to-fine recursion is given by

$$\hat{x}(s) = \arg\max_k \left\{\ell_s(k) + \log p(x(s) = k \,|\, \hat{x}(s\hat{\gamma}))\right\}. \tag{60}$$

However, as we have pointed out, and as is also pointed out in [42], use of such a tree model can lead to artifacts, and this leads to the introduction in [42] of an alternative to the tree-based Potts model of Example 3, namely, a directed graphical model from coarse-to-fine scales in which the discrete state $x(s)$ at any node s depends on the values at *three* nodes at the preceding, scale, namely, the parent node $s\bar{\gamma}$ and two of its neighbors at the same scale, where for simplicity here we denote this set of three nodes as $\mathcal{P}(s)$. The model used is a straightforward variation of (8) in which the state $x(s)$ is influenced by all three of its parents (see [42] for details). As a result of adopting this nontree model, the computation of the likelihood in (58) and in fact the entire structure of the computation of optimal estimates becomes much more complex thanks to the loopy structure of the graph corresponding to this three-parent model. In principle, maximization cannot be performed node by node at each scale; rather, all nodes at each scale need to be considered simultaneously, a combinatorially explosive requirement at finer scales.

Thus, as with any inference problem on a complex graph with loops, an approximation or iterative scheme is needed. The approach taken in [42] is to define a noniterative, two-pass algorithm that is motivated by the perspective of MR processing but that is decidedly different from approaches found in the general graphical model literature. In particular, [42] assumes that *both* a tree-based Potts model as in Example 3 *and* a three-parent nontree model are available for the discrete state process. The tree-based model is used on the fine-to-coarse sweep, computing the likelihoods in (58). The optimal root node estimate is then computed as in (59), and the estimates at other nodes are computed node by node in a coarse-to-fine sweep using the following recursion in place of (60):

$$\hat{x}(s) = \arg\max_k \left\{\ell_s(k) + \log p(x(s) = k | \hat{x}(t), t \in \mathcal{P}(s))\right\}. \tag{61}$$

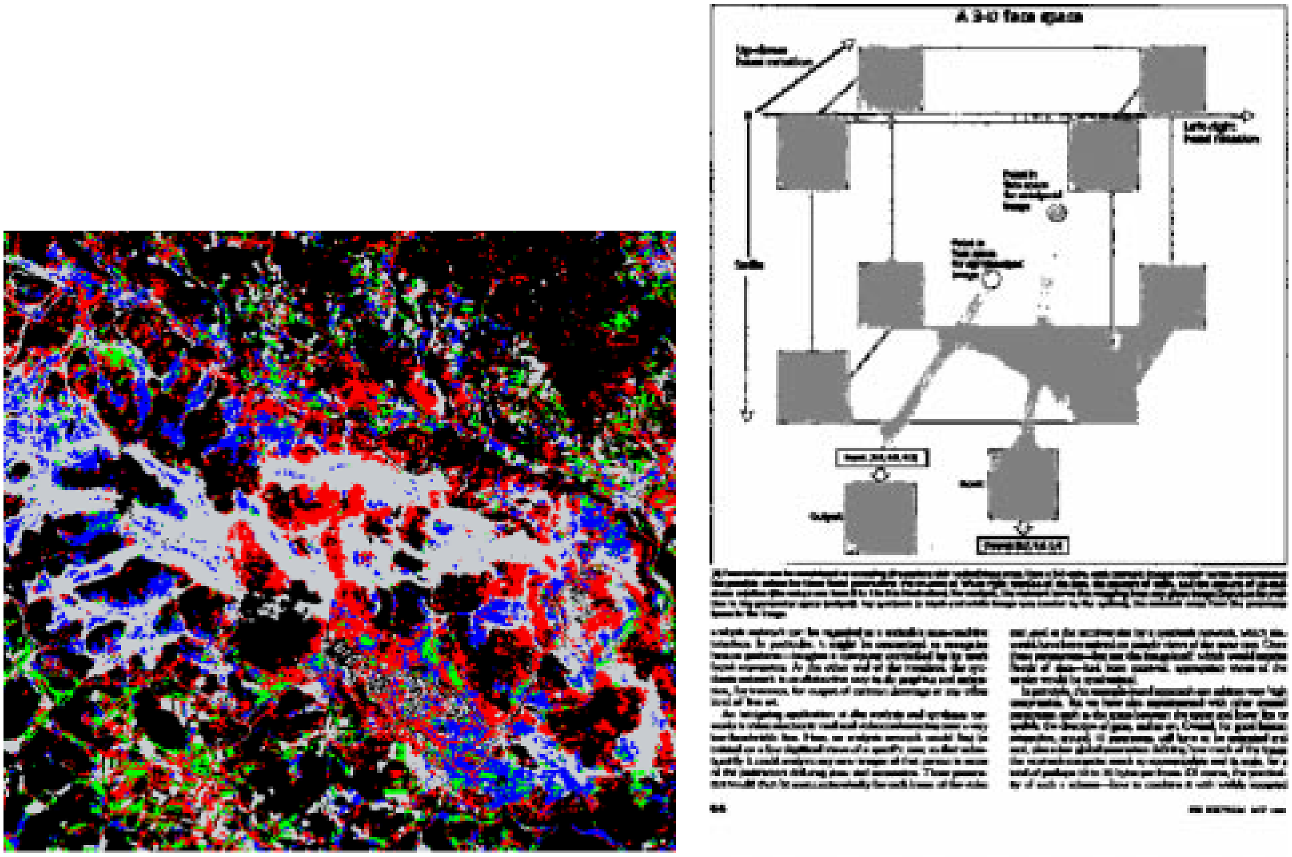
Comparing (60) with (61), we see that the only difference is the conditioning on the set of three parents rather than simply $\hat{x}(s\bar{\gamma})$ in the transition distribution. As discussed in [42], at each scale what this algorithm resembles (but does not equal) is the optimal estimate based on using a three-parent Potts model for coarser scales but a tree-based Potts model for finer scales.

Fig. 24(a) shows the result from [42] of applying this algorithm to the multispectral image in Fig. 5(a), in which the data at each pixel consists of a vector of multispectral measurements, and the image is segmented into five regions, the pixels in each of which are modeled as a multivariate Gaussian mixtures. Fig. 24(b) shows corresponding results for the segmentation of the document page in Fig. 5(b) into three regions (text, picture, and background). The algorithm used for this second example and developed in [53] uses the same structure (tree for fine-to-coarse likelihood computation and nontree for coarse-to-fine estimation) but much more sophisticated and involved data and model structures. We refer the reader to [53] for details.[60]

Finally, we briefly describe research motivated both by MR loopy graphs as in Fig. 23 and also by the efficient algorithms described in Section IV for inference on trees. Specifically, consider estimation for a graphical model on a (connected, loopy) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, based on noisy measurements of variables at some (or all) of the nodes on the graph. The basic idea behind the algorithms in [332] is to carry out this estimation process using inference on tree models as a basic engine. In particular, suppose that we identify a set of *spanning trees*, $\mathcal{S}_1, \ldots, \mathcal{S}_T$ for the graph $\mathcal{G}$. That is, each of these trees connects all of the nodes in $\mathcal{V}$ and has a set of edges that is a subset of the edge set $\mathcal{E}$ of $\mathcal{G}$. For example, Fig. 23 (b) and (c) depict two spanning trees for the graph in Fig. 23(a). The general structure of the algorithms in this class involve iterative application of tree-based inference using the statistical structure implied by using each of these trees individually.

Specifically, let $x$ be a zero-mean, Gaussian, graphical process on $\mathcal{G}$ with covariance matrix $P_x$, and suppose that

[60] The algorithm in [53] involves the use of Haar wavelets to transform the raw measurements into detail coefficients, an affine class-dependent MR model for the scale-to-scale dynamics for these coefficients, and a directed graph model for the multiscale class labels with transition probabilities that are in general much more complex than those used in [42]. All of these aspects of the model are learned by training on a small set of sample images.



(a) (b)

**Fig. 24.** Segmentation results on the images in Fig. 5: (a) from [42] using the technique described in Example 10 (each multispectral pixel is classified into one of five region types); (b) from [53] using a more sophisticated MR algorithm building on the framework described in Example 10 (here each pixel is classified into one of three classes: text, picture, and background).

we have linear measurements $y$ as in (38) with $C$ and the covariance $R$ of the measurement noise $v$ block-diagonal. As discussed in Section IV-D, the optimal estimate $\hat{x}_s$ is the solution of (39), and the corresponding error covariance $P_e$ is given by (40). Further, as we also discussed previously, $P_x^{-1}$ has a nonzero element in the off-diagonal $(s,\, t)$ block only if $(s,\, t) \in \mathcal{E}$. If $\mathcal{G}$ were a tree, this would allow us to apply the fast estimation algorithm of Section IV-B to calculate both $\hat{x}_s$ and the diagonal blocks of $P_e$. Since $\mathcal{G}$ is not a tree, we cannot do this; however, for each of the spanning trees $\mathcal{S}_1,\, \ldots,\, \mathcal{S}_T$, we can write

$$P_x^{-1} = P_i^{-1} - K_i, \qquad i = 1,\, \ldots,\, T \tag{62}$$

where the only nonzero off-diagonal blocks of $P_i^{-1}$ correspond to edges in the spanning tree $\mathcal{S}_i$, and $K_i$ has nonzero elements only in blocks corresponding to edges eliminated from $\mathcal{G}$ in order to form $\mathcal{S}_i$ (and possibly in the diagonal blocks corresponding to nodes involved with the edges eliminated).[61] As a result, $K_i$ has rank proportional to the number of edges removed from $\mathcal{G}$.

Using (62), we can rewrite (39) as

$$\left(P_i^{-1} + C^T R^{-1} C\right) \hat{x}_s = K_i \hat{x}_s + C^T R^{-1} y \tag{63}$$

which suggests an iterative algorithm of the following form. Let $i(n)$ denote a sequence that designates which of the $T$ spanning trees is used at the $n$th iteration (e.g., chosen to cycle periodically through these trees or chosen randomly), and let $\hat{x}_n$ denote the approximation to the optimal estimate at the $n$th iteration, which is the solution to

$$\left(P_{i(n)}^{-1} + C^T R^{-1} C\right) \hat{x}_n = K_{i(n)} \hat{x}_{n-1} + C^T R^{-1} y. \tag{64}$$

Since the matrix on the left-hand side of (64) has a tree structure corresponding to the $i(n)$th tree, this equation can be solved efficiently using the MR estimation algorithm in Section IV-B.

