LIDS-P-1893

1990 IFAC World Congress

WAVELET TRANSFORMS, MULTIRESOLUTION DYNAMICAL MODELS, AND MULTIGRID ESTIMATION ALGORITHMS

A.S. Willsky, K.C. Chou[1]
Laboratory for Information and Decision Systems, MIT, Cambridge, MA 02139
A. Benveniste, M. Basseville[2]
Institut de Recherche en Informatique et Systemes Aleatoires,
Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France

Abstract The work described in this paper is motivated by the need for a probabilistic theory for multiresolution stochastic models which can then provide the basis for optimal multiscale signal processing algorithms. As we develop in this paper, wavelet transforms and multiscale representations lead naturally to the study of stochastic processes indexed by nodes on lattices and trees, where different depths in the tree or lattice correspond to different spatial scales or resolutions in representing the signal. This modeling paradigm also has close ties to self-similar stochastic processes, fractals, and renormalization concepts. Using this framework we introduce several classes of dynamic models for multiscale stochastic processes in which the direction of recursion is from coarse-to-fine resolution. This leads to a theory of optimal estimation for multiresolution stochastic models. Some of the algorithms that arise in this context have a multigrid relaxation structure while others lead to new classes of Riccati equations involving the usual predict and update steps and a new "merge" step as information is propagated from fine-to-coarse scales. This framework also allows us to solve problems of the optimal fusion of multispectral(and hence multiresolution) data. In addition a theory of modeling for multiscale processing is in the process of being developed. Generalizations will be described of methods such as Levinson's algorithm for recursive generation of multiscale models of increasing order.

Keywords. Multi-scale systems; trees; smoothing; multidimensional systems; signal processing; optimal estimation; lattice structures; parameter estimation; Markov processes.

1 MULTISCALE REPRESENTATIONS AND SYSTEMS ON TREES

In the past few years there has been a renewed interest in multiscale representations of signals in one or several dimensions, thanks, in large part to the emerging theory of wavelet transforms (see, for example, Daubechies(1988), Mallat(1987)). The development of optimal multiscale signal processing algorithms - e.g. for the reconstruction of noise-degraded signals or for the detection and localization of transient signals of different durations - requires the development of a corresponding theory of stochastic processes and their estimation. The research presented in this and several other papers and reports(Chou,1990;Basseville,1989) has the development of this theory as its objective.

The multi-scale representation of a continuous signal $f(x)$ consists of a sequence of approximations of that signal at finer and finer scales where the approximation of $f(x)$ at the mth scale is given by

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

As $m \to \infty$ the approximation consists of a sum of many highly compressed, weighted, and shifted versions of the function $\phi(x)$ whose choice is far from arbitrary. In particular in order for the $(m+1)$st approximation to be a refinement of the mth, we require that $\phi(x)$ be exactly representable at the next scale:

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

Furthermore in order for (1.1) to be an orthogonal series, $\phi(x)$ and its integer translates must form an orthogonal set. As shown in Daubechies(1988), $h(n)$ must satisfy several conditions for this and several other properties of the representation to hold. In particular $h(n)$ must be the impulse response of a quadrature mirror filter. 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

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

Multiscale represention are closely related to wavelet transforms. Such a transform is based on a single function $\psi(x)$ that has the property that the full set of its scaled translates $\{2^{m/2}\psi(2^m x - n)\}$ form a complete orthonormal basis for L^2. Daubechies(1988) shows 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, and that

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

$f_m(x)$ 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 eq. (1.3), eq. (1.4), then

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

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

From the preceding remarks we see that we have 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 $f(m+1,k)$. In particular the Haar representation naturally defines a dyadic tree structure on the points (m,n) in which each point has two descendents corresponding to the two subdivisions of the support interval of $\phi(2^m x - n)$, namely those of $\phi(2^{(m+1)}x - 2n)$ and $\phi(2^{(m+1)}x - 2n - 1)$. This observation provides our motivation for the development of models for stochastic processes on dyadic trees as the basis for a statistical theory of multiresolution stochastic processes. In the next section we describe a class of dynamic state models on dyadic trees and discuss the elements of the estimation and system theory for these models. In Section 3 we describe our research on models for the class of isotropic processes. We conclude in Section 4 with a discussion of research in progress at the present time.

