March, 1991

LIDS- P 2022

Research Supported By:

AFOSR grant 88-0032
NSF grant ECS-8700903
ARO contract DAAL03-86-K-0171
INRIA

Modeling and Estimation of Multiresolution Stochastic Processes*

Basseville, M.
Benveniste, A.
Chou, K.C
Golden, S.A.
Nikoukhah,R.

MODELING AND ESTIMATION OF MULTIRESOLTUTION STOCHASTIC PROCESSES

Michele Basseville[1], Albert Benveniste[1], Kenneth C. Chou[2],
Stuart A. Golden[2], Ramine Nikoukhah[3] and Alan S. Willsky[2]

Abstract

In this paper, we provide an overview of the several components of a research effort aimed at the development of a theory of multiresolution stochastic modeling and associated techniques for optimal multiscale statistical signal and image processing. As we describe, a natural framework for developing such a theory is the study of stochastic processes indexed by nodes on lattices or trees in which different depths in the tree or lattice correspond to different spatial scales in representing a signal or image. In particular we will see how the wavelet transform directly suggests such a modeling paradigm. This perspective then leads directly to the investigation of several classes of dynamic models and related notions of "multiscale stationarity" in which scale plays the role of a time-like variable. In this paper we focus primarily on the investigation of models on homogeneous trees. In particular we describe the elements of a dynamic system theory on trees and introduce two notions of stationarity. One of these leads naturally to the development of a theory of multiscale autoregressive modeling including a generalization of the celebrated Schur and Levinson algorithms for order-recursive model building. The second, weaker notion of stationarity leads directly to a class of state space models on homogeneous trees. We describe several of the elements of the system theory for such models and also describe the natural, extremely efficient algorithmic structures for optimal estimation that these models suggest: one class of algorithms has a multigrid relaxation structure; a second uses the scale-to-scale whitening property of wavelet transforms for our models; and a third leads to a new class of Riccati equations involving the usual predict and update steps and a new "fusion" step as information is propagated from fine to coarse scales. As we will see, this framework allows us to consider in a very natural way the fusion of data from sensors with differing resolutions. Also, thanks to the fact that wavelet transforms do an excellent job of "compressing" large classes of covariance kernels, we will see that these modeling paradigms appear to have promise in a far broader context than one might expect.

[1]Institut de Recherche en Informatique et Systemes Aleatoires (IRISA), Campus de Beaulieu, 35042 Rennes, CEDEX, FRANCE. M.B. is also with the Centre National de la Recherche Scientifique (CNRS) and A.B. is also with the Institut National de Recherche en Informatique et en Automatique (INRIA). The research of these authors was also supported in part by Grant CNRS G0134.

[2]Laboratory for Information and Decision Systems and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge,MA 02139, USA. The work of these authors was also supported in part by the Air Force Office of Scientific Research under Grant AFOSR-88-0032, by the National Science Foundation under Grant ECS-8700903 and by the US Army Research Office under Contract DAAL03-86-K-0171. In addition some of this research was performed while KCC and ASW were visitors at IRISA and while ASW received support from INRIA.

[3]INRIA, Domaine de Voluceau, Rocquencourt, BP105, 78153 Le Chesnay, CEDEX, FRANCE.

1 Introduction

In recent years there has been considerable interest and activity in the signal and image processing community in developing multi-resolution processing algorithms. Among the reasons for this are the apparent or claimed computational advantages of such methods and the fact that representing signals or images at multiple scales is an evocative notion– it seems like a "natural" thing to do. One of the more recent areas of investigation in multiscale analysis has been the emerging theory of multiscale representations of signals and wavelet transforms [10, 21, 22, 23, 24, 28, 33, 34, 38, 49]. This theory has sparked an impressive flurry of activity in a wide variety of technical areas, at least in part because it offers a common unifying language and perspective and perhaps the promise of a framework in which a rational methodology can be developed for multiscale signal processing, complete with a theoretical structure that pinpoints when multiresolution methods might be useful and why.

It is important to realize, however, that the wavelet transform by itself is not the only element needed to develop a methodology for signal analysis. To understand this one need only look to another orthonormal transform, namely the Fourier transform which decomposes signals into its frequency components rather than its components at different resolutions. The reason that such a transform is useful is that its use simplifies the description of physically meaningful classes of signals and important classes of transformations of those signals. In particular stationary stochastic processes are *whitened* by the Fourier transform so that individual frequency components of such a process are statistically uncorrelated. Not only does this greatly simplify their analysis, but, it also allows us to deduce that frequency-domain operations such as Wiener or matched filtering–or their time domain realizations as linear shift-invariant systems–aren't just convenient things to do. *They are in fact the right–* i.e., the statistically optimal– *things to do.* In analogy, what is needed to complement wavelet transforms for the construction of a rational framework for multi-resolution signal analysis is the identification of a rich class of signals and phenomena whose description is simplified by wavelet transforms. Having this, we then have the basis for developing a methodology for *scale domain* filtering and signal processing, for

deducing that such operations are indeed the right ones to use, and for developing a new and potentially powerful set of insights and perspectives on signal and image analysis that are complementary to those that are the heritage of Fourier.

In this paper we describe the several components of our research into the development of a theory for multiresolution stochastic processes and models aimed at achieving the objectives of describing a rich class of phenomena and of providing the foundation for a theory of optimal multiresolution statistical signal processing. In developing this theoretical framework we have tried to keep in mind the three distinct ways in which multi-resolution features can enter into a signal or image analysis problem. First, the phenomenon under investigation may possess features and physically significant effects at multiple scales. For example, fractal models have often been suggested for the description of natural scenes, topography, ocean wave height, textures, etc. [5, 35, 36, 41]. Also, anomalous broadband transient events or spatially-localized features can naturally be thought of as the superposition of finer resolution features on a more coarsely varying background. As we will see, the modeling framework we describe is rich enough to capture such phenomena. For example, we will see that 1/f -like stochastic processes as in [50, 51] are captured in our framework as are surprisingly useful models of many other processes. Secondly, whether the underlying phenomenon has multi-resolution features or not, it may be the case that the <u>data</u> that has been collected is at several different resolutions. For example the resolutions of remote sensing devices operating in different bands– such as IR, microwave, and various band radars– may differ. Furthermore, even if only one sensor type is involved, measurement geometry may lead to resolution differences (for example, if zoomed and un-zoomed data are to be fused or if data is collected at different sensor-to-scene distances). As we will see, the framework we describe provides a natural way in which to design algorithms for such multisensor fusion problems.

Finally, whether the phenomenon or data have multi-resolution features or not, the signal analysis *algorithm* may have such features motivated by the two principal manifestations of the at least superficially daunting complexity of many image processing problems. The first and more well-known of these is the use of multi-resolution algorithms to combat the computational demands of such problems by solving coarse

(and therefore computationally simpler) versions and using these to guide (and hopefully speed up) their higher resolution counterparts. Multigrid relaxation algorithms for solving partial differential equations are of this type as are a variety of computer vision algorithms. As we will see, the stochastic models we describe lead to several extremely efficient computational structures for signal processing.

The second and equally important issue of complexity stems from the fact that a multi-resolution formalism allows one to exercise very direct control over "greed" in signal and image reconstruction. In particular, many imaging problems are, in principle, ill-posed in that they require reconstructing more degrees of freedom then one has elements of data. In such cases one must "regularize" the problem in some manner, thereby guaranteeing accuracy of the reconstruction at the cost of some resolution. Since the usual intuition is precisely that one should have higher confidence in the reconstruction of lower resolution features, we are led directly to the idea of reconstruction at multiple scales, allowing the resolution-accuracy tradeoff to be confronted directly. As we will see the algorithms arising in our framework allow such multi-scale reconstruction and provide the analytical tools both for assessing resolution versus accuracy and for correctly accounting for fine scale fluctuations as a source of "noise" in coarser scale reconstructions.

While there are several ways in which to introduce and motivate our modeling framework, one that provides a fair amount of insight begins with the wavelet tranforms. However, the key for modeling is not to view the transform as a method for analyzing signals but rather as a mechanism for *synthesizing* or *generating* such signals beginning with coarse representations and adding fine detail one scale at a time. Specifically let us briefly recall the structure of multiscale representations associated with orthonormal wavelet transforms [22, 33]. For simplicity we do this in the context of $1-D$ signals (i.e. signals with one independent variable), but the extension to multidimensional signals and images introduces only notational rather than mathematical complexity.

The multiscale representation of a continuous signal $f(x)$ consists of a sequence of approximations of that signal at finer and finer scales where the approximations of $f(x)$ at the mth scale consists of a weighted sum of shifted and compressed (or

dilated) versions of a basic scaling function $\phi(x)$:

$$f_m(x) = \sum_{n=-\infty}^{+\infty} f(m,n)\phi(2^m x - n) \tag{1.1}$$

In order for the $(m+1)$st approximation to be a refinement of the mth, we require $\phi(x)$ to be representable at the next scale:

$$\phi(x) = \sum_n h(n)\phi(2x - n) \tag{1.2}$$

As shown in [22], $h(n)$ must satisfy several conditions for (1.1) to be an orthonormal series and for several other properties of the representation to hold. In particular $h(n)$ must be the impulse response of a quadrature mirror filter (QMF) [22, 44]. The simplest example of such a ϕ, h pair is the Haar approximation with

$$\phi(x) = \begin{cases} 1 & 0 \le x < 1 \\ 0 & otherwise \end{cases} \tag{1.3}$$

and

$$h(n) = \begin{cases} 1 & n = 0, 1 \\ 0 & otherwise \end{cases} \tag{1.4}$$

By considering the incremental detail added in obtaining the $(m+1)$st scale approximation from the mth, we arrive at the wavelet transform. Such a transform is based on a single function $\psi(x)$ that has the property that the full set of its scaled translates $\left\{2^{m/2}\psi(2^m x - n)\right\}$ form a complete orthonormal basis for L^2. In [22] it is shown that ϕ and ψ are related via an equation of the form

$$\psi(x) = \sum_n g(n)\phi(2x - n) \tag{1.5}$$

where $g(n)$ and $h(n)$ form a *conjugate mirror filter* pair [44], and that

$$f_{m+1}(x) = f_m(x) + \sum_n d(m,n)\psi(2^m x - n) \tag{1.6}$$

Thus, $f_m(x)$ is simply the partial orthonormal expansion of $f(x)$, up to scale m, with respect to the basis defined by ψ. For example if ϕ and h are as in (1.3), (1.4), then

$$\psi(x) = \begin{cases} 1 & 0 \le x < 1/2 \\ -1 & 1/2 \le x < 1 \\ 0 & otherwise \end{cases} \tag{1.7}$$

$$g(n) = \begin{cases} 1 & n = 0 \\ -1 & n = 1 \\ 0 & otherwise \end{cases} \tag{1.8}$$

and $\left\{2^{m/2}\psi(2^m x - n)\right\}$ is the *Haar basis.*

One of the appealing features of the wavelet transforms for the analysis of signals is that they can be computed recursively in scale, from fine to coarse. Specifically, if we have the coefficients $\{f(m+1,\cdot)\}$ of the $(m+1)$st-scale representation we can "peel off" the wavelet coefficients at this scale and at the same time carry the recursion one complete step by calculating the coefficients $\{f(m,\cdot)\}$ at the next somewhat coarser scale:

$$f(m,n) = \sum_k h(2n-k) f(m+1,k) \tag{1.9}$$

$$d(m,n) = \sum_k g(2n-k) f(m+1,k) \tag{1.10}$$

Reversing this process we obtain the synthesis form of the wavelet transform in v hich we build up finer and finer representations via a coarse-to-fine scale recursion:

$$f(m+1,n) = \sum_k h(2k-n) f(m,k) + \sum_k g(2k-n) d(m,k) \tag{1.11}$$

Thus we see that the synthesis form of the wavelet transform defines a *dynamical* relationship between the coefficients $f(m,n)$ at one scale and those at the next. Indeed this relationship defines a lattice on the points (m,n), where $(m+1,k)$ is connected to (m,n) if $f(m,n)$ influences $f(m+1,k)$. The simplest example of such a lattice is the dyadic tree illustrated in Figure 1, where each node t corresponds to a particular scale/shift pair (m,n). As with all these lattices, the scale index is indeed time-like, with each level of the tree corresponding to a representation of signals or phenomena at a particular scale. In this paper we focus for the most part on this tree structure and on dynamic models and stochastic processes defined on it[1]. Note that this setting has a natural association with the Haar transform in which the value at a particular

[1]In Sections 4 and 5 we briefly describe some aspects of the more general case.

node $t = (m, n)$ is obtained from the average of the values at the two descendant nodes $(m+1, 2n)$ and $(m+1, 2n+1)$. However, while the Haar transform indeed plays an important role in our analysis, the dyadic tree and the pyramidal structure it captures should be viewed in a broader sense as providing a natural setting for capturing representations of signals at multiple resolutions where the relationships between the representations at different resolutions need not be constrained to the rigid equalities in (1.9) - (1.11). Rather, if we view these multiscale representations more abstractly, much as in the notion of state, as capturing the features of signals up to a particular scale that are relevant for the "prediction" of finer-scale approximations, we can define rich classes of stochastic processes and models that contain the multiscale wavelet representations of (1.9) - (1.11) as special (and in a sense degenerate) cases.

Carrying this a bit farther, let us return to the point made earlier that for wavelet transforms to be useful it should be the case that their application simplifies the description or properties of signals. For example, this clearly would be the case for a stochastic process that is whitened by (1.9), (1.10), i.e. for which the wavelet coefficients $\{d(m, \cdot)\}$ at a particular scale are white and uncorrelated with the lower resolution version $\{f(m, \cdot)\}$ of the signal. In this case (1.11) represents a first-order recursion in scale that is driven by white noise. However, as we know from time series analysis, white-noise-driven first-order systems yield a comparatively small class of processes which can be broadened considerably if we allow higher-order dynamics. Also, in sensor fusion problems one wishes to consider collectively an entire set of signals or images from a suite of sensors. In this case one is immediately confronted with the need to use higher-order models in which the actually observed signals may represent samples from such a model at *several* scales, corresponding to the differing resolutions of individual sensors.

In this paper we describe two stochastic modeling paradigms for multiresolution processes that have as their motivation the preceding observations as well as the desire to investigate and develop multiscale counterparts to the notions of stationarity and rationality that have proven to be of such value in time series analysis. The first step in doing this is the introduction of dynamics and concepts of shift-invariance on dyadic trees, and in the next section we outline the elements of this formalism and

in particular introduce two notions of (second-order) shift-invariance for stochastic processes on dyadic trees. In Section 3 we then use the stronger of these two notions to develop a theory of multiscale autoregressive modeling and in particular we describe a generalization of the celebrated Schur and Levinson algorithms for the efficient construction of such models. Figure 2 illustrates the output of a third-order model of this type displaying some of the fractal-like, multi-scale characteristics that can be captured by this class of models. An alternate modeling paradigm—coinciding with that of Section 3 only for first-order models—is described in Section 4. This formalism, which generalizes finite-dimensional state models to dyadic trees, also can be used to capture fractal-like behavior and indeed includes the 1/f-like models developed in [50, 51] as a special case. Moreover these models provide surprisingly accurate descriptions of a broad variety of stochastic processes and also lead to extremely efficient and highly parallelizable algorithms for optimal estimation and for the fusion of multiresolution measurements using multiscale, scale-recursive generalizations of Kalman filtering and smoothing. For example, Figure 3(a) illustrates the sample path of a process with a 1/f-like spectrum and its optimal estimation based on noisy measurements of the process collected only at the two ends of the data interval. Figure 3(b) illustrates the use of our methodology for the estimation of the process based on these noisy data augmented with coarser resolution measurements– i.e. the formalism we describe allows us, with relative ease, to use coarse scale data to optimally guide the interpolation of fine-scale but sparsely-collected data. Figures 3(c) and 3(d) display analogous results for the case of a standard Gauss-Markov process in which an approximate multiscale model for this process is used to design the coarse/fine data fusion and interpolation algorithm.

Due to the limitations of space our presentation of the various topics we have mentioned is of a summary nature. References to complete treatments are given, and, in addition, in Section 5 we briefly discuss several important issues, current lines of investigation, and open questions.