In [334], examples are given demonstrating that such *embedded tree* (ET) algorithms can lead to very efficient methods for computing the optimal estimates. As with BP [285], [338], if an ET algorithm converges, it does so to the optimal estimate. Moreover, while BP does not yield the correct error covariances, it is shown in [334] that the computations performed in an ET algorithm can be used to compute a sequence of approximations that do converge to the correct error covariances. Furthermore, experimental evidence reported in [310] and [334] indicates that ET algorithms converge for a broader class of processes than the general, local message-passing version of BP. Roughly speaking, this is due to the fact that the algorithm takes

[61] That is, if the edge $(s,\, t)$ has been eliminated, the $(s,\, t)$, $(s,\, s)$, and $(t,\, t)$ blocks of $K_i$ may be nonzero.

advantage of much more global structure of the process, as captured by each of the spanning trees. We refer the reader to [310] and [334] for examples and theoretical analysis including conditions guaranteeing convergence of this tree-based algorithm.

We also refer the reader to [332] for the introduction and investigation of a set of algorithms for discrete-state processes that also use embedded trees but are much closer to BP. Roughly speaking, the idea behind this work is to take advantage of an interpretation discussed at the end of Section IV-B, that exact computations of conditional probabilities on a tree correspond to a refactorization of the probability distribution for a graphical model, essentially in terms of a distribution at a root node and parent–child transition distributions. The tree reparameterization algorithm developed in [332] then corresponds to iterative refactorization of the entire distribution on a loopy graph, where at each step the factorization involves only those edges corresponding to one of the spanning trees used in the algorithm. We refer the reader to [332] for the theoretical analysis of this algorithm and for examples that show its promise for inference on loopy graphs.[62] While the methods in these references have only recently been introduced, the results obtained so far and their explicit use of global rather than local graph structure suggests that there may be much more that will result from further investigation over the next few years.

## VIII. Conclusion

In this paper, we have described a framework for MR modeling and processing of signals and images. As we have seen, this framework, based on Markov models on pyramidally structured MR trees, admits efficient processing algorithms and also is rich enough to capture broad classes of statistical phenomena. As a result, these methods have found application in a variety of very different contexts. Moreover, the formalism on which this methodology is based is of deep interest intellectually, as it makes contact with a variety of topics, including wavelets, graphical models, HMMs, multigrid and coarse-to-fine algorithms, inverse problems, data fusion, state space system theory, stochastic realization theory, and maximum entropy modeling and covariance extensions.

We believe that the theory and methodology that we have described can be of value to researchers and practitioners in many different fields. In addition, we also believe that this area remains fertile ground for further basic research. For example, while the MR methods that have been developed have been successfully applied to many problems and much is known anecdotally about the problems to which they can be applied, there is still more that can be done to deepen our understanding of the problems for which these methods are appropriate and the limits to their applicability. In particular, in many signal and image processing problems (such as the texture discrimination and groundwater hydrology examples discussed in the paper), while the underlying phenomenon may be extremely complex, the available data *and* the inference objectives are much simpler and lower dimensional. This suggests (and these examples support) the idea that, for such inference problems, it may be acceptable to use relatively simple and therefore crude MR approximations which nevertheless yield near-optimal performance for specific inference problems of interest. Formalizing this idea and developing more general methods for constructing models suited for particular processing tasks remain to be accomplished.

There are many other theoretical topics that also remain to be explored. One of these is the development of data-driven algorithms for model construction analogous to the ones we have described based on explicit knowledge of the covariance structure of the process to be modeled. Here we are motivated by the fact that standard temporal modeling methods (e.g., for AR or autoregressive moving average (ARMA) models) have versions that work directly from covariance specifications and other versions that work recursively from data. We expect that data-driven algorithms can be developed that have considerable computational advantages, and we refer the reader to [167] and [325] for some initial efforts in this direction.

In addition, for a variety of reasons, it is of considerable interest to develop methods for MR modeling on graphs with cycles. One of these reasons is the recently developed set of algorithms for loopy graphs described in Section VII-B, which show that one can make use of the power of tree-based algorithms for many graphical models on loopy graphs such as in Fig. 23(a). How do we build such models? Are there variations on the stochastic realization or covariance extension methods described in Section VI-B for such loopy graphs? For example, the covariance extension results in [121] that are described in Section VI-B4 assume that the known covariance elements form a chordal graph among the variables to be modeled. However, in many applications, especially in remote sensing, that will almost never be the case. In particular, in such applications, we are likely to have knowledge of correlations among fine-scale variables (e.g., temperature variations in the ocean) that are in close proximity spatially as well as correlations among coarser spatial averages of these variables across longer distances, forming a nonchordal graph of known covariance values. Such a problem, in which we have both local fine-scale and more distant coarse-scale statistical characterizations that are of importance, is reminiscent of the structure that is exploited in multipole algorithms [256], [280] for the solution of PDEs. We refer the reader to [115] for a first attempt to adapt multipole ideas to MR estimation and also to [22], [167], and [325] for some results on MR modeling and covariance extension on graphs with cycles.

Other directions for further research can be found in virtually every corner of this paper. One is the investigation of what we have called noninternal MR realizations, a topic that offers the possibility of additional flexibility not present if we constrain ourselves to internal models (see [325] for some

[62] Note that the tree reparameterization algorithm developed in [332] is fundamentally different than the ET algorithm in [310] and [334], as the specialization of the former to linear-Gaussian models does not yield the ET algorithm. Further, as shown in [332], BP itself can be viewed as a very special variant of tree reparametrization in which very simple embedded (but not spanning) two-node trees in the graph are used at each step of the iteration.

initial results along these lines). Another is the further investigation of methods for space–time problems, either in which time is treated in an MR fashion as well (e.g., as is found in so-called multirate Kalman filtering and estimation theory [51], [79], [96], [150], [151], ) or, as in the methods discussed in Section VII-A, in which time is treated as a sequential variable but space is treated in an MR graphical manner. The results we have presented (and others in the literature, such as [274]) represent a start to this very important area which extends well beyond MR modeling to the investigation of DBNs.

As this discussion and the results summarized in the preceding sections illustrate, MR statistical modeling and inference remains a fertile, active, and important area of investigation. It is the author's hope that this paper will help to stimulate further use of the methods that already exist and inquiry into extensions that can enhance our understanding of these methods as well as the range of problems to which they can be successfully applied.

## ACKNOWLEDGMENT

The author is indebted to many individuals for the contents of this paper. First and foremost are his good friends and valued colleagues A. Benveniste and M. Basseville, not only for hosting his 1988 sabbatical but also for suggesting that they use that time to see if they could make sense of multiresolution methods and, in particular, wavelets, from a statistical perspective. Without that suggestion and his continuing and rewarding collaboration with them, this paper would certainly not exist. The author also would like to acknowledge all of the other students and colleagues with whom he has worked and interacted on research in this area: K. Chou, R. Nikoukhah, B. Levy, B. Claus, S. Golden, M. Luettgen, W. C. Karl, R. Tenney, E. Fabre, W. Irving, P. Fieguth, C. Wunsch, D. Menememlis, H. Krim, S. Mallat, T. Allen, E. Miller, M. Bhatia, L. Novak, C. Fosgate, R. Chaney, D. Taylor, A. Kim, T. M. Chin, P. Perez, F. Heitz, P. Bouthemy, M. Daniel, D. McLaughlin, D. Rossi, B. Washburn, M. Schneider, T. Ho, A. Frakt, K. Daoudi, H. Lev-Ari, A. Tsai, J. Zhang, R. Learned, J. Fisher, E. Simoncelli, D. Tucker, M. Wainwright, E. Sudderth, T. Jaakkola, J. Johnson, and M. Cetin. He would also like to thank many of these individuals as well for allowing him to use figures from their jointly authored papers. The author is also indebted to C. Bouman and R. Baraniuk for early discussions about this paper, to R. Baraniuk and E. Kolaczyk for feedback on an earlier draft of the paper, and to these three as well as J. Romberg, H. Choi, N. Kingsbury, M. Shapiro, H. Cheng, R. Nowak, P. Viola, and J. DeBonet for their willingness to let him use figures from their research. In addition, the author owes a debt of gratitude to M. Jordan whose scholarship and intellectual enthusiasm has helped to broaden and deepen his perspective on this rich research area. Thanks also to M. Cetin and T. Kelly for their efforts in helping to put this paper together and to the anonymous reviewers of this paper for their very careful reading and valuable suggestions (as well as for their impressive stamina!). Finally, to all of the researchers and authors who have worked and are working in this area, the author owes both a profound thank you for research that makes this field as interesting and rich as it is and an apology both to those authors whose work the author has cited for his undoubtedly imperfect retelling of their work and to the authors whose work the author has inadvertently omitted due to his own ignorance or forgetfulness.

## REFERENCES

[1] K. Abend, T. J. Hartley, and L. N. Kanal, "Classification of binary patterns," *IEEE Trans. Inform. Theory*, vol. IT-11, pp. 538–544, 1965.

[2] F. Abramovich, T. Sapatinas, and B. W. Silverman, "Wavelet thresholding via a Bayesian approach," *J. Roy. Stat. Soc. B*, vol. 60, pp. 725–749, 1998.

[3] F. Abramovich and B. W. Silverman, "Wavelet decomposition approaches to statistical inverse problems," *Biometrika*, vol. 85, pp. 115–129, 1998.

[4] M. B. Adams, A. S. Willsky, and B. C. Levy, "Linear estimation of boundary value stochastic processes Part I: The role and construction of complementary models," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 803–810, Sept. 1984.

[5] ——, "Linear estimation of boundary value stochastic processes—Part II: Smoothing problems," *IEEE Trans. Automat. Contr.*, vol. AC-29, pp. 811–821, Sept. 1984.

[6] E. Adelson and P. Burt, "Image data compression with the Laplacian pyramid," in *Proc. Pattern Recognition Information Processing Conf.*, 1981, pp. 218–223.

[7] S. M. Aji and R. J. McEliece, "The generalized distributive law," *IEEE Trans. Inform. Theory*, vol. 46, pp. 325–343, Mar. 2000.

[8] H. Akaike, "Stochastic theory of minimal realizations," *IEEE Trans. Automat. Contr.*, vol. AC-19, pp. 667–674, Dec. 1974.

[9] ——, "Markovian representation of stochastic processes by canonical variables," *SIAM J. Control*, vol. 12, no. 1, pp. 162–173, Jan. 1975.

[10] T. G. Allen, M. R. Luettgen, and A. S. Willsky, "Multiscale approaches to moving target detection in image sequences," *Opt. Eng.*, vol. 33, no. 7, pp. 2248–2254, July 1994.

[11] L. Ambrosio and V. M. Tortorelli, "On the approximation of free discontinuity problems," *Bolletino U.M.I.*, vol. 7, no. 6-B, pp. 105–123, 1992.