[1]The work of these authors was supported in part by the Air Force Office of Scientific Research under Grant AFOSR-88-0032, in part by the National Science Foundation under Grant ECS-8700903, and in part by the US Army Research Office under Contract DAAL03-86-K-0171. In addition some of this research was performed while these authors were visitors at Institut de Recherche en Informatique et Systemes Aleatoires(IRISA), Rennes, France during which time A.S.W. received partial support from Institut National de Recherche en Informatique et en Automatique(INRIA).

[2]A.B. is also with INRIA, and M.B. is also with Centre National de la Recherche Scientifique(CNRS). The research of these authors was also supported in part by Grant CNRS GO134.

2 DYNAMIC STATE MODELS ON TREES

To begin, let us define some notation needed to describe dynamic systems on trees. Let $\mathcal{T}$ denote the index set of the tree and we use the single symbol t for nodes on the tree. The scale associated with t is denoted by $m(t)$, and we write $s \preceq t (s \prec t)$ if $m(s) \leq m(t) (m(s) < m(t))$. We also let $d(s,t)$ denote the distance between s and t, and $s \wedge t$ the common "parent" node of s and t (e.g. $(2^{-m}, n)$ is the parent of $(2^{-(m+1)}, 2n)$ and $(2^{-(m+1)}, 2n+1)$. In analogy with the shift operator z^{-1} used as the basis for describing discrete-time dynamics we also define several shift operators on the tree: 0, the identity operator (no move); γ^{-1}, the fine-to-coarse shift (e.g. from $(2^{-(m+1)}, 2n$ or $2n+1)$ to $(2^{-m}, n)$); α, the left coarse-to-fine shift ($(2^{-m}, n)$ to $(2^{-(m+1)}, 2n)$); β, the right coarse-to-fine shift ($(2^{-m}, n)$ to $(2^{-(m+1)}, 2n+1)$); and δ, the exchange operator ($(2^{-(m+1)}, 2n) \longrightarrow (2^{-(m+1)}, 2n+1)$). Note that 0 and δ are **isometries** in that they are one-to-one, onto maps of $\mathcal{T}$ that preserve distances. Also we have the relations

$$\delta^2 = \gamma^{-1}\alpha = \gamma^{-1}\beta = 0, \gamma^{-1}\delta = \gamma^{-1}, \delta\beta = \alpha \tag{2.1}$$

We also define for convenience the move $\delta^{(n)}$ which exchanges the nth bit:

$$\begin{aligned} &\text{If } t = \alpha\gamma^{-1}t, \text{ then } \delta^{(n)}t = \alpha\delta^{(n-1)}\gamma^{-1}t \\ &\text{If } t = \beta\gamma^{-1}t, \text{ then } \delta^{(n)}t = \beta\delta^{(n-1)}\gamma^{-1}t \end{aligned} \tag{2.2}$$

As in the synthesis description of multi-scale representations, we consider the following class of state-space models on trees, evolving from coarse-to-fine scales.

$$x(t) = A(m(t))x(\gamma^{-1}t) + B(m(t))w(t) \tag{2.3}$$

where $w(t)$ is a vector white noise process with covariance I. The model (2.3) describes a process that is Markov scale-to-scale and, because of this we can readily calculate its second order statistics. In particular the covariance of $x(t)$ depends only on scale and satisfies the Lyapunov equation

$$P_x(m+1) = A(m)P_x(m)A^T(m) + B(m)B^T(m) \tag{2.4}$$

and its correlation function is given by

$$\begin{aligned} K_{xx}(t,s) &= \\ E[x(t)x^T(s)] &= \Phi(m(t), m(s \wedge t))P_x(m(s \wedge t))\Phi^T(m(s), m(s \wedge t)) \end{aligned} \tag{2.5}$$

where $\phi(m,\mu)$ is the state transition matrix associated with $A(m)$. Note that if A and B are constant and A is stable, (2.4) has a steady-state solution satisfying the algebraic Lyapunov equation

$$P_x = AP_xA^T + BB^T \tag{2.6}$$

and in this case

$$\begin{aligned} K_{xx}(t,s) &= A^{d(t,s\wedge t)}P_x(A^T)^{d(s,s\wedge t)} \\ &= K_{xx}(d(t, s\wedge t), d(s, s\wedge t)) \end{aligned} \tag{2.7}$$