2 Stochastic Processes and Dynamic Models on Dyadic Trees

In this section we introduce the machinery needed for specifying linear models of random processes on the dyadic tree, that is for stochastic processes y_t where t is an element of the set of nodes, $\mathcal{T}$, of the tree of Figure 1. As indicated in the introduction, we have several objectives in developing such models. Our first objective is to introduce models that can be specified by finitely many parameters in order to provide associated effective algorithms. That is, we would like to develop models analogous to those specified by finite-order difference equations or finite-dimensional state models– i.e. those corresponding to rational system functions– which have provided the setting for a vast array of powerful methods of signal and system analysis. Also, recursive models of this type are naturally associated with a notion of causality. In our context we will also seek recursive structures where the associated notion of causality will be in scale, from coarse to fine as in the wavelet transform synthesis equation (1.11).

Finally, another notion from time series that we will want to adapt to our context is that of shift-invariance or stationarity. To understand what is involved in this, let us recall the usual notion of stationarity[2] for a discrete-time, zero-mean stochastic processes y_t, where in this case $t \epsilon Z$, the integers. Such a process, with covariance function

$$r_{t,s} = E[y_t y_s] \tag{2.1}$$

is stationary if $r_{t+n,s+n} = r_{t,s}$ for all integers n. That is, shifting the time index of the process by n leaves the statistics invariant. Since it is also obviously time that $r_{s,t} = r_{t,s}$, we can immediately deduce that

$$r_{s,t} = r_{d(s,t)} \tag{2.2}$$

where $d(s,t) = |t - s|$.

[2]In this paper we focus completely on linear models and second-order properties, which, of course, yield complete descriptions if the processes considered are Gaussian.

In order to understand how we might generalize these ideas to the dyadic tree, we need to make several observations. The first is that the integers Z and our dyadic tree are both examples of homogeneous trees. Specifically a homogeneous tree of multiplicity q is an infinite acyclic graph such that each node has exactly q+1 branches to other nodes representing its neighbors. In the case of Z, $q = 1$, and the neighbors of an integer t are simply $t - 1$ and $t + 1$. For the case of $\mathcal{T}$, $q = 2$. However, Figure 1 isn't the easiest way in which to see this or to understand notions of stationarity. Specifically, in considering the usual notion of stationarity we are compelled to consider processes defined on all of Z, and the same is true in our context. Thus, we must be able to extend our tree in all directions capturing in particular the fact that there is neither a finest nor a coarsest scale of description. A much more convenient representation of $\mathcal{T}$ that allows such extensions is depicted in Figure 4. As we will see, both Figures 1 and 4 will prove of use to us.

An important fact about trees is that there is a natural notion of distance $d(s,t)$ between two nodes, s and t, namely the number of branches on the path from s to t, which reduces to $|t - s|$ for Z. This allows us to define the notion of an isometry, that is a one-to-one and onto map of the tree onto itself that preserves distances. For Z the only isometries are shifts, $t \longmapsto t + n$ i.e. and reversals, i.e. $t \longmapsto -t$ (and concatenations of these), so that a useful way (for us!) in which to define the usual notion of stationarity is that the statistics of the process are invariant under any isometry on the index set, i.e. $r_{t,s} = r_{\tau(t),\tau(s)}$ for any isometry.

It is this type of notion that we seek to generalize to the dyadic tree. However, the tree $\mathcal{T}$ has many isometries. For example consider an isometry pivoting on the node denoted "$s \wedge t$" in Figure 4, where all nodes below and to the right of this point are left unchanged but the upper left-hand portion of the tree is "flipped" in that the two branches extending from $s \wedge t$ are interchanged (so that, for example, u is mapped into s). Obviously we can do the same thing pivoting at any node. We refer the reader to [14] for complete treatments of the nature and structure of isometries.

The preceding discussion suggests a first notion of shift-invariance for a stochastic process y_t which we refer to as isotropy. Specifically y_t is an isotropic process if its statistics remain invariant under any isometry on the index set. As shown in [3, 6, 7, 8]

y_t is isotropic if any only if its covariance $r_{t,s}$, as defined in (2.1) (with $t, s \epsilon \mathcal{T}$), satisfies (2.2). Thus, as with a standard temporally-stationary process, an isotropic process on $\mathcal{T}$ is characterized by a covariance sequence $r_0, r_1, r_2, \ldots$ and, as in the standard case we have two natural questions: (1) when does such a sequence of numbers correspond to a valid covariance sequence for a process on $\mathcal{T}$; and (2) how can we construct dynamic models for the construction of an isotropic process corresponding to such a valid sequence. A first form of the answer to the first question can actually be stated a bit more generally. Specifically, if S is any index set, and if $\{y_t, t \epsilon S\}$ is a zero-mean process defined on S then its covariance $r_{s,t}$ must satisfy the following : select an arbitrary finite family $\{t_i\}_{i=1,\ldots,I}$ in S; then the $I \times I$ matrix whose (i,j)-element is r_{t_i,t_j} must be non-negative definite since

$$\text{cov} \begin{bmatrix} y_{t_1} \\ \ldots \\ y_{t_I} \end{bmatrix} = \left[r_{t_i,t_j} \right]_{i,j=1,\ldots,I} \tag{2.3}$$

This property of r, which is necessary and sufficient for it to be the covariance of such a process, will be referred to as *positive definiteness* in the sequel. For general index sets it is not possible to find more useful criteria or characterizations of positive definiteness. However for stationary time series, i.e. for $S = Z$ and $r_{t,s}$ satisfying (2.2) much more can be said. In particular the celebrated Bochner spectral representation theorem states that a sequence $r_n, n = 0, 1, \ldots$ is the covariance function of a stationary time series if and only if there exists a nonnegative, symmetric spectral measure $S(d\omega)$ so that

$$r_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{j\omega n} S(d\omega)$$

As shown in [2, 3] there is a corresponding generalized Bochner theorem for a sequence r_n to be the covariance of an isotropic process on $\mathcal{T}$. Note that we can obviously find a subset of $\mathcal{T}$ isomorphic to Z — i.e. a sequence of nodes extending infinitely in both directions, and y_t restricted to such a set is essentially a temporally-stationary process. Thus for r_n to be a valid covariance of an isotropic process on $\mathcal{T}$ it must certainly be a valid covariance for a temporally-stationary process. However

there are additional constraints for isotropic processes– for example in $\mathcal{T}$ we can find three nodes which are all a distance two from one another (e.g. u, v, and $s \wedge t$ in Figure 4), and this implies an additional constraint on r_n. The impact of these additional constraints can be seen in the Bochner theorem in [2, 3] and also in the results described in the next section.

While the Bochner theorem is a powerful characterization result for time series and for processes on trees, it does not provide a computational procedure for testing positive definiteness or for constructing models for such processes. However for time series we do have such a method, namely the Wold representation of stationary processes via causal, autoregressive (AR) models. This representation and the well-known Levinson algorithm for its construction not only provide a procedure for testing positive-definiteness but also for constructing rational, finite-order models for stationary processes. The subject of Section 3 of this paper is the extension of this methodology to isotropic processes on trees. An important point in doing this is to realize that such a construction for time series produces a model that treats time asymmetrically (by imposing causality) in order to represent a process whose statistics do not have inherent temporal asymmetry. This is not a point that is typically highlighted since the geometry of Z is so simple. However the situation for $\mathcal{T}$ is decidedly more complex, and to carry out our program we need the following development which in essence relates the pictorial representations of Figures 1 and 4 and provides the basis for defining causal systems in scale.

An important concept associated with any homogeneous tree is the notion of a *boundary point* [2, 3, 6, 14, 15] of a tree. Consider the set of infinite sequences of nodes on such a tree, where any such sequence consists of a set of distinct nodes $t_1, t_2, \ldots$ where $d(t_i, t_{i+1}) = 1$. A boundary point is an equivalence class of such sequences where two sequences are equivalent if they differ by a finite number of nodes. For the case of Z there are two boundary points corresponding to paths toward $\pm\infty$. For $\mathcal{T}$ there are many. Let us choose one boundary point in $\mathcal{T}$ which we denote by $-\infty$. Note that from any node t there is a unique path in the equivalence class defined by $-\infty$ (i.e. a unique path from t "towards" $-\infty$ - see Figure 4). Then if we take any two nodes s and t, their paths to $-\infty$ must differ only by a finite number of points

and thus must meet at some node which we denote by $s \wedge t$ (see Figure 4). Thus, we can define a notion of *relative distance* of two nodes to $-\infty$:

$$\delta(s,t) = d(s, s \wedge t) - d(t, s \wedge t) \tag{2.4}$$

so that

$$s \preceq t \text{ ("s is at least as close to } -\infty \text{ as t") if } \delta(s,t) \leq 0 \tag{2.5}$$

$$s \prec t \text{ ("s is closer to } -\infty \text{ than t") if } \delta(s,t) < 0 \tag{2.6}$$

This also yields an equivalence relation on nodes of $\mathcal{T}$:

$$s \asymp t \leftrightarrow \delta(s,t) = 0 \tag{2.7}$$

For example, the points s, v, and u in Figure 4 are all equivalent. The equivalence classes of such nodes are referred to as *horocycles*. These equivalence classes are best visualized as in Figure 1 by redrawing the tree, in essence by picking the tree up at $-\infty$ and letting the tree "hang" from this boundary point. In this case the horocycles appear as points on the same horizontal level and $s \preceq t$ means that s lies on a horizontal level above or at the level of t. Note that in this way we make explicit the dyadic structure of the tree as depicted in Figure 1 and provide the basis for defining multiscale dynamic models.

In order to define dynamics on trees, let us again step back to take a more careful look at the usual formalism that is used for time series. Specifically, in specifying a temporal system in terms of a difference equation we make essential use of the notion of shifts or moves – e.g. in an AR model we relate y_t to y_{t-1}, y_{t-2}, *etc.* where the backward shift $z^{-1} : t \longmapsto t-1$ obviously plays an essential role in expressing the "local" dynamics, i.e. the relationship of a signal at a particular point to its values at nearby points. Moreover, thanks to the simple structure of Z, we have the luxury of using the symbol z^{-1} for two additional purposes. In particular, the backward shift z^{-1} is an isometry and in fact it and its inverse, the forward shift, generate all translations. Furthermore,we also use the symbol z^{-1} and its positive and negative powers to code signals— i.e. we represent the signal y_t by its z-transform – and all of these properties provides us with the powerful transform domain formalism for analyzing stationary, i.e. translation-invariant systems.

The situation is decidedly more complex on $\mathcal{T}$. To see this let us begin by defining moves on $\mathcal{T}$ that will be needed to provide a "calculus" for stochastic processes, i.e. for specifying local dynamics. Such moves are illustrated in Figure 1 and are introduced next :

- 0 the identity operator (no move)
- $\overline{\gamma}$ the backward shift (move one step toward $-\infty$)
- α the left forward shift (move one step away from $-\infty$ toward the left)
- β the right forward shift (move one step away from $-\infty$ toward the right)
- δ the interchange operator (move to the nearest point in the same horocycle)

Note that the richer structure of $\mathcal{T}$ requires a richer collection of moves. Also, unlike its counterpart z^{-1}, the backward shift $\overline{\gamma}$ is <u>not</u> an isometry (it is onto but not one-to-one), and it has <u>two</u> forward shift counterparts, α and β, which are one-to-one but not onto. Also, while these shifts allow us to move up and down in scale, (i.e. from one horocycle to the next), it is necessary to introduce another operator, δ, in order to define purely translational shifts at a given level. Note also that 0 and δ are isometries and that these operators satisfy the following relations (where the convention is that the <u>left-most</u> operator is applied first)[3]:

$$\alpha\overline{\gamma} = \beta\overline{\gamma} = 0 \tag{2.8}$$

$$\delta\overline{\gamma} = \overline{\gamma} \tag{2.9}$$

$$\delta^2 = 0 \tag{2.10}$$

$$\beta\delta = \alpha \tag{2.11}$$

Arbitrary moves on the tree can then be encoded via finite strings or *words* using these symbols as the alphabet and the formulas (2.8)–(2.11). Specifically define the language

$$\mathcal{L} = (\overline{\gamma})^* \cup (\overline{\gamma})^*\delta\{\alpha,\beta\}^* \cup \{\alpha,\beta\}^* \tag{2.12}$$

[3]Our convention will be to write operators on the right, e.g. $t\alpha, t\delta\beta$

where K^* denotes arbitrary sequences of symbols in K including the empty sequence which we identify with the operator 0. Then any move on $\mathcal{T}$ is uniquely represented by a word of this language. It is straightforward to define a *length* $|w|$ for each word in $\mathcal{L}$, corresponding to the number of shifts required in the move specified by w. Note that

$$|\overline{\gamma}| = |\alpha\ | = |\beta| = 1$$
$$|0| = 0 \quad , \quad |\delta| = 2 \tag{2.13}$$

Thus $|\overline{\gamma}^n| = n, |w_{\alpha\beta}|$ = the number of α's and β's in $w_{\alpha\beta} \in \{\alpha, \beta\}^*$, and $|\overline{\gamma}^n \delta w_{\alpha\beta}| = n + 2 + |w_{\alpha\beta}|$. This notion of length will be useful in defining the *order* of dynamic models on $\mathcal{T}$. We will also be interested exclusively in *causal* models, i.e. in models in which the output at some scale (horocycle) does not depend on finer scales. For this reason we are most interested in moves that either involve pure ascents on the tree, i.e. all elements of $\{\overline{\gamma}\}^*$, or elements $\overline{\gamma}^n \delta w_{\alpha\beta}$ of $\{\overline{\gamma}\}^* \delta \{\alpha, \beta\}^*$ in which the descent is no longer than the ascent, i.e. $|w_{\alpha\beta}| \leq n$. We use the notation $w \preceq 0$ to indicate that w is such a causal move. Note that we include moves in this causal set that are not strictly causal in that they shift a node to another on the *same* horocycle. We use the notation $w \asymp 0$ for such a move. The reasons for this will become clear when we examine autoregressive models.

Also, on occasion we will find it useful to use a simplified notation for particular moves. Specifically, we define $\delta^{(n)}$ recursively, starting with $\delta^{(1)} = \delta$ and

$$\text{If } t = t\overline{\gamma}\alpha, \text{ then } t\delta^{(n)} = t\overline{\gamma}\delta^{(n-1)}\alpha$$
$$\text{If } t = t\overline{\gamma}\beta, \text{ then } t\delta^{(n)} = t\overline{\gamma}\delta^{(n-1)}\beta \tag{2.14}$$

What $\delta^{(n)}$ does is to map t to another point on the same horocycle in the following manner: we move up the tree n steps and then descend n steps; the first step in the descent is the opposite of the one taken on the ascent, while the remaining steps are the same. That is if $t = t\overline{\gamma}^{n-1} w_{\alpha,\beta}$ then $t\delta^{(n)} = t\overline{\gamma}^{n-1}\delta w_{\alpha\beta}$. For example, referring to Figure 1, $s = u\delta^{(2)}$.

The preceding development provides us with the move structure required for the specification of local dynamics on trees. Let us turn next to the specification of "shift-invariant" systems and processes. The most general linear input/output relationship

for signals defined on the tree is simply

$$y_t = \sum_{s \in \mathcal{T}} h_{t,s} u_s \triangleq (Hu)_t \tag{2.15}$$

As with temporal systems, one would expect the requirements of various notions of shift-invariance to impose constraints on the weighting coefficients $h_{t,s}$. To see this let us first adopt an abuse of notation commonly used for time series. Specifically, if τ is an isometry of $\mathcal{T}$, we use the same notation to denote an operation on signals over $\mathcal{T}$, i.e.

$$\tau(y)_t = y_{\tau(t)} \tag{2.16}$$

(analogous to $z^{-1}y_t = y_{t-1}$). A first, rather strong notion of shift-invariance might be that if $\tau(u)$ is applied to the system for any isometry τ, then the output is $\tau(y)$, where y is the response to u. It is not difficult to check that for this to be the case we must have that

$$h_{t,s} = h(d(s,t)) \tag{2.17}$$

Note, however, that this is an exceedingly strong condition and indeed generalizes the notion of zero-phase LTI systems, i.e. systems with impulse responses such that $h(t,s) = h(|t-s|)$. Such systems obviously are not causal, and in fact are far too constrained in that they require invariance to too many isometries. In particular such an LTI system has the property that it is not only translation-invariant but also reversal invariant (i.e. $u(-t)$ yields $y(-t)$). In the case of time series we overcome this by using the smaller group of isometries generated by the shift z^{-1}. On $\mathcal{T}$, however, the shifts $\overline{\gamma}, \alpha$, and β are not isometries. For this reason it is necessary to introduce a subgroup of isometries of $\mathcal{T}$ corresponding to the other role played by z^{-1}, that of defining backward, causal, translations.