[12] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

[13] H. C. Andrews and B. R. Hunt, *Digital Image Restoration*. Englewood Cliffs, NJ: Prentice-Hall, 1977.

[14] H. Antonisse, "Image segmentation in pyramids," *Comput. Vis. Graph. Image Process.*, vol. 19, pp. 367–383, 1982.

[15] A. Arneodo, E. Bacry, and J. F. Muzy, "Random cascades on wavelet dyadic trees," *J. Math. Phys.*, vol. 39, no. 8, pp. 4142–4164, Aug. 1998.

[16] K. Arun and S. Kung, "Balanced approximations of stochastic systems," *SIAM J. Matrix Anal. Appl.*, vol. 11, no. 1, pp. 42–68, Jan. 1990.

[17] F. A. Badawi, A. Lindquist, and M. Pavon, "A stochastic realization approach to the smoothing problem," *IEEE Trans. Automat. Contr.*, vol. AC-24, pp. 878–887, Dec. 1979.

[18] B. R. Bakshi, H. Zhong, and P. Jiang, "Analysis of flow in gas-liquid bubble-columns using multiresolution methods," *Chem. Eng. Res. Des.*, vol. 73, no. A6, pp. 608–614, Aug. 1995.

[19] M. R. Banham and A. K. Katsaggelos, "Spatially adaptive wavelet-based multiscale image restoration," *IEEE Trans. Image Processing*, vol. 5, pp. 619–634, Apr. 1996.

[20] Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman, "Texture mixing and texture movie synthesis using statistical learning," *IEEE Trans. Visual. Comput. Graphics*, vol. 7, pp. 120–135, Apr.–June 2001.

[21] W. Barrett, C. Johnson, and M. Lundquist, "Determinental formulae for matrix completion associated with chordal graphs," *Linear Alg. Its Appl.*, vol. 121, pp. 265–289, 1989.

[22] W. W. Barrett, C. R. Johnson, and R. Loewy, "Critical graphs for the positive definite completion problem," *SIAM J. Matrix Anal. Appl.*, vol. 20, no. l, pp. 117–130, 1998.

[23] M. Basseville, A. Benveniste, K. C. Chou, S. A. Golden, R. Nikoukhah, and A. S. Willsky, "Modeling and estimation of multiresolution stochastic processes," *IEEE Trans. Inform. Theory*, vol. 38, pp. 766–784, Mar. 1992.

[24] M. Basseville, A. Benveniste, and A. S. Willsky, "Multiscale autoregressive processes, Part I: Schur–Levinson parametrizations," *IEEE Trans. Signal Processing*, vol. 40, pp. 1915–1934, Aug. 1992.

[25] ——, "Multiscale autoregressive processes, part II: Lattice structures for whitening and modeling," *IEEE Trans. Signal Processing*, vol. 40, pp. 1935–1954, Aug. 1992.

[26] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics*. New York: Academic, 1990.

[27] M. G. Bello, A. S. Willsky, and B. C. Levy, "Construction and applications of discrete-time smoothing error models," *Int. J. Control*, vol. 50, no. 1, pp. 203–223, July 1989.

[28] M. G. Bello, A. S. Willsky, B. C. Levy, and D. A. Castañon, "Smoothing error dynamics and their use in the solution of smoothing and mapping problems," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 483–495, July 1986.

[29] A. B. Benveniste, R. Nikoukhah, and A. S. Willsky, "Multiscale system theory," *IEEE Trans. Circuits Syst. I*, vol. 41, pp. 2–14, Jan. 1994.

[30] J. Beran, *Statistics for Long-Memory Processes*. New York: Chapman and Hall, 1994.

[31] L. M. Berliner, C. K. Wikle, and N. Cressie, "Long-lead prediction of Pacific SST's via Bayesian dynamic modeling," *J. Climate*, vol. 13, pp. 3953–3968, 2000, to be published.

[32] U. Bertelè and F. Brioschi, *Nonserial Dynamic Programming*. New York: Academic , 1972.

[33] M. Bertero, "Linear inverse and ill-posed problems," *Adv. Electron. Electron Phys.*, vol. 75, no. 1, pp. 1–120, 1989.

[34] M. Bertero, T. Poggio, and V. Torre, "Ill-posed problems in early vision," *Proc. IEEE*, vol. 76, pp. 868–889, Aug. 1988.

[35] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Roy. Stat. Soc. B*, vol. 36, no. 2, pp. 192–236, 1974.

[36] ——, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc. B*, vol. 48, no. 3, pp. 259–302, 1986.

[37] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast wavelet transforms and numerical algorithms I," *Commun. Pure Appl. Math.*, vol. 44, pp. 141–183, Mar. 1991.

[38] M. Bhatia, W. C. Karl, and A. S. Willsky, "A wavelet-based method for multiscale tomographic reconstruction," *IEEE Trans. Med. Imag.*, vol. 15, pp. 92–101, Jan. 1996.

[39] ——, "Tomographic reconstruction and estimation based on multiscale natural pixel bases," *IEEE Trans. Image Processing*, vol. 6, pp. 463–478, Mar. 1997.

[40] C. A. Bouman and B. Liu, "Multiple resolution segmentation of textured images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 99–113, Feb. 1991.

[41] C. A. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Processing*, vol. 2, pp. 296–310, July 1993.

[42] C. A. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Processing*, vol. 3, pp. 162–177, Mar. 1994.

[43] X. Boyen and D. Koller, "Tractable inference for complex stochastic processes," in *Proc. 14th Annu. Conf. Uncertainty in AI*, Madison, WI, July 1998, pp. 33–42.

[44] A. Brandt, "Multi-level adaptive solutions to boundary value problems," *Math. Comp.*, vol. 13, pp. 333–390, 1977.

[45] W. Briggs, *A Multigrid Tutorial*. Philadelphia, PA: SIAM, 1987.

[46] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Processing*, vol. 8, pp. 1688–1701, Dec. 1999.

[47] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. COM-31, pp. 532–540, Apr. 1983.

[48] P. J. Burt, T. Hong, and A. Rosenfeld, "Segmentation and estimation of image region properties through cooperative hierarchical computation," *IEEE Trans. Syst., Man Cybern.*, vol. SMC-11, pp. 802–809, Dec. 1981.

[49] S. G. Chang, B. Yu, and M. Vetterli, "Spatially adaptive wavelet thresholding with context modeling for image denoising," in *Proc. ICIP*, 1998, pp. 535–539.

[50] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 959–963, Aug. 1985.

[51] B.-S. Chen, C.-W. Lin, and Y.-Li. Chen, "Optimal signal reconstruction in noisy filter bank systems: Multirate Kalman synthesis filtering approach," *IEEE Trans. Signal Processing*, vol. 43, pp. 2496–2504, Nov. 1995.

[52] S. Chen and D. Donoho, "Atomic decomposition by basis pursuit," presented at the SPIE Int. Conf. Wavelets, San Diego, CA, July 1995.

[53] H. Cheng and C. A. Bouman, "Multiscale Bayesian segmentation using a trainable context model," *IEEE Trans. Image Processing*, vol. 10, pp. 511–525, Apr. 2001.

[54] T. M. Chin, W. C. Karl, and A. S. Willsky, "Sequential filtering for multi-frame visual reconstruction," *Signal Process.*, vol. 28, pp. 311–333, Aug. 1992.

[55] ——, "Probabilistic and sequential computation of optical flow using temporal coherence," *IEEE Trans. Image Processing*, vol. 3, pp. 773–788, Nov. 1994.

[56] T. M. Chin and A. J. Mariano, "Space–time interpolation of oceanic fronts," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 734–746, 1997.

[57] H. Chipman, E. Kolaczyk, and R. McCulloch, "Adaptive Bayesian wavelet shrinkage," *J. Amer. Stat. Assoc.*, p. 92, 1997.

[58] H. Choi and R. G. Baraniuk, "Multiscale image segmentation using wavelet-domain hidden Markov models," *IEEE Trans. Image Processing*, vol. 10, pp. 1322–1331, Sept. 2001.

[59] H. Choi, J. K. Romberg, R. G. Baraniuk, and N. G. Kingsbury, "Multiscale classification using complex wavelets and hidden Markov tree models," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Vancouver, BC, Canada, Oct. 2000, pp. 371–374.

[60] K. Chou, A. S. Willsky, and A. B. Benveniste, "Multiscale recursive estimation, data fusion, and regularization," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 464–478, Mar. 1994.

[61] K. Chou, A. S. Willsky, and R. Nikoukhah, "Multiscale systems, Kalman filters, and Riccati equations," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 479–492, Mar. 1994.

[62] K. C. Chou, S. A. Golden, and A. S. Willsky, "Multiresolution stochastic models, data, fusion, and wavelet transforms," *Signal Process.*, vol. 34, no. 3, pp. 257–282, Dec. 1993.

[63] K. C. Chou and A. S. Willsky, "A multiresolution probabilistic approach to 2D inverse conductivity problems," *Signal Process.*, vol. 18, no. 3, pp. 291–311, 1989.

[64] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 462–467, May 1968.

[65] B. Claus, "Multiscale statistical signal processing identification of a multiscale AR process from a sample of an ordinary signal," *IEEE Trans. Signal Processing*, vol. 41, pp. 3266–3274, Dec. 1993.

[66] B. Claus and G. Chartier, "Multiscale signal processing: Isotropic random fields on homogeneous trees," *IEEE Trans. Circuits Syst. II*, vol. 41, pp. 506–517, Aug. 1994.

[67] S. C. Clippingdale and R. G. Wilson, "Least squares image estimation on a multiresolution pyramid," in *Proc. ICASSP89*, vol. 3, 1989, pp. 1409–1412.

[68] M. Clyde, G. Parmigiani, and B. Vidakovic, "Multiple shrinkage and subset selection in wavelets," *Biometrika*, vol. 85, pp. 391–402, 1998.

[69] A. Cohen, J. Froment, and J. Istas, "Analyse multirésolution des signaux aléatoires," *C.R. Acad. Sci., Paris*, vol. 312, pp. 567–570, 1991.

[70] F. S. Cohen and D. B. Cooper, "Simple parallel hierarchical and relaxation algorithms for segmenting noncausal Markovian random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, pp. 195–219, Mar. 1987.

[71] F. S. Cohen, Z. Fan, and M. A. Patel, "Classification of rotated and scaled textured images using Gaussian Markov random field models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 192–202, Feb. 1991.

[72] R. Coifman and D. Donoho, "Translation invariant denoising," in *Lecture Notes in Statistics: Wavelets and Statistics*. New York: Springer Verlag, 1995, pp. 125–150.

[73] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Inform. Theory*, vol. 38, pp. 713–718, Mar. 1992.