Note that (2.7) is different from the case of dynamic systems evolving in time, thanks to the tree structure of the index set. In particular the dependence on $d(t, s\wedge t)$ and $d(s, s\wedge t)$ captures the fact that the correlation between two samples of our multiresolution process depends both on the temporal displacement between these points and the relation between the scales of the samples. Note also that if $AP_x = P_xA^T$ (e.g. this is true in the scalar case), then $K_{xx}(t,s)$ depends only upon $d(t,s)$. Such processes are referred to as **isotropic processes** which we discuss in the next.

Consider now the estimation of $x(t)$ based on measurements

$$y(t) = C(m(t))x(t) + v(t) \tag{2.8}$$

where $v(t)$ is white noise of covariance $R(m(t))$, independent of x. In many problems we may only have data at the finest level; however in some applications such as geophysical signal processing or the fusion of multispectral data, data at multiple scales are collected and must be combined.

We now briefly describe three algorithms for computing the optimal estimate of x based on $Y = \{y(t)\}$(see Chou(1990) for details). The first of these relies heavily on the structure of the covariance of x. Specifically, let x_k denote the vector constructed by stacking the state values $x(t)$ for all t for which $m(t) = k$. The ordering of these values is given by the dyadic representation of the points at this scale. That is, the order is of the form $t, \delta t, \delta^{(2)}t, \delta\delta^{(2)}t, \delta^{(3)}t, ...$, where t is any point at the kth scale. Let $\mathcal{P}_k$ denote the covariance of x_k and let $\mathcal{P}_{k,k+1} = E[x_kx_{k+1}^T]$. These matrices have very special structure: off-diagonal blocks of geometrically increasing size consist of matrices all of whose block-elements are identical. For example, the value of $E[x(t)x^T(wt)]$ is the same for $w = \delta^{(2)}$ and $w = \delta\delta^{(2)}$ and, similarly, this quantity has a single value for $w = \alpha$ and $w = \beta$. Thus, for example

$$\tilde{\mathcal{P}}_2 = \begin{bmatrix} P_x(2) & N_1 & \vdots & N_2 & N_2 \\ N_1 & P_x(2) & \vdots & N_2 & N_2 \\ \cdots & \cdots & & \cdots & \cdots \\ N_2 & N_2 & \vdots & P_x(2) & N_1 \\ N_2 & N_2 & \vdots & N_1 & P_x(2) \end{bmatrix} \tag{2.9}$$

where all blocks are of size equal to the dimension of $x(t)$. Similarly,

$$\tilde{\mathcal{P}}_{2,3} = \begin{bmatrix} P & P & M_2 & M_2 & \vdots & M_3 & M_3 & M_3 & M_3 \\ M_2 & M_2 & P & P & \vdots & M_3 & M_3 & M_3 & M_3 \\ \cdots & \cdots & \cdots & \cdots & & \cdots & \cdots & \cdots & \cdots \\ M_3 & M_3 & M_3 & M_3 & \vdots & P & P & M_2 & M_2 \\ M_3 & M_3 & M_3 & M_3 & \vdots & M_2 & M_2 & P & P \end{bmatrix} \tag{2.10}$$

where

$$P = P_x(2) \tag{2.11}$$

An important fact is that these matrices can be block-diagonalized by the **discrete Haar transform**. For simplicity let us first describe this for the case in which x and y are scalar processes. The discrete Haar basis is an orthonormal basis for $\mathcal{R}^N$ where $N = 2^k$. The matrix V_k whose columns form this basis consists of vectors representing "dilated, translated, and scaled" versions of the vector $[1, -1]^T$. For example for $k = 3$,

$$V_3 = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & -\frac{1}{2} & 0 & \frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & \frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & 0 & \frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & 0 & -\frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & 0 & 0 & -\frac{1}{\sqrt{2}} & 0 & -\frac{1}{2} & -\frac{1}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \end{bmatrix} \tag{2.12}$$