Specifically, let $(t_n)_{n \in Z}$ denote an infinite path extending in $\mathcal{T}$ back toward $-\infty$ (as $n \to -\infty$). A (one step) translation with skeleton (t_n) is an isometry of $\mathcal{T}$ that has the property that

$$\tau(t_n) = t_{n+1} \tag{2.18}$$

Since there are many such paths (t_n) there obviously are many translations, and indeed for any particular (t_n) there are numerous translations (see Figure 5). Never-

theless the class of translations represents a proper subset of all isometries, and does allow us to define a very useful notion of shift invariance:

Definition 1 (stationary systems) *A linear system H as in (2.15), acting on signals on $\mathcal{T}$, is said to be a stationary system if* [4]

$$H \circ \tau = \tau \circ H \tag{2.19}$$

for any translation τ.

A fundamental result proven in [9] is that H is stationary if and only if its weighting pattern satisfies.

$$h_{t,s} = h\left[d(t, s \wedge t), d(s, s \wedge t)\right] \tag{2.20}$$

Thus a stationary system is specified by a 2-D sequence $h(n,m), n, m \geq 0$ and, referring to Figure 1, we see that (2.20) has an intuitively appealing interpretation. Specifically $s \wedge t$ denotes the "parent" node of s and t, i.e. the finest scale node that has both s and t as descendants, and (2.20) states that $h_{t,s}$ depends only on the distances in scale from this parent node to s and to t. Roughly speaking the influence of the input at node s on the output at node t in a stationary system depends on the differences in scale and in temporal offset of the scale/shift pairs represented by t and s.

Obviously, a system satisfying (2.17) (and thus corresponding to a system that commutes with all isometries) also satisfies (2.20) (this is easily seen since $d(s,t) = d(s, s \wedge t) + d(t, s \wedge t)$). The reverse is certainly not true indicating that we have a far larger class of stationary systems as defined in Definition 1. Similarly, we can define a larger class of shift-invariant processes:

Definition 2 (stationary stochastic processes) *A zero mean (scalar) stochastic process y is said to be stationary if its covariance function is translation–invariant, i.e.*

$$r_{s,t} = r_{\tau(s),\tau(t)} \tag{2.21}$$

for any translation τ.

[4] $\circ$ denotes the composition of maps.

As shown in [9] a process is stationary if and only if

$$r_{s,t} = r[d(s, s \wedge t), d(t, s \wedge t)] \tag{2.22}$$

Thus a stationary process is specifies by a 2-D sequence $r(n, m), n, m \geq 0$. Also isotropic processes— i.e. processes for which (2.21) is satisfied for all isometries and for which (2.2) holds– are obviously stationary, but the reverse implication is not true, so that stationary processes represent a richer class of processes. Furthermore the covariance structure (2.22) in essence says that the statistical relationship between the values of a stationary process at two nodes depends on the differences in scale and in temporal offset of the two nodes. In particular from (2.22) it follows that the statistical behavior of the restriction of a stationary process to any scale (i.e. horocycle) does not depend on the scale, indicating that the concept of stationarity on the tree appears to be a natural and convenient one for capturing a notion of statistical self-similarity. Moreover, as we will see, the Haar transform yields the eigenstructure of the process at any scale, providing another tie back to wavelet transforms. In Section 4, we expand on these and related points in the context of the investigation of a class of finite-dimensional state models on dyadic trees that, in the constant-coefficient case, provides us with the class of rational linear systems satisfying the notion of stationarity we have introduced.

Let us close this discussion with a few comments. First, as shown in [9], the notions of systems and stochastic stationarity introduced in Definitions 1 and 2 are compatible in the sense that the output of a stationary system driven by a stationary input is itself stationary. In general, however, an isotropic process driving an arbitrary stationary system does not yield an isotropic output, and thus we might expect that we will have to work harder to pinpoint the class of systems that does generate isotropic processes. Furthermore, as we have indicated we are interested in constructing causal models, i.e. systems as in (2.15) with

$$h_{t,s} = 0 \text{ for } t \prec s \tag{2.23}$$

For stationary systems this corresponds to requiring

$$h(d(t, s \wedge t), d(s, s \wedge t)) = 0 \text{ for } d(t, s \wedge t) < d(s, s \wedge t) \tag{2.24}$$

Finally, let us make a brief comment about the generalization of the third use of z^{-1}, namely to define transforms. Specifically, as discussed in [6, 7, 8, 9], natural objects to consider in this context are noncommutative formal power series of the form:

$$S = \sum_{w \in \mathcal{L}} s_w \cdot w \tag{2.25}$$

We will use such tranforms in the next section in order to encode correlation functions in our generalization of the Schur recursions. In addition transforms of this type can be used to encode convolutional systems. Specifically, we can think of (2.25) as defining the system function of a system in the following manner: if the input to this system is $u_t, t \in \mathcal{T}$, then the output is given by the generalized convolution:

$$(Su)_t = \sum_{w \in \mathcal{L}} s_w u_{tw} \tag{2.26}$$

Note that in this context causality corresponds to $s_w = 0$ for all $0 \prec w$. Also it is important to realize that while (2.25), (2.26) would seem to correspond to a general class of shift-invariant systems, both classes of systems we have described– stationary and isotropic– require further restrictions. In particular for S in (2.25), (2.26) to be stationary we must have that if $\omega = \overline{\gamma}^n \delta \omega_{\alpha\beta}$, then s_ω depends only on n and $|\omega_{\alpha\beta}|$. Similarly, S is isotropic if s_ω depends only on $|w|$. Finally, for future reference we use the notation $S(0)$ to denote the coefficient of the empty word in S. Also it will be necessary for us to consider particular shifted versions of S:

$$\overline{\gamma}[S] = \sum_{w \in \mathcal{L}} s_{w\overline{\gamma}} \cdot w \tag{2.27}$$

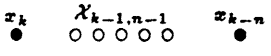
$$\delta^{(k)}[S] = \sum_{w \in \mathcal{L}} s_{w\delta^{(k)}} \cdot w \tag{2.28}$$

where we use (2.8)–(2.11) and (2.14) to write $w\overline{\gamma}$ and $w\delta^{(k)}$ as elements of $\mathcal{L}$. Notice that, because of the relations (2.8)–(2.11), the operators $S \longrightarrow \overline{\gamma}[S]$ and $S \longrightarrow \delta[S]$ can *not* be thought of as multiplication operators on formal power series.

respectively. The formulae

$$\begin{aligned} e_{k,n+1} &= x_k - E\{x_k|\mathcal{X}_{k-1,n}\} \\ &= x_k - E\{x_k|\mathcal{X}_{k-1,n-1}\} \\ &\quad + E\{x_k|\mathcal{X}_{k-1,n-1}\} - E\{x_k|\mathcal{X}_{k-1,n}\} \\ &= e_{k,n} - E\{x_k|\mathcal{X}_{k-1,n} \ominus \mathcal{X}_{k-1,n-1}\}\} \\ &= e_{k,n} - E\{e_{k,n}|f_{k-1,n}\}] \\ &= e_{k,n} - k_n f_{k-1,n} \end{aligned} \tag{3.5}$$

where $\mathcal{U} \ominus \mathcal{V}$ denotes the orthogonal complement of $\mathcal{V}$ in $\mathcal{U}$, show that the key to the calculation of the $(n+1)$st-order prediction error $e_{k,n+1}$ is the computation of the prediction of the forward residual $e_{k,n}$ given the backward one $f_{k-1,n}$. Similarly, the prediction of the backward residual given the forward one is needed for the calculation of backward residuals of increasing order. It is a remarkable property of stationary time series that *both prediction operators are identical*, i.e. that the same coefficient k_n in (3.5) also appears in the corresponding equation for the backward residual. This fact, which then leads to the celebrated Levinson recursions, stems from the fact that the statistics of a stationary time series are invariant under the isometry $k \longmapsto -k$. The correlation coefficient k_n of the two involved residuals is also known as the *PARCOR* coefficient of x_k and x_{k-n} given $\mathcal{X}_{k-1,n-1}$. This is illustrated in the following diagram :



Since $e_{k,0} = f_{k,0} = x_k$, we find that (3.5) and the associated Levinson recursion provide us with a method for constructing models for x_n of increasing order. In particular,if $e_{k,n}$ and $f_{k,n}$ are white, (so that all higher-order PARCOR coefficients are 0), we obtain an nth order AR model for x_n constructed in lattice form, i.e. one first-order section (specified by one PARCOR coefficient) at a time.

Let us now consider the extension of these ideas to the dyadic tree. As one might expect from the preceding discussion of AR(2) and as developed in detail in [6, 7, 8], construction of models of increasing order requires the consideration of vectors of forward and backward residuals of dimension that increases with model order. To

begin, let y_t be an isotropic process on a tree, and define the (nth-order) past of the node t on $\mathcal{T}$:

$$\mathcal{Y}_{t,n} \stackrel{\Delta}{=} \mathcal{H}\{y_{tw} : w \preceq 0, |w| \leq n\} \tag{3.6}$$

In analogy with the time series case, the backward innovations or prediction error space, which we denote by $\mathcal{F}_{t,n}$, are defined as the variables spanning the new information in $\mathcal{Y}_{t,n}$ which are orthogonal to $\mathcal{Y}_{t,n-1}$:

$$\mathcal{Y}_{t,n} = \mathcal{Y}_{t,n-1} \oplus \mathcal{F}_{t,n} \tag{3.7}$$

so that $\mathcal{F}_{t,n}$ is the orthogonal complement of $\mathcal{Y}_{t,n-1}$ in $\mathcal{Y}_{t,n}$ (i.e. $\mathcal{F}_{t,n} = \mathcal{Y}_{t,n} \ominus \mathcal{Y}_{t,n-1}$ for $n > 0$, while $\mathcal{F}_{t,0} = \mathcal{Y}_{t,0}$). A basis for $\mathcal{F}_{t,n}$ can be obtained by defining the backward prediction errors for the "new" elements of the "past" introduced at the nth step, i.e. for $w \preceq 0$ and $|w| = n$, define

$$F_{t,n}(w) \stackrel{\Delta}{=} y_{tw} - E\left(y_{tw} | \mathcal{Y}_{t,n-1}\right) \tag{3.8}$$

Then

$$\mathcal{F}_{t,n} = \mathcal{H}\{F_{t,n}(w) : |w| = n, w \preceq 0\} \tag{3.9}$$

Similarly we introduce the forward innovations or prediction error space, which we denote by $\mathcal{E}_{t,n}$. For $n = 0$, $\mathcal{E}_{t,0} = \mathcal{H}\{y_t\}$, while for $n > 0$

$$\mathcal{E}_{t,n} \stackrel{\Delta}{=} (\mathcal{Y}_{t,n-1} + \mathcal{Y}_{t\overline{\gamma},n-1}) \ominus \mathcal{Y}_{t\overline{\gamma},n-1} \tag{3.10}$$

Note that $\mathcal{Y}_{t,n-1} + \mathcal{Y}_{t\overline{\gamma},n-1}$ is used here instead of $\mathcal{Y}_{t,n}$; while both spaces are equal in the case of ordinary time series (in which $\overline{\gamma}$ is replaced by z^{-1}), they differ here[5]. To obtain a basis for $\mathcal{E}_{t,n}$, we define the forward innovations

$$E_{t,n}(w) \stackrel{\Delta}{=} y_{tw} - E\left(y_{tw} | \mathcal{Y}_{t\overline{\gamma},n-1}\right) \tag{3.11}$$

where w ranges over a set of words such that tw is on the same horocycle as t and at a distance at most $n-1$ from t (so that $\mathcal{Y}_{t\overline{\gamma},n-1}$ is the past of that point as well), i.e. $|w| < n$ and $w \succ 0$. Then

$$\mathcal{E}_{t,n} = \mathcal{H}\{E_{t,n}(w) : |w| < n \text{ and } w \succ 0\} \tag{3.12}$$

[5]For example $\mathcal{Y}_{t,2}$ consists of y_t, $y_{t\overline{\gamma}}$, $y_{t\overline{\gamma}^2}$, and $y_{t\delta}$. However, $\mathcal{Y}_{t,1}$ consists of y_t and $y_{t\overline{\gamma}}$, while $\mathcal{Y}_{t\overline{\gamma},1}$ consists of $y_{t\overline{\gamma}}$ and $y_{t\overline{\gamma}^2}$.

Let $E_{t,n}$ and $F_{t,n}$ denote column vectors of the elements $E_{t,n}(w)$ and $F_{t,n}(w)$, respectively. As n increases the dimensions of these residual vectors grow geometrically. Levinson recursions for isotropic processes involve the recursive computation of $F_{t,n}$ and $E_{t,n}$ as n increases. Since $F_{t,0}$ and $E_{t,0}$ both equal y_t, these recursions yield lattice structures for AR models of increasing order. As developed in [6] and as the reader may guess from the results for time series, the key to these recursions are all PARCOR coefficients involving an arbitrary pair $\{\square, \diamondsuit\}$ given the space spanned by the $\bigcirc$ in Figure 6. Furthermore, it can be verified that suitable combinations of the elementary isometries shown in this figure provide isometries

- leaving the space $\mathcal{Y}_{t\overline{\gamma},3}$ (circles) globally invariant
- exchanging two arbitrary $\square$'s or the two $\diamondsuit$.

From this it follows that *all pairs* $\{\square, \diamondsuit\}$ *possess the same PARCOR coefficients given the space spanned by the circles.* Hence, as for time series, we can show in general that *a single PARCOR or reflection coefficient is involved in each stage of the Levinson recursions.* Similar uses of the symmetries of the tree and the correlation structure of isotropic processes allows us to show that only the barycenters of the forward and backward prediction error vectors are needed to compute these reflection coefficients. These barycenters are defined as follows :

$$\begin{aligned} e_{t,n} &= 2^{-[\frac{n-1}{2}]} \sum_{|w|<n, w \asymp 0} E_{t,n}(w) \\ f_{t,n} &= 2^{-[\frac{n}{2}]} \sum_{|w|=n, w \preceq 0} F_{t,n}(w) \end{aligned}$$

In particular in [6] the following results are proven providing a generalization of the Levinson recursions to the barycentric prediction errors for isotropic processes on $\mathcal{T}$:

Theorem 1 (barycentric Levinson recursions) *For n even:*

$$e_{t,n} = e_{t,n-1} - k_n f_{t\overline{\gamma},n-1} \tag{3.13}$$

$$f_{t,n} = \frac{1}{2}\left(f_{t\overline{\gamma},n-1} + e_{t\delta(\frac{n}{2}),n-1}\right) - k_n e_{t,n-1} \tag{3.14}$$

where

$$\begin{aligned} k_n &= cor(e_{t,n-1}, f_{t\overline{\gamma},n-1}) \\ &= cor\left(e_{t\delta(\frac{\pi}{2}),n-1}, e_{t,n-1}\right) \\ &= cor\left(e_{t\delta(\frac{\pi}{2}),n-1}, f_{t\overline{\gamma},n-1}\right) \end{aligned} \tag{3.15}$$

and $cor\,(x,y) = E(xy)/\left[E(x^2)E(y^2)\right]^{1/2}$.