[74] M. L. Comer and E. J. Delp, "Segmentation of textured images using a multiresolution Gaussian autoregressive model," *IEEE Trans. Image Processing*, vol. 8, pp. 408–420, Mar. 1999.

[75] T. Constantinescu, *Schur Parameters, Factorization, and Dilation Problems*. Berlin, Germany: Birkhauser, 1991.

[76] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artif. Intell.*, vol. 42, pp. 393–405, 1990.

[77] M. Costantini, A. Farina, and F. Zirilli., "The fusion of different resolution SAR images," *Proc. IEEE*, vol. 85, pp. 164–180, Jan. 1997.

[78] N. A. C. Cressie and J. L. Davidson, "Image analysis with partially ordered Markov models," *Comput. Stat. Data Anal.*, vol. 29, pp. 1–26, 1998.

[79] R. Cristi and M. Tummala, "Multirate, multiresolution recursive Kalman filter," *Signal Process.*, vol. 80, no. 9, pp. 1945–1958, Sept. 2000.

[80] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, Apr. 1998.

[81] M. M. Daniel and A. S. Willsky, "Efficient implementations of two-dimensional noncausal IIR filters," *IEEE Trans. Circuits Syst. II*, vol. 4, pp. 549–563, July 1997.

[82] ——, "A multiresolution methodology for signal-level fusion and data assimilation with applications in remote sensing," *Proc. IEEE*, vol. 85, pp. 164–183, Jan. 1997.

[83] ——, "The modeling and estimation of statistically self-similar processes in a multiresolution framework," *IEEE Trans. Inform. Theory*, vol. 45, pp. 955–970, Apr. 1999.

[84] M. M. Daniel, A. S. Willsky, and D. McLaughlin, "A multiscale approach for estimating solute travel time distributions," *Adv. Water Resources*, vol. 23, pp. 653–665, 2000.

[85] K. Daoudi, A. B. Frakt, and A. S. Willsky, "Multiscale autoregressive models and wavelets," *IEEE Trans. Inform. Theory*, vol. 45, pp. 828–845, Apr. 1999.

[86] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

[87] I. Daubechies, S. M. Mallat, and A. S. Willsky, "Special issue on wavelet transforms and multiresolution signal analysis," *IEEE Trans. Inform. Theory*, vol. 38, pp. 529–860, Mar. 1992.

[88] J. L. Davidson, N. A. C. Cressie, and X. Hua, "Texture synthesis and pattern recognition for partially ordered Markov models," *Pattern Recognit.*, vol. 32, pp. 1475–1505, 1999.

[89] A. Dawid, "Applications of a general propagation algorithm for probabilistic expert systems," *Stat. Comput.*, vol. 2, pp. 25–36, 1992.

[90] J. S. DeBonet, "Multiresolution sampling procedures for analysis and synthesis of texture images," in *Proc. SIGGRAPH'97*, 1997, pp. 361–368.

[91] J. S. DeBonet and P. Viola, "A nonparametric multiscale statistical model for natural images," *Neural Inform. Process. Syst.*, vol. 13, Dec. 1997.

[92] ——, "Texture recognition using a nonparametric multiscale statistical model," in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, 1998.

[93] L. DeCola, "Simulating and mapping spatial complexity using multiscale techniques," *Int. J. Geogr. Inform. Syst.*, vol. 8, no. 5, pp. 411–427, Sept./Oct. 1994.

[94] N. Decoster, S. G. Roux, and A. Arneodo, "A wavelet-based method for multifractal image analysis. II. Applications to synthetic multifractal rough surfaces," *Eur. Phys. J. B*, vol. 15, no. 4, pp. 739–764, June 2000.

[95] A. Delaney and Y. Bresler, "Multiresolution tomographic reconstruction using wavelets," *IEEE Trans. Image Processing*, vol. 4, pp. 799–813, June 1995.

[96] A. N. Delopoulos and S. D. Kollias, "Optimal filter banks for signal reconstruction from noisy subband components," *IEEE Trans. Signal Processing*, vol. 44, pp. 212–214, Feb. 1996.

[97] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.

[98] H. Derin, H. Elliott, R. Cristi, and D. Geman, "Bayes' smoothing algorithms for segmentation of binary images modeled by Markov random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 707–720, June 1984.

[99] X. Descombes, M. Sigelle, and F. Prêteux, "Estimating Gaussian Markov random field parameters in a nonstationary framework: Application to remote sensing imaging," *IEEE Trans. Image Processing*, vol. 8, pp. 490–503, Apr. 1999.

[100] R. W. Dijkerman and R. R. Mazumdar, "Wavelet representations of stochastic processes and multiresolution stochastic models," *IEEE Trans. Signal Processing*, vol. 42, pp. 1640–1652, July 1994.

[101] R. W. Dijkerman, R. R. Mazumdar, and A. Bagchi, "Reciprocal processes on a tree—Modeling and estimation issues," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 330–335, Feb. 1995.

[102] R. W. Dijkerman and R. R. Mazumdar, "On the correlation structure of the wavelet coefficients of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1609–1612, Oct. 1994.

[103] D. C. Dobson, "Convergence of a reconstruction method for the inverse conductivity problem," *SIAM J. Appl. Math.*, vol. 52, pp. 442–458, 1992.

[104] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Stat. Assoc.*, vol. 90, pp. 1200–1224, 1995.

[105] D. L. Donoho, "Nonlinear solution of linear inverse problems by wavelet–vaguelette decomposition," *Appl. Comput. Harmonic Anal.*, vol. 2, pp. 101–126, 1995.

[106] I. S. Duff, A. M. Erisman, and J. K. Reid, *Direct Methods for Sparse Matrices*. Oxford, U.K.: Oxford Univ. Press, 1987.

[107] R. M. Dufour and E. L. Miller, "Statistical signal restoration with $1/f$ wavelet domain prior models," *Signal Process.*, vol. 78, no. 3, pp. 289–307, Nov. 1999.

[108] D. Edwards, *Introduction to Graphical Modeling*. New York: Springer, 1993.

[109] W. Enkelmann, "Investigations of multigrid algorithms for the estimation of optical flow fields," *Comput. Vis. Graph. Image Process.*, vol. 43, pp. 150–177, 1988.

[110] E. Fabre, "New fast smoothers for multiscale systems," *IEEE Trans. Signal Processing*, vol. 44, pp. 1893–1911, Aug. 1996.

[111] P. W. Fieguth, W. C. Karl, and A. S. Willsky, "Efficient multiresolution counterparts to variational methods for surface reconstruction," *Comput. Vis. Image Understanding*, vol. 70, no. 2, pp. 157–176, May 1998.

[112] P. W. Fieguth, W. C. Karl, A. S. Willsky, and C. Wunsch, "Multiresolution optimal interpolation and statistical analysis of TOPEX/POSEIDON satellite altimetry," *IEEE Trans. Geosci. Remote Sensing*, vol. 33, pp. 280–292, Mar. 1995.

[113] P. W. Fieguth, D. Menememlis, T. Ho, and A. S. Willsky, "Mapping Mediterranean altimeter data with multiresolution optimal interpolation algorithms," *J. Atmos. Oceanic Technol.*, vol. 15, pp. 535–546, Apr. 1998.

[114] P. W. Fieguth and A. S. Willsky, "Fractal estimation using models on multiscale tress," *IEEE Trans. Signal Processing*, vol. 44, pp. 1297–1300, May 1996.

[115] P. W. Fieguth, "Multipole-motivated reduced-state estimation," in *Proc. ICIP*, vol. 1, 1998, pp. 635–638.

[116] P. Flandrin, "On the spectrum of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 35, pp. 197–199, Jan. 1989.

[117] ——, "Wavelet analysis and synthesis of fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 38, pp. 910–917, Mar. 1992.

[118] G. D. Forney, "The Viterbi algorithm," *Proc. IEEE*, vol. 61, pp. 268–278, 1973.

[119] C. H. Fosgate, H. Krim, W. W. Irving, W. C. Karl, and A. S. Willsky, "Multiscale segmentation and anomaly enhancement of SAR imagery," *IEEE Trans. Image Processing*, vol. 6, pp. 7–20, Jan. 1997.

[120] A. B. Frakt, W. C. Karl, and A. S. Willsky, "A multiscale hypothesis testing approach to anomaly detection and localization from tomographic data," *IEEE Trans. Image Processing*, vol. 7, pp. 825–837, June 1998.

[121] A. B. Frakt, H. Lev-Ari, and A. S Willsky, "A generalized Levinson algorithm for covariance extension with applications to multiscale autoregressive modeling," *IEEE Trans. Inform. Theory*, to be published.

[122] A. B. Frakt and A. S. Willsky, "A scale-recursive method for constructing mutiscale stochastic models," *Multidimensional Signal Process.*, vol. 12, pp. 109–142, 2001.

[123] W. Freeman and Y. Weiss, "On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs," *IEEE Trans. Inform. Theory*, vol. 47, pp. 736–744, 2001.

[124] T. Frese, C. A. Bouman, N. C. Rouze, G. D. Hutchins, and K. Sauer, "Bayesian multiresolution algorithm for PET reconstruction," in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Vancouver, BC, Canada, Sept. 10–13, 2000, pp. 613–616.

[125] T. Frese, C. A. Bouman, and K. Sauer, "Multiscale Bayesian methods for discrete tomography," in *Discrete Tomography: Foundations, Algorithms and Applications*, G.T. Herman and A. Kuba, Eds. Cambridge, MA: Birkhauser, 1999.

[126] ——, "Multiscale models for Bayesian inverse problems," in *Proc. SPIE Conf. Wavelet Applications in Signal Image Processing VII*, vol. 3813, Denver, CO, July 19–23, 1999, pp. 85–96.

[127] ——, "Adaptive wavelet graph model for Bayesian tomographic reconstruction," *IEEE Trans. Image Processing*, vol. 11, pp. 756–770, July, 2002.

[128] B. Frey, *Graphical Models for Machine Learning and Digital Communication*. Cambridge, MA: MIT Press, 1998.

[129] N. Gantert, "Self-similarity of Brownian-motion and a large deviation principle for random fields on a binary tree," *Prob. Theory Related Fields*, vol. 98, no. 1, pp. 7–20, Jan. 1994.

[130] P. Gaspar and C. Wunsch, "Estimates from altimeter data of barytropic Rossby waves in the northwestern Atlantic ocean," *J. Phys. Oceanogr.*, vol. 19, no. 12, pp. 1821–1844, 1989.

[131] D. Geiger and J. Kogler, "Scaling images and image features via the renormalization group," in *Proc. IEEE CVPR 93*, June 1993, pp. 47–53.

[132] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, Nov. 1984.

[133] A. George, "Nested dissection of a regular finite element mesh," *SIAM J. Numer. Anal.*, vol. 10, no. 2, pp. 345–363, Apr. 1973.