As shown in Chou(1990),

$$\mathcal{P}_k = V_k\Lambda_kV_k^T \tag{2.13}$$

$$\tilde{\mathcal{P}}_{k,k+1}V_{k+1} = \left[0 \mid V_k\tilde{\Lambda}_k\right] \tag{2.14}$$

where Λ_k and $\tilde{\Lambda}_k$ are diagonal matrices of dimension 2^k. In the vector case we use "dilated, translated, and scaled" versions of the block matrix $[I \; -I]^T$ instead of the vector $[1, -1]^T$. It is important to note that the discrete Haar transform, i.e. the computation of V_kz can be performed in an extremely efficient manner(in the block case as well), by successive additions and subtractions of pairs of elements.

Because of this structure we can obtain an extremely efficient algorithmic structure: by taking scale-by-scale block-Haar transforms of the observations we obtain a sequence of decoupled estimation problems for the estimation of each Haar component of x_k based on corresponding components of the y's. As an illustration, consider the scalar case and the problem of predicting x_k based on an estimate $\hat{x}_{k+1}$ of x_{k+1} (it is this prediction step that is the key to the decoupling; the measurement incorporation step at any scale is readily seen to be decoupled). Let $\hat{x}_k = \mathcal{P}_{k,k+1}\mathcal{P}_{k+1}^{-1}\hat{x}_{k+1}$ denote the desired estimate and let

$$\hat{z}_k = V_k^T\hat{x}_k \tag{2.15}$$

From (2.13) and (2.14), we see that the fine scale components of $\hat{x}_k$ are unneeded at the coarser scale; i.e. only the lower half, $\hat{z}_{k+1}^2$, of $\hat{z}_{k+1}$,

which depends only on pairwise sums of the elements of $\hat{x}_k$, enters in the computation. So, if we let

$$\Lambda_{k+1} = diag(M_{k+1}, D_{k+1}) \tag{2.16}$$

where M_{k+1} and D_{k+1} each have $2^k \times 2^k$ blocks, we see that

$$\hat{z}_k = \tilde{\Lambda}_k D_{k+1}^{-1} \hat{z}_{k+1}^2 \tag{2.17}$$

The tree structure of our problem also provides us with an iterative algorithm that has the structure of multigrid methods for solving boundary-value problems(Briggs,1987). In particular thanks to Markovianity we have the following:

$$\begin{aligned} E[x(t)|Y] &= E\left\{E[x(t)|x(\gamma^{-1}t), x(\alpha t), x(\beta t), Y]|Y\right\} \\ &= E\left\{E[x(t)|x(\gamma^{-1}t), x(\alpha t), x(\beta t), y(t)]|Y\right\} \end{aligned} \tag{2.18}$$

Assuming that $A(m)$ is invertible for all m we can directly apply the results of Verghese(1979):

$$x(\gamma^{-1}t) = F(m(t))x(t) - A^{-1}(m(t))B(m(t))\tilde{w}(t) \tag{2.19}$$

with

$$F(m(t)) = A^{-1}(m(t))[I - B(m(t))B^T(m(t))P_x^{-1}(m(t))] \tag{2.20}$$

and where $\tilde{w}(t)$ is a white noise process with covariance

$$\begin{aligned} E[\tilde{w}(t)\tilde{w}^T(t)] &= I - B(m(t))P_x^{-1}(m(t))B^T(m(t)) \\ &\triangleq \tilde{Q}(m(t)) \end{aligned} \tag{2.21}$$