For n odd:

$$e_{t,n} = \frac{1}{2}\left(e_{t,n-1} + e_{t\delta(\frac{n-1}{2}),n-1}\right) - k_n f_{t\overline{\gamma},n-1} \tag{3.16}$$

$$f_{t,n} = f_{t\overline{\gamma},n-1} - \frac{1}{2}k_n\left(e_{t,n-1} + e_{t\delta(\frac{n-1}{2}),n-1}\right) \tag{3.17}$$

where

$$k_n = cor\left(\frac{1}{2}\left(e_{t,n-1} + e_{t\delta(\frac{n-1}{2}),n-1}\right), f_{t\overline{\gamma},n-1}\right) \tag{3.18}$$

Corollary: *The variances of the barycenters satisfy the following recursions.*
For n even

$$\sigma^2_{e,n} = E\left(e^2_{t,n}\right) = \left(1 - k_n^2\right)\sigma^2_{n-1} \tag{3.19}$$

$$\sigma^2_{f,n} = E\left(f^2_{t,n}\right) = \left(\frac{1+k_n}{2} - k_n^2\right)\sigma^2_{n-1} \tag{3.20}$$

where k_n *must satisfy*

$$-\frac{1}{2} \le k_n \le 1 \tag{3.21}$$

For n odd

$$\sigma^2_{e,n} = \sigma^2_{f,n} = \sigma^2_n = \left(1 - k_n^2\right)\sigma^2_{f,n-1} \tag{3.22}$$

where

$$-1 \le k_n \le 1 \tag{3.23}$$

As we had indicated previously, the constraint of isotropy represents a significantly more severe constraint on the covariance sequence $r(n)$ of an isotropic process than on that for a stationary time series. It is interesting to note that these additional

constraints appear in the preceding development only in the form of the simple modification (3.21) of the constraint on k_n for n even over the form (3.23) that one also finds in the corresponding theory for time series. Also, as with the usual Levinson recursions for time series we can use the formulae in Theorem 1 and its corollary to obtain explicit recursions for the computation of the k_n sequence directly from the given covariance data, $r(n)$. These recursions also contain some differences from the usual results reflecting the constraints of isotropy on the tree. Rather than displaying these we describe here an alternative computational procedure generalizing the so-called Schur recursions [30, 43] for the cross-spectral densities between a given time series and its forward and backward prediction errors. In considering the generalization of these recursions to isotropic processes on trees, we must replace the z-transform power series for cross-spectral densities by corresponding formal power series of the type introduced in Section 2. Specifically for $n \geq 0$ define P_n and Q_n as:

$$P_n \triangleq \operatorname{cov}(y_t, e_{t,n}) \triangleq \sum_{w \preceq 0} E(y_t e_{tw,n}) \cdot w \tag{3.24}$$

$$Q_n \triangleq \operatorname{cov}(y_t, f_{t,n}) \triangleq \sum_{w \preceq 0} E(y_t f_{tw,n}) \cdot w \tag{3.25}$$

where we begin with P_0 and Q_0 specified in terms of the correlation function r_n of y_t:

$$P_0 = Q_0 = \sum_{w \preceq 0} r(|w|) \cdot w \tag{3.26}$$

Recalling the definitions (2.27), (2.28) of $\overline{\gamma}[S]$ and $\delta^{(k)}[S]$ for S a formal power series and letting $S(0)$ denote the coefficient of $w = 0$, we have the following generalization of the Schur recursions, proven in [6]:

Theorem 2 (Schur recursions) *The following Schur recursions on formal power series yield the sequence of reflection coefficients.*
For n even

$$P_n = P_{n-1} - k_n \overline{\gamma}[Q_{n-1}] \tag{3.27}$$

$$Q_n = \frac{1}{2}\left(\overline{\gamma}[Q_{n-1}] + \delta^{(\frac{n}{2})}[P_{n-1}]\right) - k_n P_{n-1} \tag{3.28}$$

where

$$k_n = \frac{\overline{\gamma}[Q_{n-1}](0) + \delta^{(\frac{n}{2})}[P_{n-1}](0)}{2P_{n-1}(0)} \tag{3.29}$$

For n odd

$$P_n = \frac{1}{2}\left(P_{n-1} + \delta^{(\frac{n-1}{2})}[P_{n-1}]\right) - k_n\overline{\gamma}[Q_{n-1}] \tag{3.30}$$

$$Q_n = \overline{\gamma}[Q_{n-1}] - k_n\frac{1}{2}\left(P_{n-1} + \delta^{(\frac{n-1}{2})}[P_{n-1}]\right) \tag{3.31}$$

where

$$k_n = \frac{2\overline{\gamma}[Q_{n-1}](0)}{P_{n-1}(0) + \delta^{(\frac{n-1}{2})}[P_{n-1}](0)} \tag{3.32}$$

Theorems 1 and 2 provide us with the right way in which to parametrize isotropic processes. Furthermore, as developed in [6, 7, 8], we can build on these results to provide a complete generalization of the Wold decomposition of an isotropic process. In particular, lattice structures can be constructed for whitening filters, i.e. for the computation of the prediction error vectors $E_{t,n}$ and $F_{t,n}$ as outputs when y_t is taken as input. Similarly lattice forms are derived in [7] for modeling filters, i.e. systems whose output is the isotropic process when the input is the corresponding-order prediction error. Figure 2 illustrates the output, along one horocycle of a third-order modeling filter (i.e. an AR(3)-model) driven by a white $E_{t,3}$ process. We note that a major difference between these lattice structures and the usual ones for time series is that they involve lattice blocks of growing dimension, capturing the coupling along a horocycle for AR processes of higher order. Also, as with time series, statistical properties of isotropic processes may be checked using the parametrization via reflection coefficients. The main results are now listed and we again refer the reader to [7, 8] for more precise formulations of these results and their proofs.

Theorem 3 (checking properties via reflection coefficients)

1. **Characterization of AR processes :** *an isotropic process is AR(n) if and only if its reflection coefficients of order > n are all zero.*

2. **Schur criterion :** *if the sequence* (r_n) *is the covariance function of an isotropic process, then the Schur recursions must yield reflection coefficients satisfying the inequalities*

$$-1 \leq k_{2n+1} \leq +1 \ , \ -\frac{1}{2} \leq k_{2n} \leq +1 \tag{3.33}$$

3. **Parametrizing AR processes :** *conversely, a finite family of coefficients satisfying the above* **strict** *inequalities (3.33) defines a unique isotropic AR process.*

4. **Regular and singular processes :** *If the sequence* (r_n) *satisfies the strict inequalities (3.33) and furthermore the condition*

$$\sum_{n=1}^{\infty} k_{2n+1}^2 + |k_{2n}| \ < \infty$$

holds true, then it is the reflection coefficient sequence of a regular (i.e. purely nondeterministic) isotropic process.

The first three of these results represent easily understood generalizations of results for time series. For example they imply that the nth and higher-order prediction error vectors of an AR(n) process are white noise processes. The fourth statement concerns itself with the issue of whether or not the value of y_t can be perfectly prediction based on data in its (infinite) past. Specifically, an isotropic process y_t is *regular* or *purely nondeterministic* if

$$\sigma^2 > 0 \tag{3.34}$$

holds, where

$$\sigma^2 \triangleq \inf \| \left(\sum_{w \prec 0} \mu_w y_{tw} \right) - E \left(\left(\sum_{w \prec 0} \mu_w y_{tw} \right) | \mathcal{Y}_{t\gamma^{-1},\infty} \right) \|^2 \tag{3.35}$$

and the infimum ranges over all collections of scalars $(\mu_w)_{w \prec 0}$ where only finitely many of the μ_w are nonzero and the condition $\sum \mu_w^2 = 1$ is satisfied. In other words, no nonzero linear combination of the values of y_t on any given horocycle can be predicted exactly with the aid of knowledge of Y in the strict past, $\mathcal{Y}_{t\gamma^{-1},\infty}$ and the associated

prediction error is uniformly bounded from below. It is interesting to note that the condition for regularity for isotropic processes involves the absolute sum rather than sum of squares of the even reflection coefficients and thus is a stronger condition. This implies that there is apparently a far richer class of singular processes on $\mathcal{T}$ than on Z. This appears to be related to the complications arising in the Bochner theorem for isotropic processes on $\mathcal{T}$ and to the large size of its boundary. We refer the reader to [6, 7, 8] for further discussions of these and other points related to isotropic processes and their AR representations.

4 System Theory and Estimation for Stationary Processes and State Models

In this section we describe some of the basic concepts associated with the analysis of stationary systems and processes on the dyadic tree. To begin, let us introduce the following basic systems on $\mathcal{T}$:

$$(\gamma.u)_t = \frac{1}{2}(u_{t\alpha} + u_{t\beta}) \tag{4.1}$$

$$(\overline{\gamma}.u)_t = u_{t\overline{\gamma}} \tag{4.2}$$

It is not difficult to check that each of these systems is stationary. The system $\overline{\gamma}$ can be naturally thought of as a "backward" shift towards $-\infty$, corresponding to the coarse-to-fine interpolation operation in the fine-to-coarse Haar transform, whereas γ is a "forward–and–average" shift corresponding to the "Haar smoother". Using these operators, it is not difficult to show that a stationary system can be represented as

$$H = \sum_{i,j \geq 0} s_{i,j}\, \overline{\gamma}^i \gamma^j \tag{4.3}$$

Such a system is causal if and only if $s_{i,j}$ is nonzero only over the set $\{(i,j) \ : i \geq j\}$, i.e. only past inputs can influence the considered output.

The representation in (4.3) is one of two extremely useful transform-like representations of stationary systems. This one is, in particular, of use in providing a generalization of time series results on the effect of linear systems on power spectra and cross-spectra. Specifically, consider two jointly stationary processes x and y, with covariance function

$$E(x_s y_t) = r^{xy}[d(s, s \wedge t), d(t, s \wedge t)] \tag{4.4}$$

Let us define the *cross–spectrum* of x and y as the following power series:

$$R^{xy} \stackrel{\Sigma}{=} \sum_{i,j \geq 0} r^{xy}[i,j]\, \overline{\gamma}^i \gamma^j$$

Also, given a stationary transfer function as in (4.3), we introduce the following notion of an "adjoint" :

$$H^* = \sum s_{j,i} \overline{\gamma}^i \gamma^j \tag{4.5}$$

Then as shown in [9], if H and K are stationary transfer functions, the processes Hx and Ky are also jointly stationary[6], and we have the following generalization of a well-known result :

$$R^{(H_x)(K_y)} = H^* R^{xy} K \tag{4.6}$$

Let us now turn to the question of internal, "state" realizations of stationary systems. In this case an alternate representation to (4.3) is also of value. To define this we introduce the following family of operators which perform a smoothing of data on the same horocycle:

$$\sigma^{[i]} = \overline{\gamma}^i \gamma^i \tag{4.7}$$

This operator provides an average of the values of a signal at the 2^i nearest points on the same horocycle. For example, $(\sigma.u)_t = 1/0(u_t + u_{t\delta})$ where $\sigma = \sigma^{[1]}$ and $(\sigma^{[2]}.u)_t = \frac{1}{4}(u_t + u_{t\delta} + u_{t\delta(2)} + u_{t\delta(2)\delta})$. Note also that each $\sigma^{[i]}$ is an *idempotent* operator. As shown in [9] operators may be used to encode any stationary causal system via a representation of the form :

$$H = \sum_{i,j \geq 0} h_{i,j}\, \overline{\gamma}^i \sigma^{[j]} \tag{4.8}$$

In order to develop a realization theory for stationary systems, let us note that both formulae (4.3) and (4.8) are strikingly similar to the forms of system functions studied in standard 2-D system theory. While there are obvious differences – e.g. we have the relation $\gamma\overline{\gamma} = 1$ between the two variables in (4.3) and the symbol $\sigma^{[2]}$ is not simply interpretable as the square of σ– it is indeed possible to build on standard 2-D realization theories. Note in particular that even though (4.3) includes noncausal multiscale systems, it has the appearance of a 2-D quadrant-causal system, as does (4.8) since the summations are restricted to $i, j \geq 0$. Let us begin with, (4.3). Building on the 2-D analogy, if we interpret γ as the row operator and $\overline{\gamma}$ as the column generator, then it is natural to consider row–by–row scanning to define a total ordering on the $2D$ index space. This corresponds to decomposing the transfer function H according to the following two steps:

[6]This of course, is true only if Hx and Ky are well-defined, i.e. if they are finite-variance processes. As one might expect, this requires some notion of stability for the systems. We return to this point later in this section in the context of state models.

1. a bottom–up (i.e. fine–to–coarse) smoothing, followed by

2. a top-down (i.e. coarse–to–fine) propagation.

$2D$–system theory for systems having separable denominator [4, 32] may be applied here. Rational transfer functions in this latter case are of the following form:

$$H = C\,(I - \overline{\gamma} A_{\overline{\gamma}})^{-1}\, P\,(I - \gamma A_{\gamma})^{-1}\, B \tag{4.9}$$

which yields the following state space form

$$\left\{ \begin{array}{rcl} v_t & = & A_\gamma \left(\frac{v_{t\alpha} + v_{t\beta}}{2} \right) + B u_t \\ z_t & = & P_2 v_t \\ x_{t\alpha} & = & A_{\overline{\gamma}} x_t + P_1 z_{t\alpha} \\ x_{t\beta} & = & A_{\overline{\gamma}} x_t + P_1 z_{t\beta} \\ y_t & = & C x_t \end{array} \right. \tag{4.10}$$

where $P = P_1 P_2$. The first two equations define a purely "anticausal" process, whereas the last three equations define a causal zero depth process. Later in this section we describe an optimal multiscale estimation algorithm that has precisely this structure.

Now let us turn to the representation of multiscale causal systems in (4.8). Here we interpret the sequence $\sigma^{[i]}$ as the powers of the row operator and $\overline{\gamma}$ as the column operator. Then again we consider row–by–row scanning to define a total ordering of the $2D$ index space. This corresponds to decomposing the transfer function H according to the following two steps:

1. a smoothing along the considered horocycle (i.e. constant scale smoothing), followed by

2. a top-down (i.e. coarse–to–fine) propagation.

$2D$–system theory for systems having separable denominator [4, 32] may again be applied here. Rational transfer functions in this latter case are of the following form:

$$H = C\,(I - \overline{\gamma} A_{\overline{\gamma}})^{-1}\, P\,(I - \sigma A_{\sigma})^{-1}\, B \tag{4.11}$$

where it is understood that, in expanding such a formula into a power series, σ^i should be replaced by $\sigma^{[i]}$. This latter unusual feature has as a consequence the fact that no simple "time domain" translation of the "frequency domain" formula (4.11) is available. However, if A_σ is nilpotent so that $(I - \sigma A_\sigma)^{-1}$ is a finite series, we do obtain the following explicit representation for what we refer to as the *finite depth* case :

$$\begin{cases} x_{t\alpha} &= A_{\overline{\gamma}} x_t + D\left(1, \sigma, ..., \sigma^{[i]}\right) u_{t\alpha} \\ x_{t\beta} &= A_{\overline{\gamma}} x_t + D\left(1, \sigma, ..., \sigma^{[i]}\right) u_{t\beta} \\ y_t &= C x_t \end{cases} \tag{4.12}$$

where $D\left(1, \sigma, ..., \sigma^{[i]}\right)$ is a linear combination of the listed operators.

The dynamics (4.12) represent a finite-extent smoothing along each horocycle and a generalized coarse-to-fine interpolation. For example, as discussed in Section 1, the synthesis form of the Haar transform can be placed exactly in this form. It can also be shown that stationary finite depth scalar transfer functions may be equivalently expressed in the following *ARMA* form

$$H = A^{-1} D \tag{4.13}$$

where A is a causal function of *finite support* and $D = D\left(1, \sigma, ..., \sigma^{[k]}\right)$ is as in (4.12). This *ARMA* form includes as a special case the AR modeling filters for "isotropic" processes introduced in Section 3.

The preceding development, as well as the interpretation of the synthesis form of the wavelet transform provides ample motivation for the studies in [16, 17, 18, 19, 20, 48, 52] of properties and estimation algorithms for multiscale state models of the form:

$$x(t) = A(t) x(t\overline{\gamma}) + B(t) w(t) \tag{4.14}$$

$$y(t) = C(t) x(t) + v(t) \tag{4.15}$$

where $w(t)$ and $v(t)$ are independent vector white noise processes with covariances I and $R(t)$, respectively. The model class described in (4.14),(4.15) represents a noise-driven generalization of the zero-depth, causal, stationary model (4.12). Specifically we obtain such a stationary model if all of the parameters, A, B, C, and R are

constant. There are, however, important reasons to consider the more general case (and, in addition, its consideration does not complicate our analysis). First of all, one important intermediate case is that in which the system parameters are constant at each scale but may vary from scale to scale. If we let $m(t)$ denote the scale, i.e. the horocycle, on which the node t lies, we abuse notation in this case by writing $A(t) = A(m(t))$, etc. Such a model is useful for capturing the fact that data may be available at only particular scales (i.e. $C(m) \neq 0$ only for particular values of m); for example in the original context of wavelet analysis, we actually have only one measurement set, corresponding to $C(m)$ being nonzero only at the finest scale in our representation. [7] Also, by varying $A(m), B(m)$, and $R(m)$ with m we can capture a variety of scale-dependent effects. For example, dominant scales might correspond to scales with larger values of $B(m)$. Also, by building a geometric decay in scale into $B(m)$ it is possible to capture 1/f-like, fractal behavior as shown and studied in [16, 47, 50]. Finally, the general case of t-varying parameters has a number of potential uses. For example such form for $C(t)$ is clearly required to capture the situation depicted in Figure 3 in which fine scale measurements are not available at all locations. Also, it is our belief that such models will prove useful in modeling transient events localized in scale and time or space and to capture changing signal or image characteristics.