[134] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Machine Learning*, vol. 29, pp. 245–273, 1997.

[135] B. Gidas, "A renormalization group approach to image processing problems," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, pp. 164–180, Feb. 1989.

[136] J. Giesemann, M. Greiner, and P. Lipa, "Wavelet cascades," *Physica A*, vol. 247, pp. 41–58, Dec. 1997.

[137] S. A. Golden, "Identifying multiscale statistical models using the wavelet transform," Master's thesis, Massachusetts Inst. Technol., Cambridge, June 1991.

[138] G. Golub and C. Van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[139] I. P. Gorenburg, D. McLaughlin, and D. Entekhabi, "Scale-recursive estimation of precipitation at the TOAGA-COARE site," *Adv. Water Resources*, vol. 24, pp. 941–953, 2001.

[140] C. Graffigne, F. Heitz, P. Perez, F. Prêteux, M. Sigelle, and J. Zerubia, "Hierarchical Markov random field models applied to image analysis: A review," in *SPIE Conf. Neural, Morphological Stochastic Methods in Image Signal Processing*, vol. 2568, San Diego, CA, July 10–11, 1995, pp. 12–17.

[141] C. D. Greene and B. C. Levy, "Some new smoother implementations for discrete-time Gaussian reciprocal processes," *Int. J. Control*, vol. 54, no. 5, pp. 1233–1247, 1991.

[142] R. Grone, C. Johnson, E. Sa, and H. Wolkowicz, "Positive definite completions of partial Hermitian matrices," *Linear Alg. Its Appl.*, vol. 58, pp. 109–124, 1984.

[143] X. Guyon, *Random Fields on a Network: Modeling, Statistics, and Applications*. New York: Springer-Verlag, 1995.

[144] F. Heitz, P. Perez, and P. Bouthemy, "Multiscale minimization of global energy functions in some visual recovery problems," *Comput. Vis. Graph. Image Process.*, vol. 59, no. 1, pp. 125–134, Jan. 1994.

[145] F. Heitz, P. Perez, E. Memin, and P. Bouthemy, "Parallel visual motion analysis using multiscale Markov random fields," in *Proc. IEEE Workshop Visual Motion*, Oct. 1991, pp. 30–35.

[146] G. A. Hirchoren and C. E. D'Attellis, "Estimation of fractal signals using wavelets and filter banks," *IEEE Trans. Signal Processing*, vol. 46, pp. 1624–1630, June 1998.

[147] T. L. Ho, P. W. Fieguth, and A. S. Willsky, "Computationally efficient multiscale estimation for 1-D diffusion processes," *Automatica*, vol. 37, pp. 325–340, Mar. 2001.

[148] Y. C. Ho and K. C. Chu, "Team decision theory and information structures in optimal control problems, Part I," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 15–22, Feb. 1972.

[149] ——, "Team decision theory and information structures in optimal control problems, Part II," *IEEE Trans. Automat. Contr.*, vol. AC-17, pp. 22–28, Feb. 1972.

[150] L. Hong, "Multiresolutional filtering using wavelet transform," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 29, pp. 1244–1251, Oct. 1993.

[151] ——, "Multiresolutional distributed filtering," *IEEE Trans. Automat. Contr.*, vol. 39, pp. 853–856, Apr. 1994.

[152] B. Horn, *Robot Vision*. Cambridge, MA: MIT Press, 1986.

[153] ——, "Height and gradient from shading," *Int. J. Comput. Vis.*, vol. 5, no. 1, pp. 37–46, 1990.

[154] C. Houdré, "Wavelets, probability and statistics: Some bridges," in *Wavelets: Mathematics and Applications*, J. Benedetto and M. Frazier, Eds. Boca Raton, FL: CRC Press, Feb. 1993.

[155] H. C. Huang and N. Cressie, "Deterministic/stochastic wavelet decomposition for recovery of signal from noisy data," *Technometrics*, vol. 42, no. 3, pp. 262–276, Aug. 2000.

[156] H. C. Huang and N. A. C. Cressie, "Multiscale graphical modeling in space: Applications to command and control," in *Spatial Statistics: Methodological Aspects and Some Applications*. ser. Springer Lecture Notes in Statistics, M. Moore, Ed. New York: Springer-Verlag, 2001, vol. 159.

[157] J. Huang and D. Mumford, "Statistics of natural images and models," in *Proc. CVPR*, vol. 1, 1999, pp. 541–547.

[158] H.-C. Huang, N. A. C. Cressie, and J. Gabrosek, "Fast, resolution-consistent spatial prediction of global processes from satellite data," *J. Comput. Graph. Stat.*, vol. 11, pp. 63–88, 2002.

[159] W. W. Irving, P. W. Fieguth, and A. S. Willsky, "An overlapping tree approach to multiscale stochastic modeling and estimation," *IEEE Trans. Image Processing*, vol. 6, pp. 1517–1529, Nov. 1997.

[160] W. W. Irving, L. M. Novak, and A. S. Willsky, "A multiresolution approach to discrimination in SAR imagery," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, pp. 1157–1169, Oct. 1997.

[161] W. W. Irving and A. S. Willsky, "Multiscale stochastic realization using canonical correlations," *IEEE Trans. Automat. Contr.*, vol. 46, pp. 1514–1528, Oct. 2001.

[162] L. M. Ivanov, A. D. Kirwan, and T. M. Margolina, "Filtering noise from oceanographic data with some applications for the Kara and Black Seas," *J. Marine Syst.*, vol. 28, no. 1–2, pp. 113–139, Feb. 2001.

[163] A. K. Jain and E. Angel, "Image restoration, modeling, reduction of dimensionality," *IEEE Trans. Comput.*, vol. C-23, pp. 470–476, 1974.

[164] F. C. Jeng, "Subsampling of Markov random fields," *J. Vis. Commun. Image Repres.*, vol. 3, no. 3, pp. 225–229, Sept. 1992.

[165] D. H. Johnson and A. R. Kumar, "Modeling and analyzing fractal point processes," in *Proc. ICASSP*, vol. 3, 1990, pp. 1353–1356.

[166] J. Johnson, "Estimation of Gaussian MRF's by recursive cavity modeling," S.M. thesis, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, Sept. 2002.

[167] J. Johnson, E. Sudderth, D. Tucker, M. Wainwright, and A. S. Willsky, "Multiresolution and graphical models for images and spatial data," in *Proc. SIAM Conf. Imaging Science*, Mar. 2002, p. 33.

[168] M. I. Jordan, Ed., *Learning in Graphical Models*. Cambridge, MA: MIT Press, 1998.

[169] M. I. Jordan, *An Introduction to Probabilistic Graphical Models*, to be published.

[170] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning Inference in Graphical Models*. Cambridge, MA: MIT Press, 1998, pp. 75–104.

[171] A. G. Journel, "Geostatistics for conditional simulation of ore bodies," *Econ. Geol.*, vol. 69, pp. 673–687, 1974.

[172] ——, *Fundamentals of Geostatistics in Five Lessons*. Washington, DC: Amer. Geophysics Union, 1989.

[173] A. G. Journel and C. T. Huijbregts, *Mining Statistics*. New York: Academic, 1978.

[174] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[175] A. Kannan, M. Ostendorf, W. C. Karl, D. A. Castañon, and R. Fish, "ML parameter estimation of multiscale stochastic processes using the EM algorithm," *IEEE Trans. Signal Processing*, vol. 48, pp. 1836–1847, June 2000.

[176] L. M. Kaplan and C.-C. J. Kuo, "Fractal estimation from noisy data via discrete fractional Gaussian noise (DFGN) and the Haar basis," *IEEE Trans. Signal Processing*, vol. 41, pp. 3354–3562, Dec. 1993.

[177] D. Karger and N. Srebro, "Learning Markov networks: Maximum bounded tree-width graphs," in *Proc. 12th ACM-SIAM Symp. Discrete Algorithms*, Jan. 2001, pp. 392–401.

[178] R. Kashyap, R. Chellappa, and A. Khotanzad, "Texture classification using features derived from random field models," *Pattern Recognit. Lett.*, vol. 1, no. 1, pp. 43–50, Oct. 1982.

[179] Z. Kato, M. Berthod, and J. Zerubia, "A hierarchical Markov random field model and multitemperature annealing for parallel image classification," *Graph. Models Image Process.*, vol. 58, no. 1, pp. 18–37, Jan. 1996.

[180] Z. Kato, J. Zerubia, and M. Berthod, "Unsupervised parallel image classification using Markovian models," *Pattern Recognit.*, vol. 32, no. 4, pp. 591–604, Apr. 1999.

[181] J. Kaufhold, M. Schneider, W. C. Karl, and A. S. Willsky, "A statistical method for efficient segmentation of MR imagery," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 11, no. 8, 1997.

[182] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall, 1993.

[183] K. W. Khawaja, A. A. Maciejewski, D. Tretter, and C. A. Bouman, "Automated assembly inspection using a multiscale algorithm trained on synthetic images," *IEEE Robot. Automat. Mag.*, vol. 3, pp. 15–22, June 1996.

[184] P. W. Fieguth, F. M. Khellah, M. J. Murray, and M. R. Allen, "Data fusion of sea-surface temperature data," in *Proc. IGARSS*, vol. 5, 2000, pp. 2111–2113.

[185] A. J. Kim and H. Krim, "Hierarchical stochastic modeling of SAR imagery for segmentation/compression," *IEEE Trans. Signal Processing*, vol. 47, pp. 458–468, Feb. 1999.

[186] N. G. Kingsbury, "Image processing with complex wavelets," *Phil. Trans. Roy. Soc. London A*, vol. 357, pp. 2543–2560, Sept. 1999.

[187] E. Kolaczyk, "A wavelet shrinkage approach to tomographic image reconstruction," *J. Amer. Stat. Assoc.*, vol. 91, pp. 1079–1990, 1996.

[188] ——, "Bayesian multi-scale models for Poisson processes," *J. Amer. Stat. Assoc.*, vol. 94, pp. 920–933, 1999.

[189] E. D. Kolaczyk and H. Y. Huang, "Multiscale statistical models for hierarchical spatial aggregation," *Geogr. Anal.*, vol. 33, no. 2, pp. 95–118, Apr. 2001.

[190] J. Konrad and E. Dubois, "Multigrid Bayesian estimation of image motion fields using stochastic relaxation," in *Proc. 2nd Int. Conf. Computer Vision*, Dec. 1988, pp. 354–362.

[191] H. Krim and J.-C. Pesquet, "Multiresolution analysis of a class of nonstationary processes," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1010–1020, July 1995.

[192] H. Krim and I. C. Schick, "Minimax description length for signal denoising and optimized representation," *IEEE Trans. Inform. Theory*, vol. 45, pp. 898–908, May 1999.