Using these computations and our model (2.3) we can show(Chou,1990) that

$$\begin{aligned} \hat{x}(t) &= \mathcal{P}^{-1}(m(t))\left\{K_1(m(t))y(t) + K_2(m(t))\hat{x}(\gamma^{-1}t)\right. \\ &+ \left. K_3(m(t))[\hat{x}(\alpha t) + \hat{x}(\beta t)]\right\} \end{aligned} \tag{2.22}$$

where

$$K_1(m) = C^T(m)R^{-1}(m) \tag{2.23}$$

$$K_2(m) = F^T(m)R_1^{-1}(m) \tag{2.24}$$

$$\begin{aligned} R_1(m) &= A^{-1}(m)B(m)\tilde{Q}(m)B^T(m)A^{-T}(m) \\ K_3(m) &= A^T(m+1)R_2^{-1}(m+1) \\ R_2(m+1) &= B(m+1)B^T(m+1) \end{aligned} \tag{2.25}$$

$$\begin{aligned} \mathcal{P}(m) &= P_x^{-1}(m) + K_1(m)C(m) + K_2(m)F(m) \\ &+ 2K_3(m)A(m+1) \end{aligned} \tag{2.26}$$

Eq.(2.22) specifies an implicit set of equations for $\{\hat{x}(t)|t \in T\}$. Note that the computation involved at each point on the tree involves only its three nearesest neighbors and the measurement at that point. This suggests the use of a Gauss-Seidel relaxation algorithm for solving this set of equations. Note that the computations of all the points along a particular scale are independent of each other, allowing these computations to be performed in parallel. One can then imagine a variety of ways to organize the Gauss-Seidel computations. For example, we can initialize all $\hat{x}$ with zero, solve (2.22) first for the finest scale and then for sequentially coarser scales, followed by sequential computation back down to the finest scale. In multigrid terminology(Briggs,1987) this is a **V-cycle**, which can be iterated until convergence is obtained or augmented to form so-called **W-cycles**.

A third algorithm is a generalization of the well-known Rauch-Tung-Striebel(RTS) smoothing algorithm for causal state models. The algorithm once again involves a pyramidal set of steps and a considerable level of parallelism, with an initial fine-to-coarse sweep followed by coarse-to-fine. To begin let us define some notation:

$$Y_t = \{y(s)|s \preceq t\} \tag{2.27}$$

$$Y_t^+ = \{y(s)|s \prec t\} \tag{2.28}$$

and let $\hat{x}(\cdot|t)$ and $\hat{x}(\cdot|t+)$ denote the best estimates of $x(\cdot)$ based on Y_t and Y_t^+, respectively. Suppose that we have computed $\hat{x}(t|t+)$ and the corresponding error covariance, $P(m(t)|m(t)+)$; then, standard estimation results yield

$$\hat{x}(t|t) = \hat{x}(t|t+) + K(m(t))[y(t) - C(m(t))\hat{x}(m(t))] \tag{2.29}$$

$$K(m(t)) = P(m(t)|m(t)+)C^T(m(t))V^{-1}(m(t)) \tag{2.30}$$

$$V(m(t)) = C(m(t))P(m(t)|m(t)+)C^T(m(t)) + R(m(t)) \tag{2.31}$$

and the resulting error covariance is given by

$$P(m(t)|m(t)) = [I - K(m(t))C(m(t))]P(m(t)|m(t)+) \tag{2.32}$$

Suppose now that we have computed $\hat{x}(\alpha t|\alpha t)$ and $\hat{x}(\beta t|\beta t)$ and their common error covariance $P(m(t)+1|m(t)+1)$. Note that $Y_{\alpha t}$ and $Y_{\beta t}$ are disjoint and these estimates can be calculated in parallel. We then compute $\hat{x}(t|\alpha t)$ and $\hat{x}(t|\beta t)$ via identical formulas. For example

$$\hat{x}(t|\alpha t) = F(m(t)+1)\hat{x}(\alpha t|\alpha t) \tag{2.33}$$

with corresponding error covariance

$$\begin{aligned} P(m(t)|m(t)+1) &= F(m(t)+1)P(m(t)+1|m(t)+1)F^T(m(t)+1) \\ &+ \mathcal{Q}(m(t)+1) \end{aligned} \tag{2.34}$$

$$\begin{aligned} \mathcal{Q}(m(t)+1) &= \Psi\tilde{Q}(m(t)+1)\Psi^T \\ \Psi &= A^{-1}(m(t)+1)B(m(t)+1) \end{aligned} \tag{2.35}$$