As with standard temporal state models, the second-order statistics of $x(t)$ are easily computed. In particular the covariance $P_x(t) = E[x(t)x^T(t)]$ evolves according to a Lyapunov equation on the tree:

$$P_x(t) = A(t)P_x(t\overline{\gamma})A^T(t) + B(t)B^T(t) \tag{4.16}$$

Specializing to the case in which $A(t) = A(m(t))$ and $B(t) = B(m(t))$, we can obtain a covariance that allows dependence only on scale, i.e. $P_x(t) = P_x(m(t))$, and indeed in this case we have a standard Lyapunov equation in scale :

$$P_x(m+1) = A(m)P_x(m)A^T(m) + B(m)B^T(m) \tag{4.17}$$

[7] It is important to emphasize here that the wavelet transform of this fine scale measurement— which we use as well as in the sequel– does not correspond to measurements as in (4.15) at several scales. Rather (4.15) corresponds to independent measurements at various nodes.

Also, as shown in [16, 19] the full covariance function in this case is given by

$$K_{xx}(t,s) = \Phi(m(t), m(s \wedge t)) P_x(m(s \wedge t)) \Phi^T(m(s), m(s \wedge t)) \tag{4.18}$$

where $\Phi(m,n)$ is the state transition matrix associated with $A(m)$. Specializing further to the constant coefficient case we have the following [16, 19]: if A is stable and if P_x is the unique solution to the algebraic Lyapunov equation

$$P_x = A P_x A^T + B B^T \tag{4.19}$$

then our state model generates the stationary covariance

$$K_{xx}(t,s) \;=\; A^{d(t,s\wedge t)} P_x (A^T)^{d(s,s\wedge t)} \tag{4.20}$$

Note that in the scalar case our constant coefficient model is exactly the AR(1) model introduced in the preceding section and indeed (4.19)-(4.20) reduce to

$$K_{xx}(t,s) = \left\{ \frac{B^2}{1-A^2} \right\} A^{d(s,t)} \tag{4.21}$$

In the vector case (4.20) is stationary but not, in general, isotropic. However, it is interesting to note that we do obtain an isotropic model if $AP_x = P_x A^T$, precisely the condition arising in the study of temporally-reversible vector stochastic models [1]. Let us turn now to the problems of estimating the state of (4.14) based on the measurements (4.15). Note that this framework allows us to consider not only the fusion of measurements at multiple resolutions but also the reconstruction of processes at multiple scales. Indeed in this way we can consider the resolution-accuracy tradeoff directly and can also assess the impact of fine-scale fluctuations on the accuracy of coarser scale reconstructions, a problem of some importance in applications such as the fusion of satellite IR measurement of ocean temperature variations with point measurements from ships in order to produce temperature maps at an intermediate scale. To be specific in the following development we consider the problem of optimal estimation on a finite portion of $\mathcal{T}$. This corresponds to estimation of a temporal process on a compact interval so that there is a coarsest scale (and hence a top to our subtree) denoted by $m = 0$, and a finest scale, denoted by $m = M$, at which

measurements may be available and/or reconstructions desired. As developed in [16, 17, 18, 52], the model structure (4.14), (4.15) leads to three efficient, highly parallelizable algorithmic structures for optimal multiscale estimation. A first of these is an iterative algorithm taking advantage of the fact that (4.14) defines a Markov random field structure on $\mathcal{T}$. Specifically, let Y denote the full set of measurements at all scales. Then, thanks to Markovianity we have that

$$\begin{aligned} E[x(t)|Y] &= E\{E[x(t)|x(t\overline{\gamma}), x(t\alpha), x(t\beta), Y]|Y\} \\ &= E\{E[x(t)|x(t\overline{\gamma}), x(t\alpha), x(t\beta), y(t)]|Y\} \end{aligned} \tag{4.22}$$

where the second equality in (4.22) states that given $x(t\overline{\gamma})$, $x(t\alpha)$, $x(t\beta)$, only the measurement at node t provides additional useful information about $x(t)$. From (4.22) we can then obtain an explicit representation for the optimal estimate of $x(t)$ in terms of the optimal estimates at its parent node, $t\overline{\gamma}$, at its immediate descendant nodes, $t\alpha$ and $t\beta$, and the single measurement at node t. This implicit specification is then perfectly set up for solution via Gauss-Seidel or Jacobi iteration which can be organized to have exactly the same structure as multigrid relaxation algorithms, with coarse-to-fine and fine-to-coarse sweeps that in multigrid terminology [11, 12, 26, 29, 37, 39] lead to so-called V- and W-cycle iterations. Furthermore, in such iterations all of the calculations at any particular scale can be carried out in parallel. In addition this methodology carries over completely not only to the case of nonzero depth models as in (4.12), with the additional inter-node connectivity implied by the coupling introduced by the horocycle-smoothing operator D, but also to state models on more general lattices corresponding to the interpretation of (1.11) as defining a scale-to-scale dynamic relationship for any finitely-supported QMF pair $h(n)$, $g(n)$ and thus for any compactly-supported wavelet transform. We refer the reader to [16, 19] for details and further development of this multigrid estimation methodology.

A second estimation structure applies to the case in which all system parameters depend only on scale (i.e. $A(t) = A(m(t))$, etc.). In this case, as shown in [16, 17, 19], the Haar transform, applied to each scale of the state process $x(t)$ and the measurement data $y(t)$ yields a decoupled set of estimation problems for each of the scale components. Specifically, let $x(m)$ denote the vector of all 2^m values of $x(t)$

at the mth scale, and let $y(m)$, $w(m)$, and $v(m)$ similarly. Then in this case (4.14), (4.15) can be rewritten in scale-to-scale form:

$$x(m+1) = \mathcal{A}_{m+1}x(m) + \mathcal{B}_{m+1}w(m+1) \tag{4.23}$$

$$y(m) = \mathcal{C}_m x(m) + v(m) \tag{4.24}$$

where

$$\mathcal{A}_{m+1} = \begin{bmatrix} A(m+1) & 0 & 0 & \cdots & 0 \\ A(m+1) & 0 & 0 & \cdots & 0 \\ 0 & A(m+1) & 0 & \cdots & 0 \\ 0 & A(m+1) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & A(m+1) \\ 0 & 0 & 0 & \cdots & A(m+1) \end{bmatrix} \tag{4.25}$$

$$\mathcal{B}_{m+1} = diag(B_{m+1}, ..., B_{m+1}) \tag{4.26}$$

$$\mathcal{C}_m = diag(C(m), ..., C(m)) \tag{4.27}$$

Note that $x(m)$ has half as many elements as $x(m+1)$, reflecting the fine-to-coarse decimation that occurs in multiscale representations. As shown in [16, 19], the covariances of $x(m)$ and $y(m)$ as well as the cross-covariance between x at different scales have (block-) eigenstructures specified by the Haar transform. For example if $x(t)$ is a scalar process and we look at $x(3)$, which is 8-dimensional, we find that the covariance of this vector has as its eigenvectors the columns of the following orthonormal matrix, corresponding to the (8-dimensional) discrete Haar basis consisting of vectors

representing "dilated, translated, and scaled" versions of the vector $[1,-1]^T$:

$$V_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & -\frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & -\frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix} \tag{4.28}$$

Analogous bases can be defined for any dimension that is a power of two, and when $x(t)$ is a vector each of the elements of matrices as in (4.28) is replaced by a correspondingly-scaled version of the identity matrix of dimension equal to that of $x(t)$ (e.g. the $(1,1)$ block of such a matrix would be $(1/\sqrt{2})I$).

As a consequence of these observations, one would expect considerable simplification if we consider the Haar-transformed version of our estimation problem. Specifically, define the transformed variables

$$s(m) = V_x^T(m)x(m), \quad z(m) = V_y^T(m)y(m) \tag{4.29}$$

where $V_x(m)$ ($V_y(m)$) is the block-Haar transform matrix of block-size equal to the dimension of $x(t)$ ($y(t)$). In this transformed representation the system and measurement equations block-decouple completely. Specifically, the vector $s(m)$ can be decomposed into 2^m subvectors each of the same dimension as $x(t)$, and we index these as $s_{00}(m)$, $s_{01}(m)$, and $s_{ij}(m)$ for $1 \leq i \leq m-1$, $1 \leq j \leq 2^i$. Here $s_{00}(m)$ is the component corresponding to the right-most (block) basis component in $V_x(m)$ (refer to (4.28))– i.e. it is the average of the $x(t)$ at the mth horocycle (scaled by $2^{-m/2}$); $s_{01}(m)$ is then the coarsest resolution first difference coefficient (see the next-to-last column in (4.28)), while for $i \geq 1$, the s_{ij} correspond to the ith resolution first difference coefficients (note in (4.28) that there are four such coefficients at the finest resolution and two at the next, coarser scale). In a similar fashion we define the components of $z(m)$. With these definitions we find that we have a set of *completely*

decoupled standard dynamic systems in the time-like variable m:

$$s_{ij}(m+1) = A(m+1)s_{ij}(m) + B(m+1)w_{ij}(m+1),\ 0 \le i \le m-1 \tag{4.30}$$

$$s_{mj}(m+1) = B(m+1)w_{mj}(m+1) \tag{4.31}$$

$$z_{ij}(m) = C(m)s_{ij}(m) + v_{ij}(m) \tag{4.32}$$

Here $w_{ij}(m)$ and $v_{ij}(m)$ are white in all three indices, with covariances I and $R(m)$, respectively.

Recall that the dimension of $x(m)$ increases with m, indicative of the increasing detail available at finer scales. In the transformed basis this is made absolutely explicit in that we see that the dynamics (4.30), (4.31) consists of two parts: the interpolation of coarse features to finer scales (4.30) and the initiation, at each scale, of new components (4.31) representing levels of detail that can be captured at this (but not at any coarser) scale. Thus for any pair of indices i, j we have a dynamic system in m, initiated at scale $m = i$, and thus we can use standard state space smoothing techniques independently for each such system, leading to a highly parallel algorithm in which (a) we transform the available measurement data $y(m)$ to obtain $z(m)$ as in (4.29); (b) we then use standard smoothing techniques on the individual components; and (c) we inverse transform the resulting estimates of $s(m)$ to obtain the optimal estimates of $x(t)$ at all nodes. Note that the fact that each s_{ij} is initiated only at the ith scale implies that the corresponding smoother works on data only from this and finer scales, leading to a set of smoothing algorithms of different (scale) length. This is consistent with the intuition that data at any particular scale provides useful information at that scale and at coarser scales (by averaging) but not at finer scales.

We refer the reader to [16, 17] for details of this procedure and for its generalization to the case of nonzero-depth models and to arbitrary lattices associated with other wavelet transforms–i.e. to dynamic system as in (1.11) (and a significant extension of these) with other choices for the QMF's $h(n)$ and $g(n)$ than the Haar pair. Again one finds that the wavelet transformed – modified appropriately to deal with the windowing effect of smoothing multiscale measurements over a compact interval – yields a set of decoupled smoothing problems in scale. Since the wavelet transform can be computed quite quickly, this leads to an extremely efficient overall procedure. We note

also that by specializing our model to the case in which process noise variances decrease exponentially in scale we obtain a generalization of the procedure developed in [51] for the estimation of 1/f-like processes. In particular, what we have just described provides a procedure for fusing multiresolution measurements of such processes. Finally, we note that the interpretation of our models as scale-to-scale Markov processes and the dual viewpoint that the wavelet transform for such a model whitens the data in scale suggest the problems of (a) optimizing wavelet transforms in order to achieve maximal scale-to-scale decorrelation; and (b) approximating stochastic processes by such scale-to-scale Markov models. The former of these problems is discussed in [27] and the latter is touched upon in [16, 17, 27]. In particular in [17, 27] we construct approximate models of this type for a standard first-order Gauss-Markov process (i.e. with temporal correlation function of the form $\sigma^2 e^{-\alpha|t|}$) and demonstrate their fidelity in several ways including their use as the basis for the fusion and smoothing of multiresolution measurements of Gauss-Markov processes. In Figure 7 we depict the correlation function of such a unit-variance first-order Gauss-Markov process – i.e. viewing a set of 2^m samples of this process as the values of $x(m)$, Figure 7 displays the matrix of correlation coefficients of the elements of this vector. In contrast in Figure 8 we display the correlation coefficients of the elements of $s(m)$ obtained as in (4.29), but using an 8-tap QMF $h(n)$ rather than the 2-tap $h(n)$ – i.e. first the corresponding orthogonal matrix for this $h(n)$ is applied to $x(m)$, and then the resulting covariance of $s(m)$ is modified by dividing its (i,j) element by the square-root of the product of the (i,i) and (j,j) elements, yielding the matrix of correlation coefficients. As one would expect from the work on transforming kernels of integral operators in [10], the result is an almost-diagonal matrix, implying nearly perfect scale-to-scale whitening. This is further substantiated in [16, 17] (see also Figure 3) by demonstration of the high quality estimates produced if such remaining inter-scale correlation is neglected.

While the preceding algorithm provides a very efficient procedure for multiscale fusion, its use does require that all model parameters vary only with scale and thus are constant on each horocycle. For example this implies that if any measurement is available at any particular scale, than a full set of measurements is available at that scale. In contrast, the result shown in Figure 3 (a),(b) corresponds to a situation in which we

have only sparse, fine scale measurements from a 1/f-like model of the type described in [50, 51], together with full-coverage, but coarser-resolution measurements, while Figure 3 (c) and (d) correspond to the analogous situation for a first-order Gauss-Markov process. In particular in each case 16 fine scale measurements are taken at each end of the 64-point signal, together with coarse measurements of 4-point averages of this signal. While the wavelet-transform-based smoothing algorithm does not apply to this case, the multigrid method described previously does (using in the case of (c) and (d) an approximate model of the form of (4.14), (4.15) for the Gauss-Markov process), as does the following approach which not only provides an extremely efficient algorithm for multiscale fusion but also illuminates several system-theoretic issues on dyadic trees. Specifically, as developed in detail in [16, 18, 19], there is a nontrivial generalization of the so-called Rauch-Tung-Striebel (RTS) smoothing algorithm for causal state models [42]. Recall that the standard RTS algorithm involves a forward Kalman filtering sweep followed by a backward sweep to compute the smoothed estimates. The generalization to our models on trees has the same structure, with several important differences. First for the standard RTS algorithm the procedure is completely symmetric with respect to time – i.e. we can start with a reverse-time Kalman filtering sweep followed by a forward smoothing sweep. For processes on trees, the Kalman filtering sweep *must* proceed from fine-to-coarse followed by a coarse-to-fine smoothing sweep[8].

Furthermore the Kalman filtering sweep, is somewhat more complex for processes on trees. In particular one full step of the Kalman filter recursion involves a measurement update, *two* parallel backward predictions (corresponding to backward prediction along both of the paths descending from a node), and the *fusion* of these predicted estimates. Specifically, as depicted in Figure 9, the fine-to-coarse Kalman filter step has as its goal the recursive computation of $\hat{x}(t|t)$, the best estimate of $x(t)$ based on data in the descendant subtree with root node t. As in usual Kalman filtering if $\hat{x}(t|t+)$ denotes the best estimate based on all of the same data *except the*

[8]The reason for this is not very complex. To allow the measurement on the tree at one point to contribute to the estimate at another point on the same level of the tree, one must use a recursion that first moves up and then down the tree.

measurement at node t, we obtain a straightforward update step to produce $\hat{x}(t|t)$:

$$\hat{x}(t|t) = \hat{x}(t|t+) + K(t)[y(t) - C(t)\hat{x}(t|t+)] \tag{4.33}$$

$$K(t) = P(t|t+)C^T(t)V^{-1}(t) \tag{4.34}$$

$$V(t) = C(t)P(t|t+)C^T(t) + R(t) \tag{4.35}$$

and

$$P(t|t) = [I - K(t)C(t)]P(t|t+) \tag{4.36}$$

Here $P(t|t)$ and $P(t|t+)$ are the error covariances associated with $\hat{x}(t|t)$ and $\hat{x}(t|t+)$, respectively. Working back one-step, we see that $\hat{x}(t|t+)$ represents the *fusion* of information in the subtree under $t\alpha$ and under $t\beta$. Thus we might expect that $\hat{x}(t|t+)$ could be computed from the one-step-backward-predicted estimates $\hat{x}(t|t\alpha)$ and $\hat{x}(t|t\beta)$ of $x(t)$ based separately on the information in the subtrees with root $t\alpha$ and root $t\beta$, respectively. Indeed as shown in [16, 19]

$$\hat{x}(t|t) = P(t|t+)[P^{-1}(t|t\alpha)\hat{x}(t|t\alpha) + P^{-1}(t|t\beta)\hat{x}(t|t\beta)] \tag{4.37}$$

$$P(t|t+) = [P^{-1}(t|t\alpha) + P^{-1}(t|t\beta) - P_x^{-1}(t)]^{-1} \tag{4.38}$$

Finally to complete the recursion, $\hat{x}(t|t\alpha)$ and $\hat{x}(t|t\beta)$ are computed from $\hat{x}(t\alpha|t\alpha)$ and $\hat{x}(t\beta|t\beta)$, respectively, in identical fashions. Specifically, each of these calculations represents a one-step-backward prediction. It is not surprising, then that a backward version of the model (4.14) plays a role here. Indeed, as shown in [16]

$$\hat{x}(t|t\alpha) = F(t\alpha)\hat{x}(t\alpha|t\alpha) \tag{4.39}$$

$$P(t|t\alpha) = F(t\alpha)P(t\alpha|t\alpha)F^T(t\alpha) + \mathcal{Q}(t\alpha) \tag{4.40}$$

where

$$F(t) = A^{-1}(t)[I - B(t)B^T(t)P_x^{-1}(t)] \tag{4.41}$$

$$\mathcal{Q}(t) = A^{-1}(t)B(t)Q(t)B^T(t)A^{-T}(t) \tag{4.42}$$

$$Q(t) = I - B^T(t)P_x^{-1}(t)B(t) \tag{4.43}$$

The prediction (4.39-4.43) and update (4.33-4.36) steps correspond to the analogous steps in the usual Kalman filter (although here we *must* use the backward model in

the prediction step), while the fusion step (4.37)-(4.38) has no counterpart in usual Kalman filtering. The interpretation of (4.37)-(4.38) is that we are fusing together two estimates each of which incorporates one set of information that is independent of that used in the other—i.e. the measurements in the $t\alpha$ and $t\beta$ subtrees– and one common information source, namely the prior statistics of $x(t)$. Eq. (4.38) ensures that this common information is accounted for only once in the fused estimate. Once the top of the overall tree is reached we, of course, have the optimal smoothed estimate at that node. As shown in [16, 18, 19], it is then possible to compute the optimal smoothed estimated in a recursive fashion moving down the tree, from coarse to fine. This recursion combines the smoothed estimate $\hat{x}_s(t\overline{\gamma})$ with the filtered estimates from the upward sweep to produce $\hat{x}_s(t)$:

$$\hat{x}_s(t) = \hat{x}(t|t) + P(t|t)F^T(t)P^{-1}(t\overline{\gamma}|t)[\hat{x}_s(t\overline{\gamma}) - \hat{x}(t\overline{\gamma}|t)] \tag{4.44}$$

Note that this algorithm also has a highly parallel, and in this case pyramidal, structure, since all calculations, on either the fine-to-coarse or coarse-to-fine sweep can be computed in parallel.