[193] H. Krim, D. Tucker, S. Mallat, and D. Donoho, "On denoising and best signal representation," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2225–2238, Nov. 1999.

[194] H. Krim, W. Willinger, A. Juditski, and D. Tse, "Special issue on multiscale statistical signal analysis and its applications," *IEEE Trans. Inform. Theory*, vol. 45, pp. 825–1062, Apr. 1999.

[195] S. Krishnamachari and R. Chellappa, "Multiresolution Gauss–Markov random field models for texture segmentation," *IEEE Trans. Image Processing*, vol. 6, pp. 251–267, Feb. 1997.

[196] A. Krishnan and K. A. Hoo, "A multiscale model predictive control strategy," *Ind. Eng. Chem. Res.*, vol. 38, no. 5, pp. 1973–1986, May 1999.

[197] F. Kschischang and B. Frey, "Iterative decoding of compound codes by probability propagation in graphical models," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 219–230, Feb. 1998.

[198] P. Kumar, "A multiple scale state-space model for characterizing subgrid scale variability of near-surface soil moisture," *IEEE Trans. Geosci. Remote Sensing*, vol. 37, pp. 182–197, 1999.

[199] J.-M. Laferte, P. Perez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. Image Processing*, vol. 9, pp. 390–404, Mar. 2000.

[200] S. Lakshmanan and H. Derin, "Gaussian Markov random fields at multiple resolutions," in *Markov Random Fields: Theory and Applications*, R. Chellappa, Ed. New York: Academic, 1993, pp. 131–157.

[201] W. M. Lam and G. W. Wornell, "Multiscale representation and estimation of fractal point processes," *IEEE Trans. Signal Processing*, vol. 43, pp. 2606–2617, Nov. 1995.

[202] ——, "Multiscale analysis and control of networks with fractal traffic," *Appl. Comput. Harmon. Anal.*, 2001.

[203] M. Lang, H. Guo, J. E. Odegard, C. S. Burrus, and R. O. Wells, "Noise reduction using an undecimated discrete wavelet transform," *IEEE Signal Processing Lett.*, vol. 3, pp. 10–12, 1996.

[204] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

[205] R. Learned, A. S. Willsky, and D. M. Boroson, "Low complexity optimal joint detection for over-saturated multiple access communications," *IEEE Trans. Signal Processing*, vol. 45, pp. 113–123, Jan. 1997.

[206] N.-Y. Lee and B. J. Lucier, "Wavelet methods for inverting the Radon transform with noisy data," *IEEE Trans. Image Processing*, vol. 10, pp. 79–94, Jan. 2001.

[207] H.-G. Leimer, "Optimal decomposition by clique separators," *Discrete Math.*, vol. 113, pp. 99–123, 1993.

[208] B. C. Levy, "Noncausal estimation for discrete Gauss–Markov random fields," in *Proc. Int. Symp. Mathematical Theory of Networks and Systems*, vol. 2, 1989, pp. 13–21.

[209] ——, "Multiscale models and estimation of discrete Gauss–Markov random fields," in *Proc. 2nd SIAM Conf. Linear Algebra in Systems, Control, and Signal Processing*, 1990.

[210] B. C. Levy, R. Frezza, and A. J. Krener, "Modeling and estimation of discrete-time Gaussian reciprocal processes," *IEEE Trans. Automat. Contr.*, vol. 35, pp. 1013–1023, Sept. 1990.

[211] P. Lévy, "Le mouvement Brownien," *Mem. Sci. Math.*, vol. 126, pp. 1–81, 1954.

[212] C.-T. Li and R. Wilson, "Image segmentation based on a multiresolution Bayesian framework," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, Chicago, IL, Oct. 4–7, 1998, pp. 761–765.

[213] J. Li, R. M. Gray, and R. A. Olshen, "Multiresolution image classification by hierarchical modeling with two-dimensional hidden Markov models," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1826–1841, Aug. 2000.

[214] A. Lindquist and G. Picci, "On the stochastic realization problem," *SIAM J. Control Optim.*, vol. 17, no. 3, pp. 365–389, May 1975.

[215] J. Liu, "A multiresolution method for distributed parameter estimation," *SIAM J. Sci. Comput.*, vol. 14, no. 2, pp. 389–405, 1993.

[216] L. Ljung, *System Identification: Theory for the User*. Upper Saddle River, NJ: Prentice-Hall, 1999.

[217] L. Ljung and T. Söderström, *Theory and Practice of Recursive Identification*. Cambridge, MA: MIT Press, 1983.

[218] S. M. LoPresto, K. Ramchandran, and M. T. Orchard, "Image coding based on mixture modeling of wavelet coefficients and a fast estimation-quantization framework," in *Proc. Data Compression Conf.*, Snowbird, UT, 1997, pp. 221–230.

[219] S. Lovejoy and D. Schartzer, "Generalized scale invariance in the atmosphere and fractal models of rain," *Water Resources Res.*, vol. 21, pp. 1233–1250, Aug. 1985.

[220] ——, "Multifractals and rain," in *New Uncertainty Concepts in Hydrology and Hydrological Modeling*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[221] S. B. Lowen and M. C. Teich, "Fractal renewal processes generate $1/f$ noise," *Phys. Rev. E*, vol. 47, pp. 992–1001, 1993.

[222] H. Lucke, "Which stochastic models allow Baum–Welch training?," *IEEE Trans. Signal Processing*, vol. 44, pp. 2746–2756, Nov. 1996.

[223] M. R. Luettgen, W. C. Karl, and A. S. Willsky, "Efficient multiscale regularization with applications to the computation of optical flow," *IEEE Trans. Image Processing*, vol. 3, pp. 41–64, Jan. 1994.

[224] M. R. Luettgen, W. C. Karl, A. S. Willsky, and R. R. Tenney, "Multiscale representations of Markov random fields," *IEEE Trans. Signal Processing*, vol. 41, pp. 3377–3396, Dec. 1993.

[225] M. R. Luettgen and A. S. Willsky, "Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination," *IEEE Trans. Image Processing*, vol. 4, pp. 194–207, Feb. 1995.

[226] ——, "Multiscale smoothing error models," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 173–175, Jan. 1995.

[227] P. Malanotte-Rizzoli, "Oceanographic data assimilation in the 1990s: Overview, motivation, and purposes," *Naval Res. Rev.*, vol. 51, no. 2, 1999.

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

[229] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency libraries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[230] F. M. Malvestuto, "Approximating discrete probability distributions with decomposable models," *IEEE Trans. Syst., Man Cybern.*, vol. 21, pp. 1287–1294, 1991.

[231] B. B. Mandelbrot, "Self-similar error clusters in communication systems and the concept of conditional stationarity," *IEEE Trans. Commun.*, vol. COM-13, pp. 71–90, 1965.

[232] B. Mandlebrot and J. W. Van Ness, "Fractional Brownian motions, fractional noises, and applications," *SIAM Rev.*, vol. 10, pp. 422–437, 1968.

[233] B. S. Manjunath and R. Chellappa, "Unsupervised texture segmentation using Markov random field models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 478–482, May 1991.

[234] J. Marroquin, S. K. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Stat. Assoc.*, vol. 82, pp. 76–89, Mar. 1987.

[235] E. Masry, "The wavelet transform of stochastic processes with stationary increments and its application to fractional Brownian motion," *IEEE Trans. Inform. Theory*, vol. 39, pp. 260–264, Jan. 1993.

[236] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, “Turbo decoding as an instance of Pearl’s ‘belief propagation’ algorithm,” *IEEE J. Select. Areas Commun.*, vol. 16, pp. 140–152, 1998.

[237] M. Meila, “Learning with mixtures of trees,” Ph.D. dissertation, MIT, Cambridge, 1998.

[238] ——, “An accelerated Chow and Liu algorithm: Fitting tree distributions to high-dimensional sparse data,” in *Proc. ICM*, 1999.

[239] M. Meila and T. Jaakkola, “Tractable Bayesian learning of tree belief networks,” in *Proc. UAI*, 2000, pp. 380–388.

[240] M. Meila and M. I. Jordan, “Estimating dependency structure as a hidden variable,” in *Proc. NIPS’97*, 1997, pp. 584–590.

[241] D. Menemenlis and M. Chechelnitsky, “Error estimates for an ocean general circulation model from altimeter and acoustic tomography data,” *Mon. Weather Rev.*, vol. 128, no. 3, pp. 763–778, Mar. 2000.

[242] D. Menemenlis, P. W. Fieguth, C. Wunsch, and A. S. Willsky, “Adaptation of a fast optimal interpolation algorithm to the mapping of oceanographic data,” *J. Geophys. Res.*, vol. 102, pp. 10 573–10 584, May 1997.

[243] M. Mihçak, I. Kozintsev, K. Ramchandran, and P. Moulin, “Low-complexity image denoising based on statistical modeling of wavelet coefficients,” *IEEE Signal Processing Lett.*, vol. 6, pp. 300–303, 1999.

[244] E. L. Miller, L. Nicolaides, and A. Mandelis, “Nonlinear inverse scattering methods for thermal-wave slice tomography: A wavelet domain approach,” *J. Opt. Soc. Amer. A*, vol. 15, pp. 1545–1556, 1998.

[245] E. L. Miller and A. S. Willsky, “Multiscale statistical anomaly detection, analysis, and algorithms for linearized inverse scattering problems,” *Multidimensional Syst. Signal Process.*, vol. 8, no. 1–2, pp. 151–184, 1994.

[246] ——, “A multiscale approach to sensor fusion and the solution of linear inverse problems,” *Appl. Comput. Harmon. Anal.*, vol. 2, pp. 127–147, 1995.

[247] ——, “A multiscale statistically-based inversion scheme for the linearized inverse scattering problem,” *IEEE Trans. Geosci. Remote Sensing*, vol. 34, pp. 346–357, Mar. 1996.

[248] ——, “Wavelet-based methods for the nonlinear inverse scattering problem using the extended Born approximation,” *Radio Sci.*, vol. 31, no. 1, pp. 51–56, 1996.

[249] B. T. Milne and W. B. Cohen, “Multiscale assessment of binary and continuous landcover variables for MODIS validation, mapping, and modeling applications,” *Remote Sens. Environ.*, vol. 70, no. 1, pp. 82–98, Oct. 1999.

[250] P. Moulin and J. Liu, “Analysis of multiresolution image denoising schemes using a generalized Gaussian and complexity priors,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 909–919, Apr. 1999.

[251] P. Müller and B. Vidakovic, Eds., *Bayesian Inference in Wavelet-Based Models*. New York: Springer, 1999.

[252] D. Mumford and J. Shah, “Optimal approximation by piecewise smooth functions and associated variational problems,” *Commun. Pure Appl. Math.*, vol. 42, pp. 577–685, 1989.