These two steps of the processing are identical to the usual Kalman filter. The difference arises because we must now **merge** the two estimates $\hat{x}(t|\alpha t)$ and $\hat{x}(t|\beta t)$ to obtain $\hat{x}(t|t+)$ and its covariance. As shown in Chou(1990) we have the following:

$$\hat{x}(t|t+) = P(m(t)|m(t)+)P^{-1}(m(t)|m(t)+1)[\hat{x}(t|\alpha t) + \hat{x}(t|\beta t)] \tag{2.36}$$

$$P(m(t)|m(t)+) = [2P^{-1}(m(t)|m(t)+1) - P_x^{-1}(t)]^{-1} \tag{2.37}$$

Once we have reached the coarsest scale(consisting of a single node), we have the best smoothed estimate $\hat{x}_s$ at that node, and we can begin our coarse-to-fine propagation. As shown in Chou(1990), the coarse-to-fine recursion has the following form analogous to that in the RTS algorithm:

$$\begin{aligned} \hat{x}_s(t) &= \hat{x}(t|t) \\ &+ P(m(t)|m(t))F(m(t))P^{-1}(m(t)-1|m(t))\left[\hat{x}_s(\gamma^{-1}t) - \hat{x}(\gamma^{-1}t|t)\right] \end{aligned} \tag{2.38}$$

Note that $\hat{x}(t|t)$ and $\hat{x}(\gamma^{-1}t|t)$ were calculated during the fine-to-coarse sweep.

3 MODELING OF ISOTROPIC PROCESSES ON TREES

A zero-mean process Y_t, $t \in \mathcal{T}$ is **isotropic** if

$$E[Y_t Y_s] = r_{d(t,s)} \tag{3.1}$$

i.e. if its second-order statistics are invariant under any isometry of $\mathcal{T}$. These processes have been the subject of some study. However, many questions remain including an explicit criterion for a sequence r_n to be the covariance of such a process and the representations of isotropic processes as outputs of systems driven by white noise. Note first that the sequence $\{Y_{\gamma^{-n}t}\}$ is an ordinary time series so that r_n must be positive semidefinite; however, the constraints of isotropy require even more. To uncover this structure we seek here the characterization of the class of autoregressive (AR) models and to do this we need a bit more notation to define dynamics on trees. In particular, it is possible to code all points on the tree via shifts from an arbitrary origin node, i.e. as wt_0, where w is a word consisting of appropriate concatenations of our basic shift operators(Basseville,1989). The **length** of a word w is denoted $|w|$ and equals $d(wt,t)$ (e.g. $|\gamma^{-1}| = 1$, $|\delta| = 2$). Also, since we will be interested in coarse-to-fine dynamic models, we define some notation for **causal** moves:

$$w \preceq 0 \ (w \prec 0) \ \text{ if } wt \preceq t \ (wt \prec t) \tag{3.2}$$

The AR model of order p for processes on trees has the form

$$Y_t = \sum_{\substack{w \prec 0 \\ |w| \le p}} a_w Y_{wt} + \sigma W_t \tag{3.3}$$

where W_t is a white noise with unit variance. Note that this model is "causal" - i.e. it has a coarse-to-fine direction of propagation - since $w \preceq 0$.

The geometry of the tree and the constraints of isotropy make the parametrization of AR models as in (3.3) unsatisfactory. As we now describe a better representation is provided by the generalization of lattice structures which involves only one new parameter as p increases

by one. Let $\mathcal{H}\{\cdots\}$ denote the Gaussian linear space spanned by the variables in braces and define the (nth order) past of the node t:

$$\mathcal{Y}_{t,n} \triangleq \mathcal{H}\{Y_{wt} : w \preceq 0, |w| \leq n\} \tag{3.4}$$

