Multidimensional Systems and Signal Processing, 12, 109–142, 2001

Computationally Efficient Stochastic Realization for Internal Multiscale Autoregressive Models*

AUSTIN B. FRAKT http://ssg.mit.edu/
Abt Associates Inc., Cambridge, MA

ALAN S. WILLSKY
Laboratory for Information and Decision Systems,
Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology,
Cambridge, MA

Received April 29, 1999; Revised September 25, 2000

Abstract. In this paper we develop a stochastic realization theory for multiscale autoregressive (MAR) processes that leads to computationally efficient realization algorithms. The utility of MAR processes has been limited by the fact that the previously known general purpose realization algorithm, based on canonical correlations, leads to model inconsistencies and has complexity quartic in problem size. Our realization theory and algorithms addresses these issues by focusing on the estimation-theoretic concept of predictive efficiency and by exploiting the scale-recursive structure of so-called *internal* MAR processes. Our realization algorithm has complexity quadratic in problem size and with an approximation we also obtain an algorithm that has complexity linear in problem size.

Keywords: multiscale autoregressive models, stochastic realization, graphical models, predictive efficiency, state-space, Markov models

1. Introduction

The power of the multiscale autoregressive (MAR) framework [5, 6] resides in its ability to compactly model a rich class of phenomena [10, 26, 37] and to efficiently address complications that arise in many one- and multi-dimensional signal processing problems (e.g., spatially irregular data, non-stationarities, and others). Fast and flexible signal processing algorithms have been developed for MAR models [5, 38] and the utility of the framework has already been established in a wide variety of applications [9, 11, 14–18, 21, 25, 28–31, 33, 36, 41, 42]. To harness this utility, of course, requires that the phenomena of interest be effectively modeled in the MAR framework. However, prior attempts [10, 22, 24, 26] to develop systematic MAR model-building methodologies have suffered from theoretical inconsistencies and computational intractability. In this paper we develop a conceptually complete realization theory for a class of MAR processes that leads to computationally efficient model realization algorithms.

A MAR process is a collection of random vectors $\{x(s)\}$, called *states*, indexed by the nodes of a tree which are organized into scales (see Figure 1). MAR states are coupled with affine coarse-scale to fine-scale dynamics that generalize those of a state-space process (see (1)). It is perhaps already evident to the reader that a MAR model is a

* This work has been sponsored by Office of Naval Research grant N00014-91-J-1004 and Air Force Office of Scientific Research grant F49620-98-1-0349.

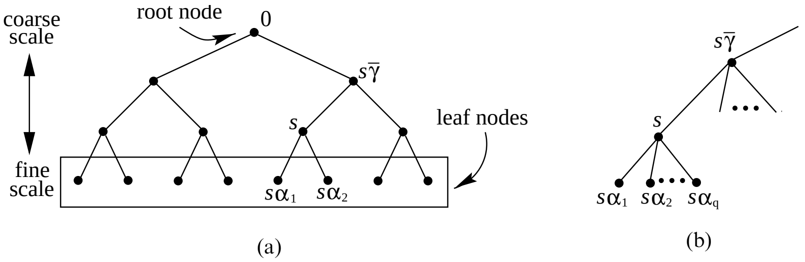


Figure 1. (a) A dyadic tree. The root node is indexed by $s = 0$. The parent of node s is denoted $s\bar{\gamma}$. The children of node s are labeled from left to right by $s\alpha_1$, $s\alpha_2$. (b) For a q-adic tree the children of node s are labeled from left to right by $s\alpha_1, s\alpha_2, \ldots, s\alpha_q$.

particular type of graphical model. As such, MAR models, and their associated algorithms, possesses properties that are similar to more general graphical models. We will highlight such similarities as they arise in the body of this paper and we also include a discussion of this point in the conclusion.

The theory and algorithms we develop in this paper address the problem of choosing the parameters for the dynamics of a MAR process to model the second-order statistics of *any* given fine-scale one- or multi-dimensional random process. We call this the MAR stochastic realization problem and our approach to it relies on the concept of *internality*.

Internality is a familiar and important concept in state-space modeling:[1] an internal state-space process is one for which each state is a linear function of the observed process [35]. The corresponding definition of an internal MAR process is one for which each state $x(s)$ is a linear function of the states indexed by leaf nodes that descend from s.

Internal MAR processes and models are important for a variety of reasons [12]; they admit exact inclusion of non-local linear functions at coarser-scale nodes and thereby permit the statistically optimal fusion of multiresolution measurements, a point we will illustrate in our examples. Internality also plays a role in overcoming the computational burdens of previous model building approaches. The precise development of internality is a primary contribution of this paper and it represents a significant advancement in the theory underlying the MAR framework.