[253] W. Munk, P. Worcester, and C. Wunsch, *Ocean Acoustic Tomography*. Cambridge, MA: Cambridge Univ. Press, 1995.

[254] K. Murphy and Y. Weiss, “The factored frontier algorithm for approximate inference in DBN’s, uncertainty in artificial intelligence,” in *Proc. 17th Conf. (UAI-2001)*, 2001, pp. 378–385.

[255] M. J. Murray, P. W. Fieguth, and M. R. Allen, “Combined ATSR/AVHRR skin sea temperature analysis,” in *Proc. Eur. Geophys. Soc.’00*, 2000.

[256] K. Nabors, F. T. Korsmeyer, F. T. Leighton, and J. White, “Preconditioned adaptive multipole-accelerated iterative methods for three-dimensional first-kind integral equations of potential theory,” *SIAM J. Sci. Comput.*, vol. 15, no. 3, pp. 713–735, 1994.

[257] G. K. Nicholls and M. Petrou, “A generalization of renormalisation group methods for multiresolution image analysis,” in *Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition*, 1992, pp. 567–570.

[258] R. Nikoukhah, M. B. Adams, A. S. Willsky, and B. C. Levy, “Estimation for boundary-value descriptor systems,” *Circuits, Syst., Signal Process.*, vol. 8, no. 1, pp. 25–48, Apr. 1989.

[259] R. Nikoukhah, D. Taylor, B. C. Levy, and A. S. Willsky, “Graph structure and recursive estimation of noisy linear relations,” *J. Math. Syst., Est., Contr.*, vol. 5, no. 4, pp. 1–37, 1995.

[260] R. Nikoukhah, A. S. Willsky, and B. C. Levy, “Kalman filtering and Riccati equations for descriptor systems,” *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1325–1342, Sept. 1992.

[261] R. Nowak, “Multiscale hidden Markov models for Bayesian image analysis,” in *Bayesian Inference in Wavelet Based Models*, B. Vidakovic and P. Müller, Eds. New York: Springer-Verlag, 1999.

[262] R. D. Nowak, “Shift invariant wavelet-based statistical models and $1/f$ processes,” in *Proc. IEEE DSP Workshop*, 1998.

[263] R. D. Nowak and E. D. Kolaczyk, “A statistical multiscale framework for Poisson inverse problems,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 1811–1825, Aug. 2000.

[264] T. Olson and J. DeStefano, “Wavelet localization of the Radon transform,” *IEEE Trans. Image Processing*, vol. 42, pp. 2055–2067, Aug. 1994.

[265] M. Ostendorf, V. Digalakis, and O. Kimball, “From HMM’s to segment models: A unified view of stochastic modeling for speech recognition,” *IEEE Trans. Speech Audio Processing*, pp. 360–378, Oct. 1996.

[266] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 1991.

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

[268] A. Pentland, “Fractal-based description of natural scenes,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 661–674, Nov. 1984.

[269] P. Perez and F. Heitz, “Restriction of a Markov random field on a graph and multiresolution statistical image modeling,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 180–190, Jan. 1996.

[270] J.-C. Pesquet, H. Krim, and H. Carfantan, “Time-invariant orthonormal wavelet representations,” *IEEE Trans. Signal Processing*, vol. 44, pp. 1964–1970, Aug. 1996.

[271] F. Peyrin, M. Zaim, and R. Goutte, “Multiscale reconstruction of tomographic images,” in *Proc. IEEE-SP Int. Symp. Time-Frequency Time-Scale Analysis*, 1992, pp. 219–222.

[272] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[273] J. Ramanathan and O. Zeitouni, “On the wavelet transform of fractional Brownian motion,” *IEEE Trans. Inform. Theory*, vol. 37, pp. 1156–1158, July 1991.

[274] R. P. N. Rao and D. H. Ballard, “Dynamic model of visual recognition predicts neural response properties in the visual cortex,” *Neural Comput.*, vol. 9, no. 4, pp. 721–763, May 1997.

[275] S. Rao and W. A. Pearlman, “Analysis of linear prediction, coding, and spectral estimation from subbands,” *IEEE Trans. Inform. Theory*, vol. 42, pp. 1160–1178, July 1996.

[276] F. Rashid-Farrokhi, K. J. R. Liu, C. A. Berenstein, and D. Walnut, “Wavelet-based multiresolution local tomography,” *IEEE Trans. Image Processing*, vol. 6, pp. 1412–1430, Oct. 1997.

[277] H. Rauch, F. Tung, and C. Striebel, “Maximum likelihood estimates of linear dynamic systems,” *AIAA J.*, vol. 3, no. 8, pp. 1445–1450, Aug. 1965.

[278] T. Richardson, “The geometry of turbo-decoding dynamics,” *IEEE Trans. Inform. Theory*, vol. 46, pp. 9–23, Jan. 2000.

[279] R. H. Riedi, M. S. Crouse, V. J. Ribeiro, and R. G. Baraniuk, “A multifractal wavelet model with application to network traffic,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 992–1018, Apr. 1999.

[280] V. Rokhlin, “Rapid solution of integral equations of classical potential theory,” *J. Comput. Phys.*, vol. 60, no. 2, pp. 187–207, 1985.

[281] J. K. Romberg, H. Choi, and R. G. Baraniuk, “Bayesian tree-structured image modeling,” *IEEE Trans. Image Processing*, vol. 10, pp. 1056–1068, July 2001.

[282] J. K. Romberg, H. Choi, R. G. Baraniuk, and N. Kingsbury, “Hidden Markov tree models for complex wavelet transforms,” *IEEE Trans. Image Processing*, submitted for publication.

[283] O. Ronen, J. R. Rohlicek, and M. Ostendorf, “Parameter estimation of dependence tree models using the EM algorithm,” *IEEE Signal Processing Lett.*, vol. 2, pp. 157–159, Aug. 1995.

[284] Yu. A. Rozanov, *Random Fields and Stochastic Partial Differential Equations*. Norwell, MA: Kluwer, 1998.

[285] P. Rusmevichientong and B. Van Roy, “An analysis of turbo decoding with Gaussian priors,” in *Proc. NIPS*, 2000, vol. 12, pp. 575–581.

[286] B. Sahiner and A. Yagle, “Image reconstruction from projections under wavelet constraints,” *IEEE Trans. Signal Processing*, vol. 41, pp. 3579–3584, Dec. 1993.

[287] ——, “Iterative inversion of the Radon transform using image-adaptive wavelet constraints,” in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1998.

[288] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes: Stochastic Models With Infinite Variance*. New York: Chapman & Hall, 1994.

[289] S. S. Saquib, C. A. Bouman, and K. Sauer, “A nonhomogeneous MRF model for multiresolution Bayesian estimation,” in *Proc. IEEE Int. Conf. Image Processing*, vol. 2, Lausanne, Switzerland, Sept. 16–19, 1996, pp. 445–448.

[290] “Purdue Univ.,”, Dept. Electrical and Computer Eng., West Lafayette, IN.

[291] L. K. Saul and M. I. Jordan, “Exploiting tractable substructures in intractable networks,” in *Proc. NIPS*, 1995, pp. 486–492.

[292] M. Schneider, P. W. Fieguth, W. C. Karl, and A. S. Willsky, “Multiscale statistical methods for the segmentation of images,” *IEEE Trans. Image Processing*, vol. 9, pp. 442–455, Mar. 2000.

[293] F. Sellan, “Synthèse de mouvements Brownien fractionnaires á l’aide de la transformation par ondelettes,” *C.R.A.S. Paris ser. I Math*, vol. 321, pp. 351–358, 1995.

[294] G. R. Shafer and P. P. Shenoy, “Probability propagation,” *Ann. Math. Artif. Intell.*, vol. 2, pp. 327–352, 1990.

[295] ——, “Valuation-based systems for Bayesian decision analysis,” *Oper. Res.*, vol. 40, pp. 463–484, 1992.

[296] J. Shapiro, “Embedded image coding using zerotrees of wavelet coefficients,” *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[297] E. P. Simoncelli and R. W. Buccigrossi, “Embedded wavelet image compression based on a joint probability model,” in *Proc. ICIP*, vol. I, Santa Barbara, CA, Oct. 26–29, 1997, pp. 640–643.

[298] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, “Shiftable multiscale transforms,” *IEEE Trans. Inform. Theory*, vol. 38, pp. 587–607, Mar. 1992.

[299] E. P. Simoncelli, “Statistical models for images: Compression, restoration, and synthesis,” in *Proc. 31st Asilomar Conf.*, Nov. 1997, pp. 673–678.

[300] ——, “Bayesian denoising of visual images in the wavelet domain,” in *Bayesian Inference in Wavelet Based Models*, P. Muller and B. Vidakovic, Eds. New York: Springer-Verlag, June 1999, vol. 141 of Lecture Notes in Statistics, pp. 291–308.

[301] E. P. Simoncelli and E. H. Adelson, “Noise removal via Bayesian wavelet coring,” in *Proc. ICIP*, vol. 1, 1996, pp. 379–382.

[302] P. Smyth, D. Heckerman, and M. Jordan, “Probabilistic independence networks for hidden Markov probability models,” *Neural. Comput.*, vol. 9, pp. 227–269, 1997.

[303] N. Srebro, “Maximum likelihood bounded tree-width Markov networks,” in *Proc. 17th Conf. Uncertainty in Art. Intell.*, Aug. 2001, pp. 504–511.

[304] F. M. J.-L. Starck and A. Bijaoui, *Multiscale Image Processing and Data Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[305] I. M. Stephanakis, “Regularized image restoration in multiresolution spaces,” *Opt. Eng.*, vol. 36, no. 6, pp. 1738–1744, June 1997.

[306] G. Stephanopoulos, O. Karsligil, and M. Dyer, “Multi-scale aspects in model-predictive control,” *J. Process. Control*, vol. 10, no. 2–3, pp. 275–282, Apr.–Jun. 2000.

[307] P. Stoica and R. Moses, *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice-Hall, 1997.

[308] V. Strela, J. Portilla, and E. Simoncelli, “Image denoising using a local Gaussian scale mixture model in the wavelet domain,” presented at the SPIE Conf., San Diego, CA, July 2000.

[309] N. S. Subotic, B. J. Thelen, J. D. Gorman, and M. F. Reilly, “Multiresolution detection of coherent radar targets,” *IEEE Trans. Image Processing*, vol. 6, pp. 21–35, Jan. 1997.

[310] E. B. Sudderth, “Embedded trees: Estimation of Gaussian processes on graphs with cycles,” S.M. thesis, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, Feb. 2002.