Equations (4.34-4.36), (4.38), and (4.40-4.43) define, in essence a Riccati equation on the dyadic tree. As for standard Riccati equations, it is possible to relate properties of the solution of this equation to system-theoretic properties. For example, one can show that suitably defined notions of uniform complete reachability and uniform complete observability imply upper and lower positive-definite bounds on the error covariance. Here since the Riccati equation propagates up the tree, the analysis of reachability and observability relate to systems defined recursively from fine-to-coarse scale—i.e. noncausal systems as in the first two equations of (4.10). One might also expect that one could obtain results on the stability of the error dynamics and asymptotic behavior in the constant parameter case. This is indeed the case, but there are several issues that complicate the analysis. Specifically, in standard Kalman filtering analysis the Riccati equation for the error covariance can be viewed simply as the covariance of the error equation, which can be analyzed directly without explicitly examining the state dynamics, since the error evolves as a state process itself. This is not the case here in general. First, while the process $x(t)$ is defined recursively moving

down the tree, the filtered estimate $\hat{x}(t|t)$ is defined by a recursion in the opposite direction. This difficulty cannot be overcome in general simply by reversing one of these processes, as the reversal process does not, in general, produce a system driven by white noise.[9] Also, unlike the standard situation, our Riccati equation explicitly involves the prior state covariance $P_x(t)$, arising as we've seen to prevent the double counting of prior information.

There is, however, a way in which these difficulties can be avoided, essentially by setting P_x^{-1} to zero. In particular, as discussed in [16, 18] if we do this in (4.33)-(4.43), the estimates produced have the interpretation as maximum likelihood (ML) estimates. A variation of the RTS algorithm we have described here uses this ML procedure to propagate to the top of the tree, at which point prior information is then incorporated, followed by the coarse-to-fine sweep (4.44). To see what happens to the Riccati equation and error dynamics in this case, let us focus on the scale-varying case, i.e. the case in which all parameters depend only on $m(t)$. In this case the same is true of the error covariances, yielding the following Riccati equation in scale:

$$\begin{aligned} P_{ML}(m|m+1) &= A^{-1}(m+1)P_{ML}(m+1|m+1)A^{-T}(m+1) \\ &+ G(m+1)Q(m+1)G^T(m+1) \qquad (4.45) \\ P_{ML}^{-1}(m|m) &= 2P_{ML}^{-1}(m|m+1) + C^T(m)R^{-1}(m)C(m) \qquad (4.46) \end{aligned}$$

where

$$G(m) = -A^{-1}(m)B(m) \qquad (4.47)$$

This Riccati equation differs from the usual equation only in the presence of the factor of 2 in (4.46), representing the doubling of information arising in the fusion step. In this case we can also write a direct fine-to-coarse state form for the ML estimation error $\tilde{x}_{ML}(t|t) = x(t) - \hat{x}_{ML}(t|t)$:

$$\tilde{x}_{ML}(t|t) = \frac{1}{2}(I - K_{ML}(m(t))C(m(t)))A^{-1}(m(t)+1)(\tilde{x}_{ML}(\alpha t|\alpha t) + \tilde{x}_{ML}(\beta t|\beta t))$$

[9] In particular the backward models used in [16, 18, 19] to write $x(t)$ in terms of $x(t\alpha)$ and in terms of $x(t\beta)$ yield driving noises which are martingale differences *with respect to the partial order defined on the tree.*

$$- \frac{1}{2}(I - K_{ML}(m(t))C(m(t)))G(m(t)+1)(w(\alpha t) + w(\beta t)) - K_{ML}(m(t))v(t) \tag{4.48}$$

$$K_{ML}(m) = P(m|m)C^T(m)R^{-1}(m) \tag{4.49}$$

In [16, 18] we provide a detailed analysis of (4.45)-(4.49). In particular the stability of the error dynamics (4.48) under reachability and observability conditions is established. The notion of stability, however, deserves further comment. Intuitively what we would like stability to mean is that the state of the recursion up the tree decays to 0 as we propagate farther and farther away from the initial level of the tree. Note, however, that as we move up the tree the state at any node is influenced by a geometrically increasing number of nodes at the initial level. Thus in order to study asymptotic stability it is *necessary* to consider an infinite dyadic tree, with an infinite set of initial conditions corresponding to all nodes at the initial level. The implications of this are most easily seen in the constant parameter case. In this case we have that if (A, B) is a reachable pair and (C, A) observable, then

$$\begin{aligned} \overline{P}_\infty &= \frac{1}{2}A^{-1}\overline{P}_\infty A^{-T} + \frac{1}{2}GQG^T \\ &- K_\infty(\frac{1}{2}CA^{-1}\overline{P}_\infty A^{-T}C^T + \frac{1}{2}CGQG^TC^T + R)K_\infty^T \end{aligned} \tag{4.50}$$

where

$$K_\infty = \overline{P}_\infty C^T R^{-1} \tag{4.51}$$

Moreover, the autonomous dynamics of the steady-state ML filter, i.e.

$$e(t) = \frac{1}{2}(I - K_\infty C)A^{-1}(e(\alpha t) + e(\beta t)) \tag{4.52}$$

is exponentially l_2 stable, i.e. the l_2 norm of all values of $e(t)$ along an entire horocycle converges exponentially to zero as $m(t) \to 0$. As shown in [16, 18] this is equivalent to all eigenvalues of the Kalman filter error dynamics matrix

$$\frac{1}{2}(I - K_\infty C)A^{-1} \tag{4.53}$$

having magnitude less than $\frac{\sqrt{2}}{2}$.

5 Conclusions

In this paper we have outlined a mathematical framework for the multiresolution modeling and analysis of stochastic processes. As we have discussed, the theory of multiscale signal analysis and wavelet transforms leads naturally to the investigation of multiscale statistical representations and dynamic models on dyadic trees and lattices. The rich structure of the dyadic tree requires that we take some care in the specification of such models and in the generalization of standard time series notions. In particular, we have seen that in this context there are two natural concepts of shift invariance which provide new ways in which to capture notions of scale-invariant statistical descriptions. In addition, the observation that the scale variable is time-like in nature leads to a natural notion of "causal" dynamics in scale: from fine to coarse; however the tree provides only a partial ordering of points, requiring that we take some care in defining the "past".

In part of our work we have described the multiscale autoregressive modeling of isotropic processes, *i.e.* processes satisfying our stronger notion of statistical shift-invariance. As we have seen, the usual AR representation of time series is not a particularly convenient one thanks both to the geometric explosion of points in the "past" as we increase system order and to the nonlinear constraints isotropy imposes on the AR coefficients. In contrast, we have seen that it is possible to construct a generalization of the reflection-coefficient-based lattice representation for such models, including generalized Levinson and Schur recursions. As we have illustrated such models can be used to generate fractal-like signals.

The other part of our work was motivated by our weaker notion of stationarity which in essence says that the correlation between two values in our multiscale representation depends on the difference in scale and location of the two points. As we have seen, this framework leads to state models evolving from coarse-to-fine scales on dyadic trees. We have described some of our work on a basic system theory for such models and have also discussed an estimation framework that allows us to capture the fusion of measurements at differing resolutions. In addition the structure of these models leads to several extremely efficient and highly parallel estimation structures:

a multiscale iterative algorithm that can be arranged so as to have the same form as well-known multigrid algorithms for solving partial differential equations; an algorithm using wavelet transforms to decouple the estimation procedure into a large set of far simpler parallel estimation algorithms; and a pyramidal algorithm that introduces a generalization of the Kalman filter and the associated Riccati equation.

As we have discussed and illustrated, these models appear to be useful for a rich variety of processes including the 1/f-like models as introduced in [50, 51] and standard first-order Gauss-Markov processes. Much, of course, remains to be done in developing this theory, in investigating the processes that can be conveniently and accurately represented within this framework, and in applying these results to problems of practical importance such as sensor fusion, noise rejection, multisensor or multiframe data registration and mapping, and segmentation. Among the theoretical topics under investigation are the development of model fitting and likelihood function-based methods for parameter estimation and segmentation and the development of a detailed theory of approximation of stochastic processes including a specification of those processes that can be "well"-approximated by models of the type we have introduced. Of particular interest is the dynamic interpretation of so-called wave packet transforms [21] in which the wavelet coefficients are subjected to further decomposition through the same filter pair used in the wavelet transform. Viewing this from our dynamic synthesis perspective, this would appear to correspond to a class of higher-order models. Identifying and analyzing this model class, however, remains for the future.

References

[1] B. Anderson and T. Kailath, "Forwards, backwards, and dynamically reversible Markovian models of second-order processes," *IEEE Trans. Circuits and Systems*, CAS-26, no. 11, 1978, pp. 956-965.

[2] J. Arnaud, "Fonctions spheriques et fonctions definies-positives sur l'arbre homogene", C.R. Acad. Sc., Serie A, 1980, pp. 99-101.

[3] J. Arnaud and B. Letac, "La formule de representation spectrale d'un processus gaussien stationnaire sur un arbre homogene," Laboratoire de Stat. et. Prob.-U.A.-CNRS 745, Toulouse.

[4] S. Attasi, "Modeling and Recursive Estimation for Double Indexed Sequences," in *System Identification: Advances and Case Studies*, R.K. Mehra and D.G. Lainiotis, eds., Academic Press, NY 1976.

[5] M. Barnsley, *Fractals Everywhere*, Academic Press, San Diego, 1988.

[6] M. Basseville, A. Benveniste, and A.S. Willsky "Multiscale Autoregressive Processes, Part I: Schur-Levinson Parametrizations", submitted to *IEEE Transactions on ASSP.*

[7] M. Basseville, A. Benveniste, and A.S. Willsky "Multiscale Autoregressive Processes, Part II: Lattice Structures for Whitening and Modeling", submitted to *IEEE Transactions on ASSP.*

[8] M. Basseville, A. Benveniste, A.S. Willsky, and K.C. Chou, "Multiscale Statistical Processing: Stochastic Processes Indexed by Trees," in *Proc. of Int'l Symp. on Math. Theory of Networks and Systems*, Amsterdam, June 1989.

[9] A. Benveniste, R. Nikoukhah, and A.S. Willsky, "Multiscale System Theory", Proceedings of the 29th IEEE Conference on Decision and Control, Honolulu, HI, December 1990.

[10] G. Beylkin, R. Coifman, and V. Rokhlin, "Fast Wavelet Transforms and Numerical Algorithms I", to appear in *Comm. Pure and Appl. Math.*

[11] A. Brandt, "Multi-level adaptive solutions to boundary value problems," Math. Comp. Vol. 13,1977, pp.333-390.

[12] W. Briggs, "*A Multigrid Tutorial*, SIAM, Philadelphia, PA, 1987.

[13] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Comm.*, vol. 31, pp. 482–540, 1983.

[14] P. Cartier, "Harmonic analysis on trees", Proc. Symos. Pure Math., Vol 26, Amer. Math. Soc. Providence, R.I., 1974, pp.419-424.

[15] P. Cartier, "Geometrie et analyse sur les arbres",Seminaire Bourbaki, 24eme annee, Expose no. 407, 1971/72.

[16] K.C. Chou, *A Stochastic Modeling Approach to Multiscale Signal Processing*, MIT, Department of Electrical Engineering and Computer Science, Ph.D. Thesis, (in preparation).

[17] K.C. Chou, S. Golden and A.S. Willsky, "Modeling and Estimation of Multiscale Stochastic Processes", *Int'l Conference on Acoustics, Speech, and Signal Processing"*, Toronto, April 1991.

[18] K.C. Chou and A.S. Willsky, ''Multiscale Riccati Equations and a Two-Sweep Algorithm for the Optimal Fusion of Multiresolution Data", *Proceedings of the 29th IEEE Conference on Decision and Control*, Honolulu, HI, December 1990.

[19] K.C. Chou, A.S. Willsky, A. Benveniste, and M. Basseville,"Recursive and Iterative Estimation Algorithms for Multi-Resolution Stochastic Processes," Proc. 28th IEEE Conf. on Dec. and Cont., Tampa, Dec. 1989.

[20] S.C. Clippingdale and R.G. Wilson, "Least Squares Image Estimations on a Multiresolution Pyramid", Proc. of the 1989 Int'l Conf. on Acoustics, Speech, and Signal Proceeding.

[21] R.R. Coifman, Y. Meyer, S. Quake and M.V. Wickehauser, "Signal Processing and Compression with Wave Packets", preprint, April 1990.

[22] I. Daubechies,"Orthonormal bases of compactly supported wavelets", Comm. on Pure and Applied Math. 91, 1988, pp. 909-996.

[23] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. on Information Theory*, 36, 1990, pp. 961-1005.

[24] I. Daubechies, A. Grossman, and Y. Meyer, "Painless non-orthogonal expansions," J. Math. Phys. 27, 1986, pp. 1271-1283.

[25] P. Flandrin, "On the Spectrum of Fractional Brownian Motions", *IEEE Transactions on Information Theory*, Vol. 35, 1989, pp. 197-199.

[26] J. Goodman and A. Sokal, "Multi-grid Monte Carlo I. conceptual foundations," Preprint, Dept. Physics New York University, New York, Nov. 1988; to be published.

[27] S. Golden, *Identifying Multiscale Statistical Models Using the Wavelet Transform*, S.M. Thesis, M.I.T. Dept. of EECS, May 1991.

[28] A. Grossman and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape", SIAM J. Math. Anal. 15, 1984, pp. 723-736.

[29] W. Hackbusch and U. Trottenberg, Eds., *Multigrid Methods and Applications*, Springer-Verlag, N.Y., N.Y., 1982.

[30] T. Kailath, "A theorem of I. Schur and its impact on modern signal processing", in *Schur Methods in Operator Theory and Signal Processing*, I. Gohberg Ed., Operator theory: advances and Applications, Vol. 18, Birkhäuser (Basel, Boston, Stuttgart), 1986.

[31] M. Kim and A.H. Tewfik, "Fast Multiscale Detection in the Presence of Fractional Brownian Motions", *Proceedings of SPIE Conference on Advanced Algorithms and Architecture for Signal Processing V*, San Diego, CA, July 1990.

[32] T. Lin, M. Kawamata and T. Higuchi, "New necessary and sufficient conditions for local controllability and observability of 2-D separable denominator systems," *IEEE Trans. Automat. Control*, AC-32, pp. 254-256, 1987.