As for time series, the development of models of increasing order involves recursions for the forward and backward prediction errors. Specifically, define the **backward residual space**:

$$\mathcal{Y}_{t,n} = \mathcal{Y}_{t,n-1} \oplus \mathcal{F}_{t,n} \tag{3.5}$$

where $\mathcal{F}_{t,n}$ is spanned by the backward prediction errors

$$F_{t,n}(w) \triangleq Y_{wt} - E\left(Y_{wt}|\mathcal{Y}_{t,n-1}\right) \tag{3.6}$$

where $w \preceq 0$, $|w| = n$. These variables are collected into a $2^{[\frac{n}{2}]}$-dimensional vector (see Basseville(1989) for the order), $F_{t,n}$. For $|w| < n$ and $w \succ 0$(i.e. $m(wt) = m(t)$) define the **forward prediction errors**:

$$E_{t,n}(w) \triangleq Y_{wt} - E\left(Y_{wt}|\mathcal{Y}_{\gamma^{-1}t,n-1}\right) \tag{3.7}$$

and let $\mathcal{E}_{t,n}$ denote the span of these residuals and $E_{t,n}$ the $2^{[\frac{n-1}{2}]}$-dimensional vector of these variables (see Basseville(1989)).

A key result(Basseville,1989) is that we can develop Levinson-like recursions for the local averages or **barycenters** of the residuals:

$$e_{t,n} = 2^{-[\frac{n-1}{2}]} \sum_{|w|<n, w\succ 0} E_{t,n}(w) \tag{3.8}$$

$$f_{t,n} = 2^{-[\frac{n}{2}]} \sum_{|w|=n, w\preceq 0} F_{t,n}(w) \tag{3.9}$$

Note first that $e_{t,0} = E_{t,0} = Y_t = F_{t,0} = f_{t,0}$; that $e_{t,1} = E_{t,1}$, $f_{t,1} = F_{t,1}$; and that a straightforward calculation yields

$$f_{t,1} = f_{\gamma^{-1}t,0} - k_1 e_{t,0} \tag{3.10}$$

$$e_{t,1} = e_{t,0} - k_1 f_{\gamma^{-1}t,0} \tag{3.11}$$

$$-1 \leq k_1 = \frac{r_1}{r_0} \leq 1 \tag{3.12}$$

For n even we find that

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

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

where the reflection coefficients k_n and the variances of the residuals satisfy

$$k_n = \operatorname{cor}\left(e_{t,n-1}, f_{\gamma^{-1}t,n-1}\right) = \operatorname{cor}\left(e_{\delta(\frac{n}{2})t,n-1}, e_{t,n-1}\right)$$
$$= \operatorname{cor}\left(e_{\delta(\frac{n}{2})t,n-1}, f_{\gamma^{-1}t,n-1}\right) \tag{3.15}$$

$$cor(x,y) = E(xy)/\left[E(x^2)E(y^2)\right]^{1/2} \tag{3.16}$$

$$\sigma^2_{e,n} = E\left(e^2_{t,n}\right) = \left(1-k_n^2\right)\sigma^2_{n-1} \tag{3.17}$$

$$\sigma^2_{f,n} = E\left(f^2_{t,n}\right) = \left(\frac{1+k_n}{2} - k_n^2\right)\sigma^2_{n-1} \tag{3.18}$$

$$-\frac{1}{2} < k_n < 1 \tag{3.19}$$

For n odd we have

$$e_{t,n} = \frac{1}{2}\left(e_{t,n-1} + e_{\delta(\frac{n-1}{2})t,n-1}\right) - k_n f_{\gamma^{-1}t,n-1} \tag{3.20}$$

$$f_{t,n} = f_{\gamma^{-1}t,n-1} - \frac{1}{2}k_n\left(e_{t,n-1} + e_{\delta(\frac{n-1}{2})t,n-1}\right) \tag{3.21}$$

$$k_n = \operatorname{cor}\left(\frac{1}{2}\left(e_{t,n-1} + e_{\delta(\frac{n-1}{2})t,n-1}\right), f_{\gamma^{-1}t,n-1}\right) \tag{3.22}$$

$$\sigma^2_{e,n} = \sigma^2_{f,n} = \sigma^2_n = \left(1-k_n^2\right)\sigma^2_{f,n-1} \tag{3.23}$$