[311] S. M. Sussman, “Analysis of the pareto model for error statistics on telephone circuits,” *IEEE Trans. Commun.*, vol. COM-11, pp. 213–221, Feb. 1963.

[312] R. Szeliski, *Bayesian Modeling of Uncertainty in Low-Level Vision*. Norwell, MA: Kluwer, 1989.

[313] M. S. Taqqu, “A bibliographic guide to self-similar processes and long-range dependencies,” in *Dependence in Probability and Statistics*, E. Eberlein and M. S. Taqqu, Eds. Boston, MA: Birkhäuser, 1986.

[314] R. E. Tarjan, “Decomposition by clique separators,” *Discrete Math.*, vol. 55, pp. 221–232, 1985.

[315] R. E. Tarjan and M. Yannakakis, “Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs,” *SIAM J. Comput.*, vol. 13, pp. 566–579, 1984.

[316] D. Taylor, “Parallel estimation on one and two dimensional systems,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, Feb. 1992.

[317] C. Taylore and A. Colchester, Eds., *Medical Image Computing and Computer-Assisted Intervention: MICCA’99*. New York: Springer, 1999.

[318] M. C. Teich, D. H. Johnson, A. R. Kumar, and R. G. Turcott, “Rate fluctuations and fractional power-law noise recorded from cells in the lower auditory pathway of the cat,” *Hearing Res.*, vol. 46, pp. 41–52, 1990.

[319] D. Terzopoulos, “Image analysis using multigrid relaxation methods,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 129–139, Mar. 1986.

[320] A. Tewfik and M. Kim, “Correlation structure of the discrete wavelet coefficients of fractional Brownian motion,” *IEEE Trans. Inform. Theory*, vol. 38, pp. 904–909, Mar. 1992.

[321] A. H. Tewfik, B. C. Levy, and A. S. Willsky, “Parallel smoothing,” *Syst. Contr. Lett.*, vol. 14, no. 3, pp. 253–259, Mar. 1990.

[322] K. E. Timmermann and R. D. Nowak, “Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging,” *IEEE Trans. Inform. Theory*, vol. 45, pp. 846–862, Apr. 1999.

[323] D. Tretter, C. A. Bouman, K. Khawaja, and A. Maciejewski, “A multiscale stochastic image model for automated inspection,” *IEEE Trans. Image Processing*, vol. 4, pp. 507–517, Dec. 1995.

[324] A. Tsai, J. Zhang, and A. S. Willsky, “EM algorithms for image processing using multiscale models and mean field theory, with applications to laser radar range profiling and segmentation,” *Opt. Eng.*, vol. 40, no. 7, pp. 1287–1301, July 2001.

[325] D. S. Tucker, “Multiresolution modeling from data and partial specifications,” Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, to be published.

[326] B. Tustison, D. Harris, and E. Foufoula-Georgiou, “Scale issues in verification of precipitation forecasts,” *J. Geophys. Res.—Atmos.*, vol. 106, no. D11, pp. 11 775–11 784, June 16, 2001.

[327] S. Ungarala and B. R. Bakshi, “A multiscale, Bayesian and error-in-variables approach for linear dynamic data rectification,” *Comput. Chem. Eng.*, vol. 24, pp. 445–451, July 15, 2000.

[328] G. C. Verghese and T. Kailath, “A further note on backward Markovian models,” *IEEE Trans. Inform. Theory*, vol. IT-25, pp. 121–124, 1979.

[329] M. Vetterli and J. Kovacevic, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[330] B. Vidakovic, “Nonlinear wavelet shrinkage with Bayes rule and Bayes factors,” *J. Amer. Stat. Assoc.*, vol. 93, pp. 173–179, 1998.

[331] S. Vignudelli, P. Cipollini, and M. Astraldi, “Integrated use of altimeter and *in situ* data for understanding the water exchanges between the Tyrrhenian and Ligurian Seas,” *J. Geophys. Res.—Oceans*, vol. 105, no. C8, pp. 19 649–19 663, Aug. 15, 2000.

[332] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, “Tree-based reparameterization framework for approximate estimation on graphs with cycles,” *IEEE Trans. Inform. Theory*, submitted for publication.

[333] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, “Random cascades on wavelet trees and their use in modeling natural images,” *Appl. Comput. Harmon. Anal.*, vol. 11, pp. 89–123, 2001.

[334] M. J. Wainwright, E. B. Sudderth, and A. S. Willsky, “Tree-based modeling and estimation of Gaussian processes on graphs with cycles,” *Neural Inform. Process. Syst.*, vol. 13, Dec. 2000.

[335] G. Wang, J. Zhang, and G.-W. Pan, “Solution of inverse problems in image processing by wavelet expansion,” *IEEE Trans. Image Processing*, pp. 579–593, May 1995.

[336] R. B. Washburn, W. W. Irving, J. K. Johnson, D. S. Avtgis, J. W. Wissinger, R. R. Tenney, and A. S. Willsky, “Multiresolution image compression and image fusion algorithms,” Alphatech, Inc., Tech. Rep., Feb. 1996.

[337] Y. Weiss, “Correctness of local probability propagation in graphical models with loops,” *Neural Comput.*, vol. 12, pp. 1–41, 2000.

[338] Y. Weiss and W. T. Freeman, “Correctness of belief propagation in Gaussian graphical models of arbitrary topology,” in *Proc. NIPS*, 2000, vol. 12, pp. 673–679.

[339] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. New York: Wiley, 1990.

[340] P. Whittle, “On stationary processes in the plane,” *Biometrika*, vol. 41, pp. 434–449, 1954.

[341] ——, “Stochastic processes in several dimensions,” *Bull. Int. Stat. Inst.*, vol. 40, pp. 974–994, 1963.

[342] C. K. Wikle, L. M. Berliner, and N. Cressie, "Hierarchical Bayesian space–time models," *Environ. Ecologic. Stat.*, vol. 5, no. 2, pp. 117–154, June 1998.

[343] A. S. Willsky, "Relationships between digital signal processing and control and estimation theory," *Proc. IEEE*, vol. 66, pp. 996–1017, Sept. 1978.

[344] J. W. Woods, "Two-dimensional discrete Markovian fields," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 232–240, 1972.

[345] J. W. Woods and C. H. Radewan, "Kalman filtering in two dimensions," *IEEE Trans. Inform. Theory*, vol. IT-23, pp. 473–482, 1977.

[346] G. Wornell, "Wavelet-based representations for the $1/f$ family of fractal processes," *Proc. IEEE*, vol. 81, pp. 1428–1450, Oct. 1993.

[347] G. Wornell and A. V. Oppenheim, "Estimation of fractal signals from noisy measurements using wavelets," *IEEE Trans. Signal Processing*, vol. 40, pp. 611–623, Mar. 1992.

[348] ——, "Wavelet-based representations for a class of self-similar signals with applications to fractal modulation," *IEEE Trans. Inform. Theory*, vol. 38, pp. 785–800, Mar. 1992.

[349] G. W. Wornell, *Signal Processing With Fractals. A Wavelet-Based Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[350] ——, "A Karhunen–Loève-like expansion for $1/f$ processes via wavelets," *IEEE Trans. Inform. Theory*, vol. 36, pp. 859–861, July 1990.

[351] C. H. Wu and P. C. Doerschuk, "Tree approximations to Markov random fields," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 391–402, Apr. 1995.

[352] C. Wunsch, "Low frequency variability of the sea," in *Evolution of Physical Oceanography: Scientific Studies in Honor of Henry Stommel*, B. A. Warren and C. Wunsch, Eds. Cambridge, MA: MIT Press, 1981.

[353] C. Wunsch and E. M. Gaposchkin, "On using satellite altimetry to determine the general circulation of the oceans with application to geoid improvement," *Revs. Geophys. Space Phys.*, vol. 18, pp. 725–745, 1980.

[354] C. Wunsch and D. Stammer, "The global frequency-wavenumber spectrum of oceanic variability estimated from TOPEX/POSEIDON altimetric measurements," *J. Geophys. Res.*, vol. 100, pp. 24 895–24 910, 1995.

[355] H.-C. Yang and R. Wilson, "Adaptive image restoration using a multiresolution Hopfield neural network," in *Proc. 5th Int. Conf. Image Processing and Its Applications*, Edinburgh, U.K., July 4–6, 1995, Inst. Elect. Eng. Conference Publication no. 410, pp. 198–202.

[356] J. C. Ye, C. A. Bouman, K. J. Webb, and R. P. Millane, "Nonlinear multigrid algorithms for Bayesian optical diffusion tomography," *IEEE Trans. Image Processing*, vol. 10, pp. 909–922, June 2001.

[357] J. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *Proc. NIPS*, 2001, vol. 13, pp. 689–695.

[358] B. A. Zeldin and P. D. Spanos, "Random field representation and synthesis using wavelet bases," *J. Appl. Mech. Trans. ASME*, vol. 63, no. 4, pp. 946–952, Dec. 1996.

[359] J. Zhang, "The mean field theory in EM procedures for Markov random fields," *IEEE Trans. Signal Processing*, vol. 40, pp. 2570–2583, Oct. 1992.

[360] J. Zhang and G. Walter, "A KL-like expansion for wide-sense stationary processes via a wavelet-like basis," *IEEE Trans. Signal Processing*, vol. 42, pp. 1737–1745, July 1992.

[361] S. Y. Zhao, "Wavelet filtering for filtered backprojection in computed tomography," *Appl. Comput. Harmon. Anal.*, vol. 6, pp. 346–373, 1999.

[362] W. Zhu, Y. Wang, Y. Deng, Y. Yao, and R. Barbour, "A wavelet-based multiresolution regularization least squares reconstruction approach for optical tomography," *IEEE Trans. Med. Imag.*, vol. 16, pp. 210–217, Apr. 1997.

**Alan Willsky** (Fellow, IEEE) joined the faculty of the Massachusetts Institute of Technology (MIT), Cambridge, in 1973 and is currently the Edwin Sibley Webster Professor of Electrical Engineering. He is a founder and member of the Board of Directors of Alphatech, Inc., and a member of the U.S. Air Force Scientific Advisory board. He has held visiting positions in England and France and various leadership positions in the IEEE Control Systems Society (which made him a Distinguished Member in 1988). He has delivered numerous keynote addresses and is coauthor of the undergraduate text *Signals and Systems* (Upper Saddle River, NJ: Prentice-Hall, 1997, 2nd ed.). His research interests are in the development and application of advanced methods of estimation and statistical signal and image processing. Methods he has developed have been successfully applied in variety of applications including failure detection, surveillance systems, biomedical signal and image processing, and remote sensing.

Dr. Willsky has received several awards including the 1975 American Automatic Control Council Donald P. Eckman Award, the 1979 ASCE Alfred Noble Prize, and the 1980 IEEE Browder J. Thompson Memorial Award.