[33] S.G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Transactions on Pattern Anal. and Mach. Intel.*, Vol. PAMI-11, July 1989, pp. 674-693.

[34] S.G. Mallat, "Multifrequency Channel Decompositions of Images and Wavelet Models", *IEEE Transactions on ASSP*, Vol. 37, December 1989, pp. 2091-2110.

[35] B. Mandelbrot, *The Fractal Geometry of Nature,* Freeman, New York, 1982.

[36] B. B. Mandelbrot and H.W. Van Ness, "Fractional Brownian Motions, Fractional Noises and Applications", *SIAM Review*, Vol. 10, October 1968, pp. 422-436

[37] S. McCormick, *Multigrid Methods*, Vol. 3 of the SIAM Frontiers Series, SIAM, Philadelphia, 1987.

[38] Y. Meyer, "L'analyse par ondelettes", Pour la Science, Sept. 1987.

[39] D. Paddon and H. Holstein, Eds. *Multigrid Methods for Integral and Differential Equations*, Clarendon Press, Oxford, England, 1985.

[40] A.J. Pentland, "Fast Surface Estimation Using Wavelet Bases", MIT Media Lab Vision and Modeling Group TR-142, June 1990.

[41] A.P. Pentland, "Fractal-Based Description of NaturalScenes", *IEEE Transactions on Patt. Anal. and Mach. Intel.,* Vol. PAMI-6, November 1989, 661-674.

[42] H. E. Rauch, F. Tung, and C. T. Striebel, "Maximum Likelihood Estimates of Linear Dynamic Systems," AIAA Journal, Vol. 3, No. 8, Aug. 1965, pp. 1445-1450.

[43] E.A. Robinson, S. Treitel, "Maximum Entropy and the Relationship of the Partial Autocorrelation to the Reflection Coefficients of a Layered System," *IEEE Trans. on ASSP*, vol. 28 Nr 2, 224–235, 1980.

[44] M.J. Smith and T.P. Barnwell, "Exact reconstruction techniques for tree-structured subband coders", IEEE Trans. on ASSP 34, 1986, pp. 434-441.

[45] R. Szeliski, "Fast Surface Interpolation Using Hierarchical Basis Function", *IEEE Transactions on PAMI*, Vol. 12, No. 6, June 1990, pp. 513-528.

[46] D. Terzopoulos, "Image Analysis Using Multigrid Relaxation Methods", *IEEE Transaction on PAMI*, Vol. PAMI-8, No. 2, March 1986, pp. 129-139.

[47] A.H. Tewfik and M. Kim, "Correlation Structure of the Discrete Wavelet Coefficients of Fractional Brownian Motions", submitted to *IEEE Transactions on Information Theory.*

[48] M. Todd and R. Wilson, "An Anisotropic Multi-Resolution Image Data Compression Algorithm", Proc. of the 1989 Int'l Conf. on Acoustics, Speech, and Signal Processing.

[49] M. Vetterli, and C. Herley, "Wavelet and Filter Banks: Relationships and New Results", *Proceedings of the ICASSP*, Albuquerque, NM, 1990.

[50] G.W. Wornell, "A Karhunen-Loeve-Like Expansion for 1/f Processes via Wavelets", *IEEE Transactions on Information Theory*, Vol. 36, No. 9, July 1990, pp. 859-861.

[51] G.W. Wornell and A.V. Oppenheim, "Estimation of Fractal Signs from Noisy Measurements Using Wavelets", submitted to *IEEE Transactions on ASSP.*

[52] A.S. Willsky, K.C. Chou, A. Benveniste, and M. Basseville,"Wavelet Transforms, Multiresolution Dynamical Models, and Multigrid Estimation Algorithms", *1990 IFAC World Congress*, Tallinn, USSR, August 1990.

[53] A. Witkin, D. Terzopoulos and M. Kass, "Signal Matching Through Scale Space", *Int. J. Comp. Vision*, Vol 1, 1987, pp. 133-144.

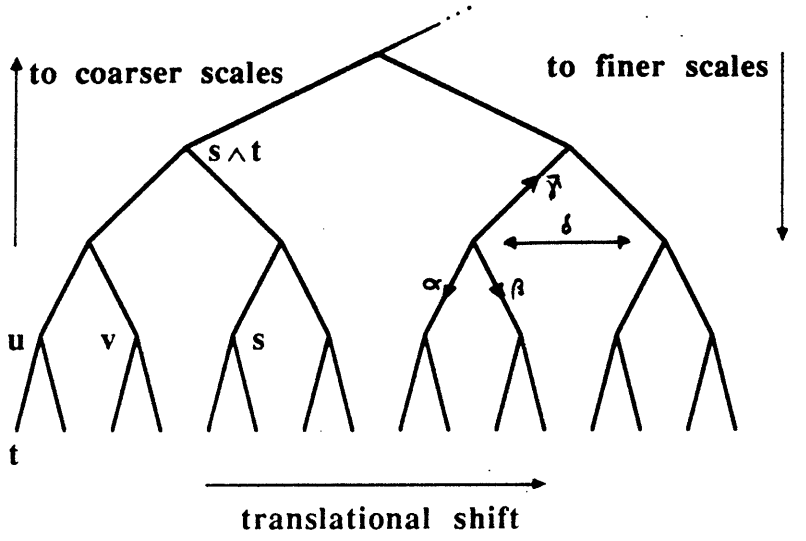


Figure 1: The dyadic tree, in which each level of the tree corresponds to a single scale in a multiscale representation. The nodes here correspond to scale/shift pairs (m, n).

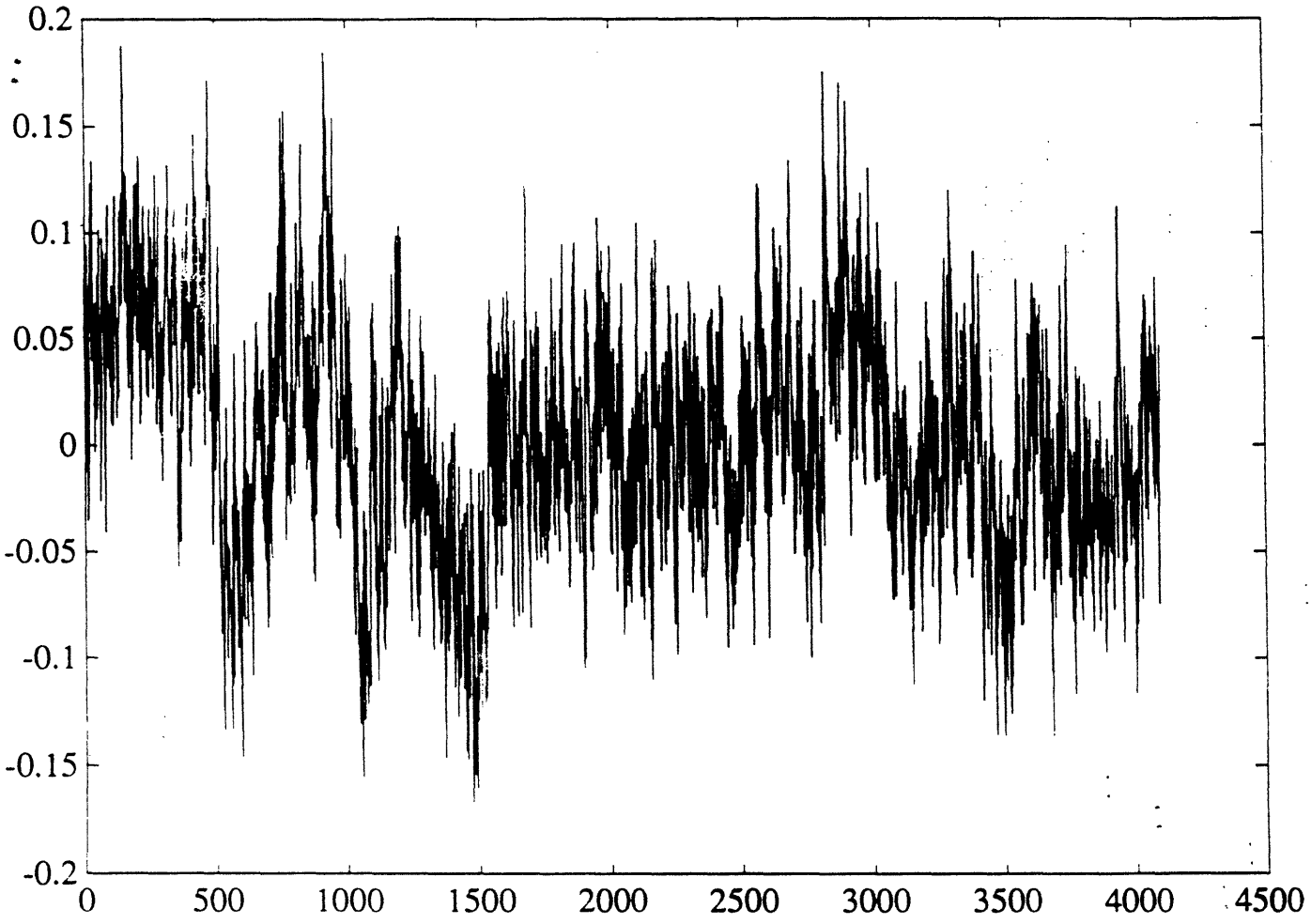


Figure 2: A signal generated by a third-order multiscale autoregressive model, as described in Section 3.

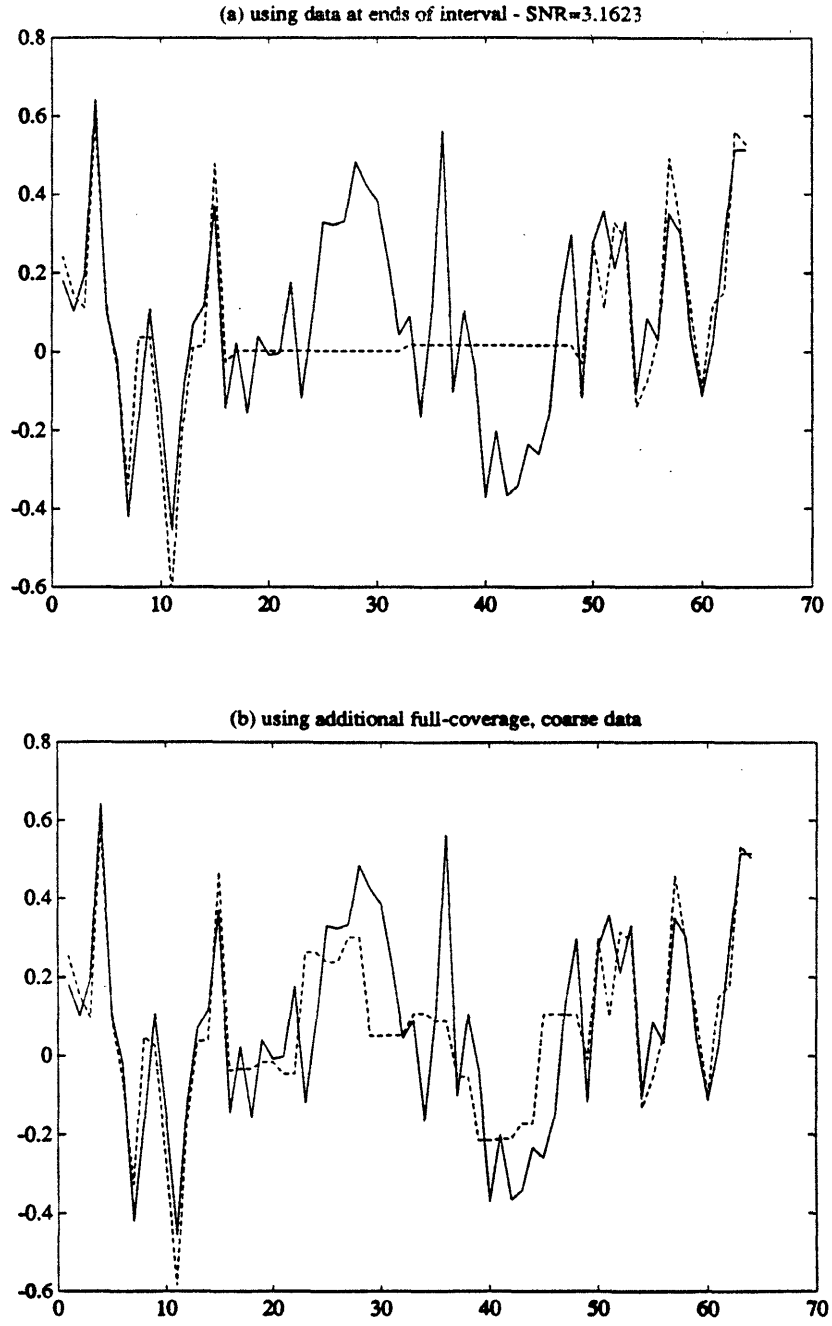
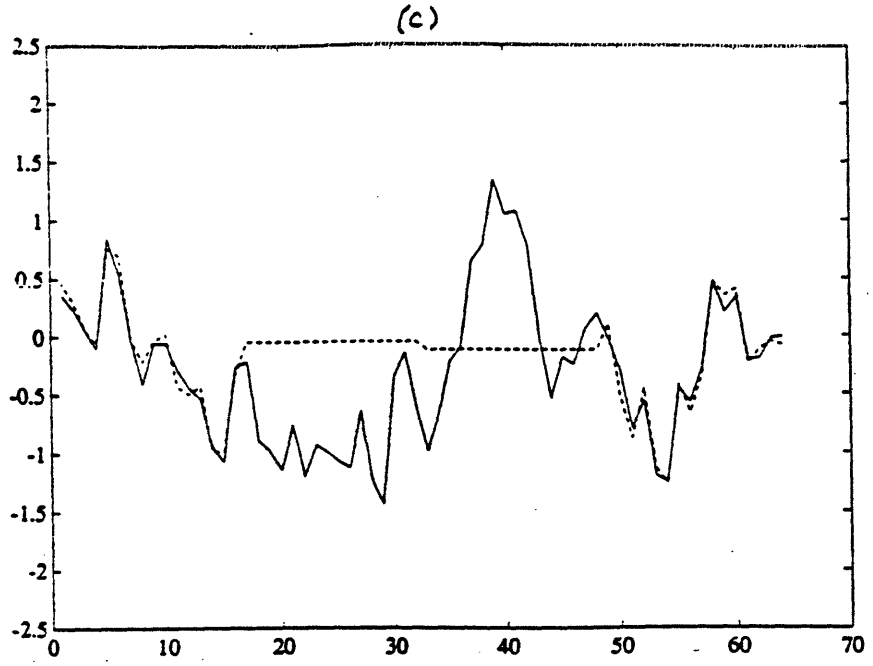


Figure 3: Illustrating multiscale data fusion using the techniques described in Section 4. In (a) and (b) a signal with a 1/f-like spectrum (as described in [50]), shown as a solid line in both plots, is reconstructed based on measurements. In (a) data is available only at the two ends of the interval, while in (b) coarse scale (i.e. locally averaged) measurements are fused to improve signal interpolation. In (c) and (d) analogous results are shown for the multiscale data fusion and interpolation of a Gauss-Markov process.