$$-1 < k_n < 1 \tag{3.24}$$

Note that the constraints on the reflection coefficients are slightly different than for time series, and these conditions are precisely those for a sequence r_n to be the covariance of an isotropic process. In addition, there exists a generalization of the Schur recursions that allows us to calculate the k_n efficiently. Further results in Basseville(1989) include a complete characterization of AR models, a stability result analogous to the time series result that k_n must not achieve its extreme values, and whitening and modeling filter structures for $Y_t = E_{t,0}$. We limit ourselves here to stating the last of these results. Let 1_* denote a unit vector all of whose components are the same, and let $\cup_* = 1_* 1_*^T$. The modeling filter for Y_t is given by the following. For n even

$$\begin{pmatrix} E_{t,n-1} \\ F_{t,n} \end{pmatrix} = \Sigma(k_n) \begin{pmatrix} E_{t,n} \\ E_{\delta(\frac{n}{2})t,n} \\ F_{\gamma^{-1}t,n-1} \end{pmatrix} \tag{3.25}$$

$$\Sigma(k_n) \triangleq \begin{bmatrix} I & 0 & k_n\cup_* \\ -k_n\cup_* & I & (k_n - k_n^2)\cup_* \\ -k_n\cup_* & 0 & (I - k_n^2\cup_*) \end{bmatrix} \tag{3.26}$$

For n odd, $n > 1$:

$$\begin{pmatrix} E_{t,n-1} \\ E_{\delta(\frac{n-1}{2})t,n-1} \\ F_{t,n} \end{pmatrix} = \Sigma(k_n) \begin{pmatrix} E_{t,n} \\ F_{\gamma^{-1}t,n-1} \end{pmatrix} \tag{3.27}$$

$$\Sigma(k_n) \triangleq \begin{bmatrix} I & k_n\cup_* \\ -k_n\cup_* & (I - k_n^2\cup_*) \end{bmatrix} \tag{3.28}$$

while for $n = 1$:

$$\begin{pmatrix} E_{t,0} \\ F_{t,1} \end{pmatrix} = \begin{pmatrix} 1 & k_1 \\ -k_1 & 1-k_1^2 \end{pmatrix} \begin{pmatrix} E_{t,1} \\ F_{\gamma^{-1}t,0} \end{pmatrix} \tag{3.29}$$

4 DISCUSSION OF FURTHER WORK

The results described so far in this paper represent our initial efforts in developing a system and estimation theory for multi-scale processes. There are a number of additional results presently under development. First, we are in the process of developing a complete system theory for the models described in Section 2. In particular in order to understand the estimation problem better and to develop a theory for our new 3-step Riccati equation on trees(update, predict, and merge) we have developed the dual system to (2.3) and the associated complementary processes and Hamiltonian form of the estimator. In addition reachability and observability results are being developed in order to develop an asymptotic theory for the Riccati equation. Secondly, we are exploring the development of a theory as in Section 3 but for a weaker notion of stationarity when $E(Y_t Y_s)$ depends only on $d(t, s\wedge t)$ and $d(s, s\wedge t)$. Note that from (2.7) we see that the model (2.3) with constant parameters is of this form so that we expect very strong ties between our work in Sections 2 and 3. In particular the transform theory we are developing for our models leads us to a notion of rationality or finite-dimensionality for systems on trees and we expect the resulting tests on Hankel matrices to be closely tied to our system theory for the models of Section 2. Finally, we are also exploring the generalization of the results described here to more general weighted lattices, with weights and structure determined by quadrature mirror filters other than the one that generates the Haar wavelet basis.

References

[1] Basseville, M., Benveniste, A., and Willsky, A.S. [1989]. "Multi-scale Autoregressive Processes," Lab. for Information and Decision Systems, MIT, June.

[2] Briggs, W. (1987). *A Multigrid Tutorial*, SIAM, Philadelphia, PA.

[3] Chou, K. (1990). *A Stochastic Modeling Approach to Multiscale Signal Processing*, MIT, Dept. Electrical Engineering, PhD Thesis(in preparation).

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

[5] Mallat, S.G. (1987). "Multiresolution approximation and wavelets," Dept. of Computer and Info. Science - U. of Penn., MS-CIS-87-87, GRASP LAB 80, Sept.

[6] Verghese, G. and Kailath, T. (1979). "A further note on backward Markovian models," IEEE Trans. on Information Theory, IT-25, pp.121-124.