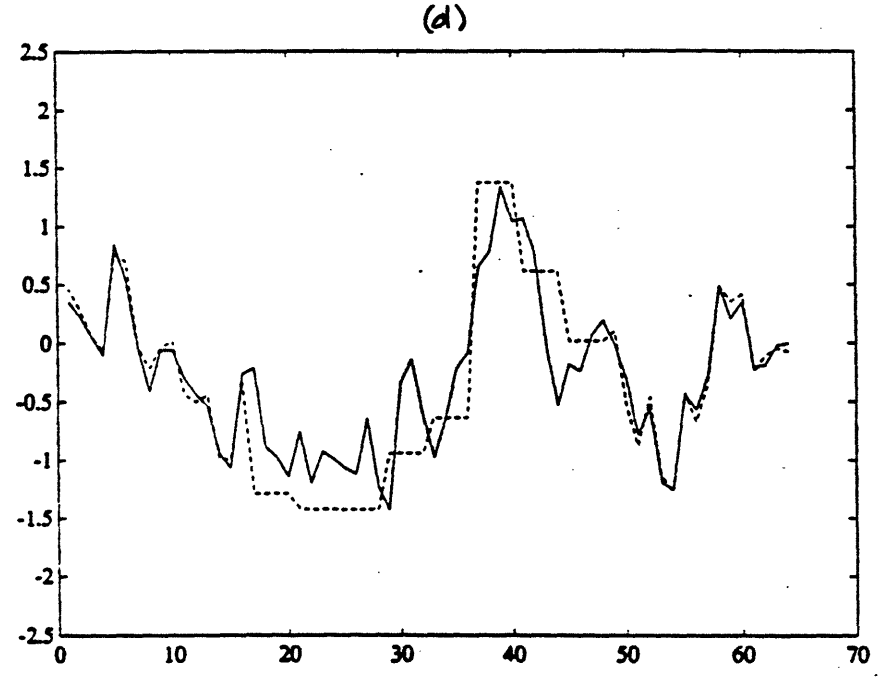


Figure 3: (continued)

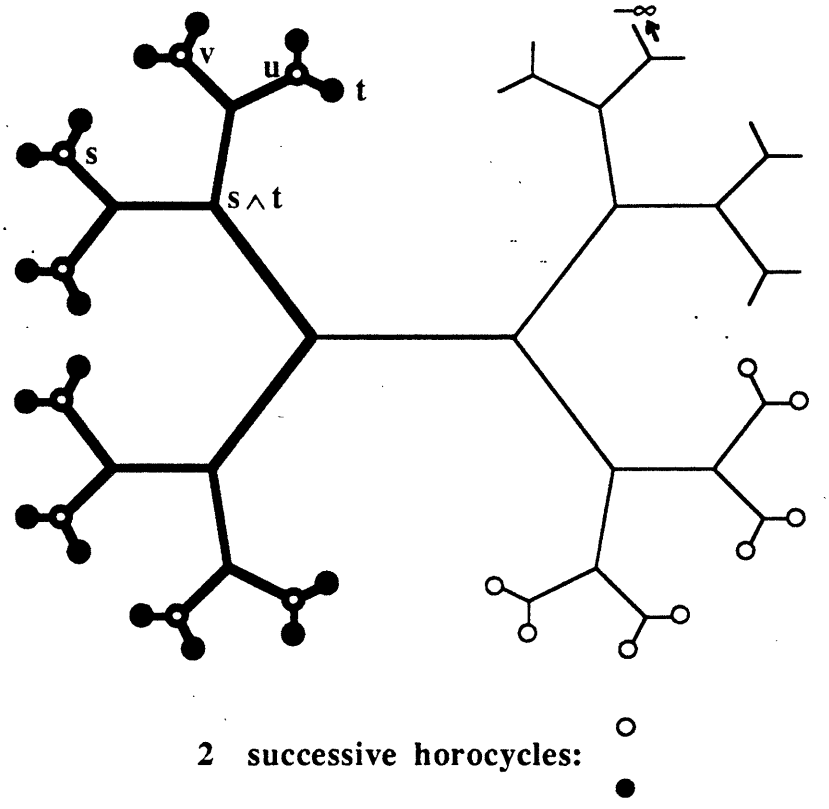


Figure 4: A more symmetric depiction of the dyadic tree, illustrating the notion of a boundary point $-\infty$, horocycles, and the "parent" $s \wedge t$ of nodes s and t (see the text for explanations).

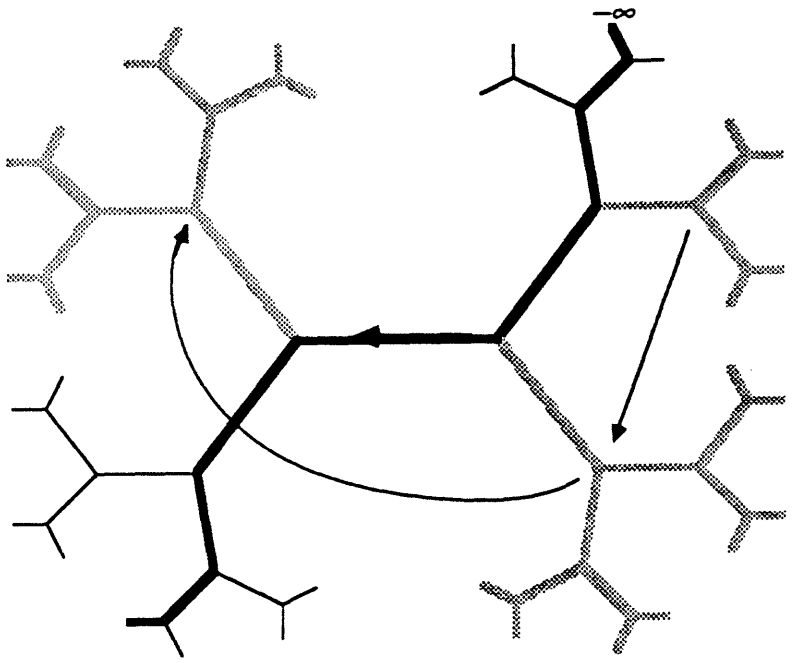


Figure 5: Illustrating (in bold) the skeleton of a translation. As indicated in the figure, any translation with this skeleton must map the subtree extending away from any node on the skeleton onto the corresponding subtree of the next node. There are, however, many ways in which this can be done (e.g. by "pivoting" isometries within any of these subtrees).

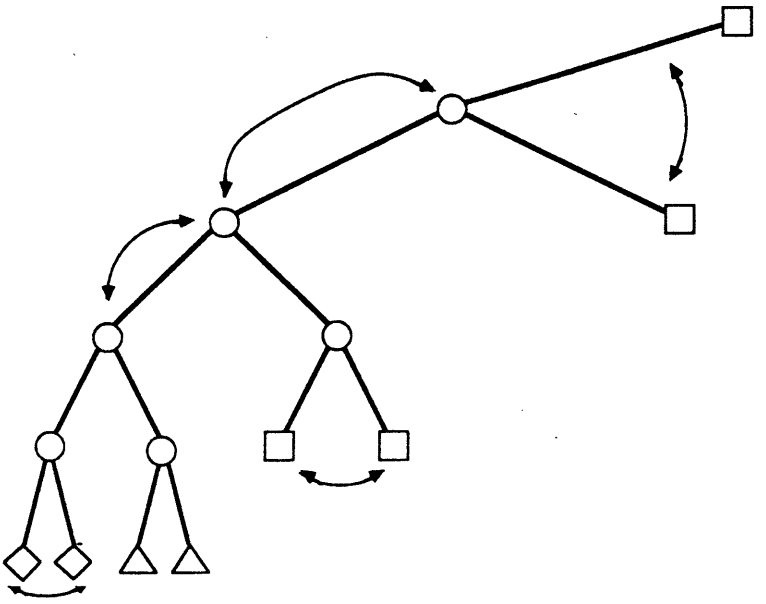


Figure 6: Illustrating the nature of the construction required in developing recursions for $E_{t,n}$ and $F_{t,n}$. Here if t is the node in the lower left-hand corner, then the elements of $E_{t,4}$ are the prediction errors at the two points indicated by diamonds given the data $\mathcal{Y}_{t\overline{\gamma},3}$ spanned by the circles. The elements of $F_{t\overline{\gamma},4}$ are the prediction errors at the four points indicated by squares given again the data in $\mathcal{Y}_{t\overline{\gamma},3}$. The elementary "pivoting" isometries indicated in the figure allow us to obtain the result on PARCOR coefficients described in the text.

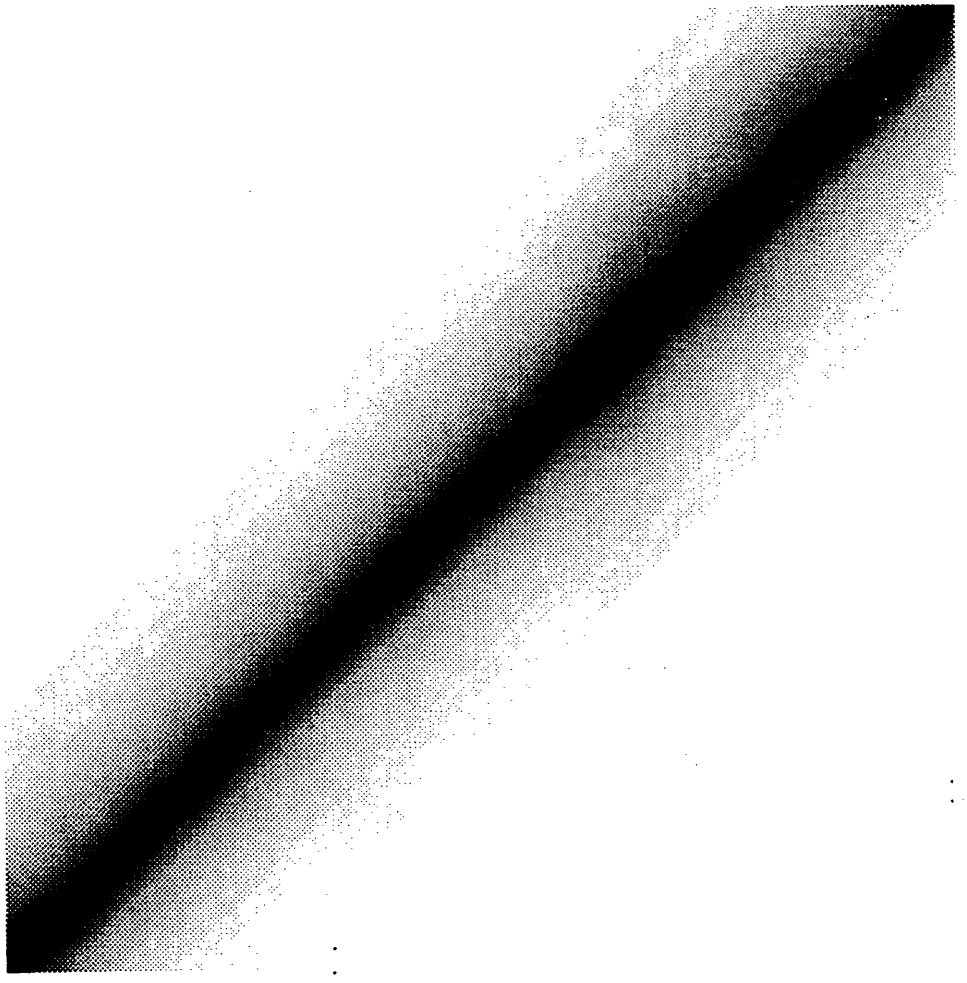


Figure 7: Illustrating the covariance matrix of a set of samples of a first-order Gauss-Markov process with covariance of the form $\exp^{-\alpha|t|}$. Black corresponds to a value of 1 with lighter shades representing smaller values. The covariance of this process decays exponentially as we move away from the main diagonal.

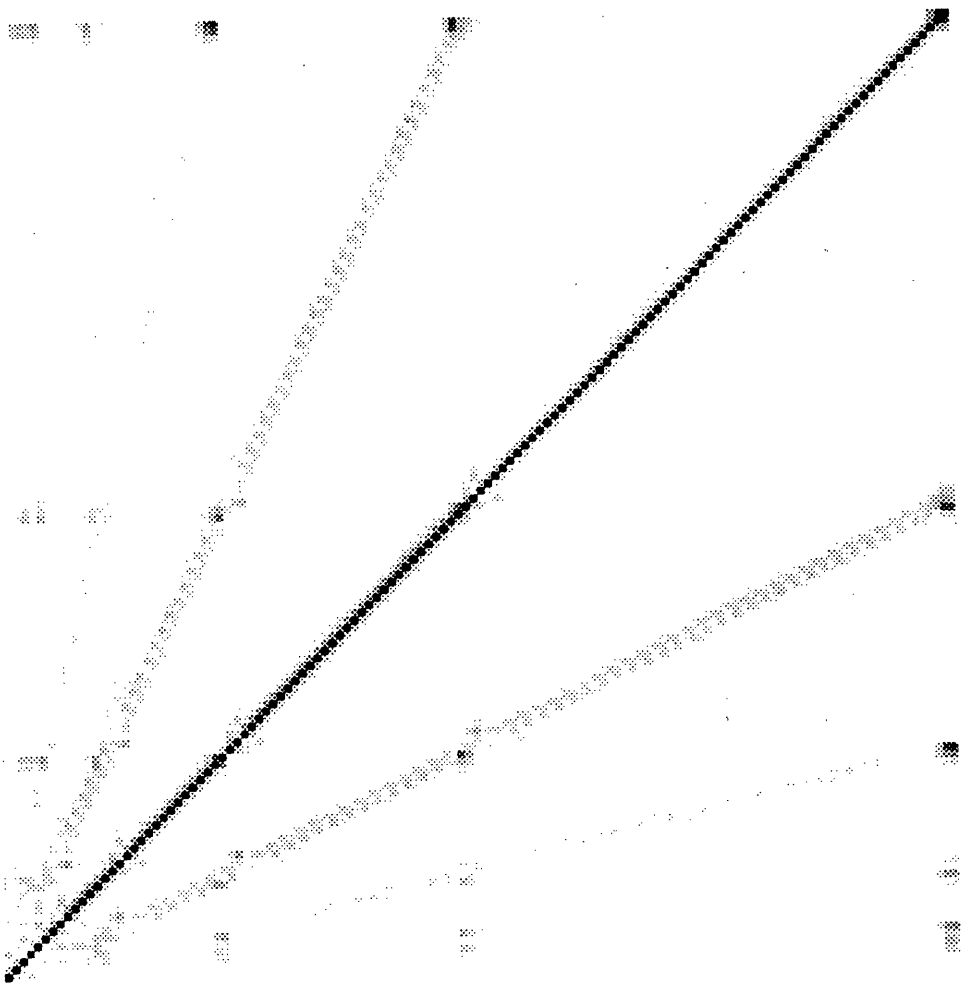


Figure 8: The matrix of correlation coefficients (i.e. covariance divided by the square root of the product of variances) for the wavelet-transform of the Gauss-Markov process of Figure 7 using an 8-tap QMF.

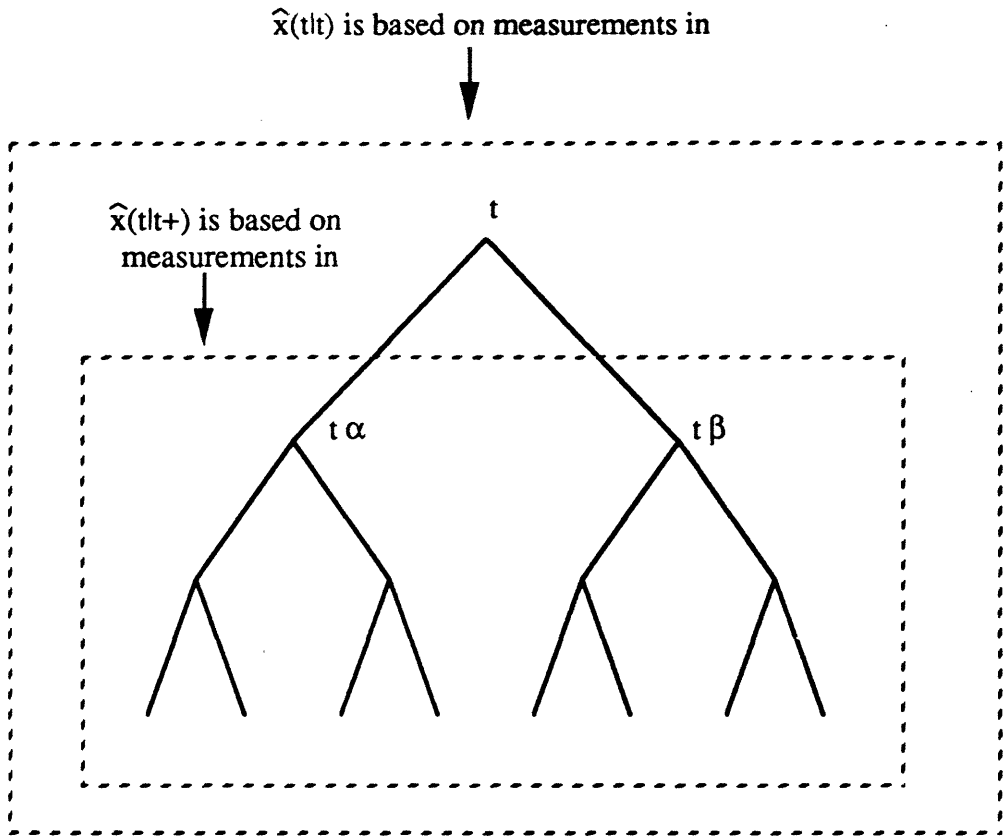


Figure 9: Illustrating the measurement sets used for the esimates $\hat{x}(t|t)$ and $\hat{x}(t|t+)$.