Previous work [10, 22, 24, 26] also attempted to focus on internal realizations but is inconsistent in the following sense. While the computation of model parameters is based on assuming that the model under construction is internal, the model resulting from these computations is not guaranteed to be internal (and usually is not). Specific examples of this phenomenon are found in [12, 19, 22]. This inconsistency has also been noted but not resolved in [26].

The methods for constructing MAR models developed in [10, 22, 24, 26] scale poorly with problem size, making them prohibitive for many problems of interest. The computational burden of these methods stems from two sources. First, they are not scale-recursive and, therefore, do not take advantage of the natural efficiency of tree data structures. Second, they are based on canonical correlations, a burdensome approach involving the inversion and singular value decomposition of large matrices.

Consequently, the approach developed in [22, 24, 26] is quartic in problem size while that of [10] is cubic in problem size.[2]

In contrast, our approach has a quadratic complexity because it *is* scale-recursive and is *not* based on canonical correlations. With respect to the former, the theoretical basis for our scale-recursive realization algorithm stems from a thorough analysis of (wide-sense) Markovianity for internal tree-indexed processes. In addition to internality, Markovianity is another important concept in the state-space setting that generalizes to MAR processes and that plays a central role in the stochastic realization problem. Moreover, and most importantly for our purpose, we show that, for internal processes, this Markov property has an *equivalent* scale-recursive definition that is vastly simpler to work with and leads to efficient model realization. Because of the structure it reveals and efficiency to which it leads, the development of this scale-recursive Markov property is one of the important contributions of this paper.

The efficiency of our realization approach stems also from the fact that it is based on the estimation-theoretic concept of *predictive efficiency* [3, 40] rather than on canonical correlations. In brief, predictive efficiency is the idea of finding and prioritizing the best (in a minimum mean-square error sense) linear functionals of one random vector for the purposes of linearly estimating another.

An important feature of predictive efficiency is the asymmetric way in which it treats data and variables to be estimated. A consequence of this for our approach to the stochastic realization problem is that state variables are chosen to provide maximal *total* reduction in estimation error variance. This is in contrast to the canonical correlations approach which provides maximal *fractional* error variance reduction and is therefore equally concerned with low- and high-variance features. Another important advantage of predictive efficiency's asymmetry is that it avoids the costly inversion and singular value decomposition of large matrices, steps which cannot be avoided in the canonical correlations approach.

While our scale-recursive, predictive efficiency-based realization algorithm is relatively efficient, it is still too computationally burdensome for many practical problems, particularly those arising in image (or higher-dimensional) processing. A major objective of our work has been to develop realization approaches that scale well with problem size. In this paper we provide an approximation with complexity that scales linearly with problem size and illustrate in several examples the degree of approximation error relative to the exact method. All of our theory and algorithms are applicable to signal processing problems of any dimension. However, for clarity of presentation and visualization, we often provide a detailed development in one-dimension only. Generalizations to higher dimensions are discussed and image processing examples are also provided.

This paper is organized as follows. In Section 2 we review MAR processes. A problem statement and overview of our theoretical development is found in Section 3. Internality and several notions of Markovianity for tree-indexed processes are the subjects of Section 4 and Section 5, respectively. Predictive efficiency is reviewed in Section 6. Our realization theory is then applied in Section 7 which develops an algorithm with complexity quadratic in problem size and, with an approximation, an algorithm with complexity linear in problem size. Examples are provided in Section 8 and concluding remarks are found in Section 9.

2. MAR Processes Background

In this section we provide our notational conventions and a brief review of MAR processes. A MAR process is a generalization of a discrete-time state-space process. Both are graphical models with affine dynamics. However, a MAR process may be indexed by the nodes of *any* tree and it reduces to a state-space process in time when the tree is monadic. Precisely, a zero-mean[3] MAR process $x(\cdot)$ has dynamics

$$x(s) = A(s)x(s\bar{\gamma}) + w(s) \tag{1}$$

where $s\bar{\gamma}$ is the parent of node s (see Figure 1) and $w(s)$ is white, uncorrelated with the root-node state $x(0)$ and has auto-covariance $Q(s)$.

For our purposes, it suffices to consider MAR processes indexed by nodes of q-adic trees. Our notation for referring to nodes of a q-adic tree is indicated in Figure 1. The root node is labeled 0 and the children of a node s are, from left to right, $s\alpha_1, s\alpha_2, \ldots, s\alpha_q$. There is a natural notion of scale associated with q-adic trees. The root node represents the coarsest scale (scale zero) while the leaf nodes constitute the finest scale (scale M). More generally, the nodes $\{s \mid s\bar{\gamma}^n = 0\}$ reside at scale n. We denote the scale of node s by $m(s)$. For dyadic trees, $q = 2$ and scale n indexes a one-dimensional vector-valued signal of length 2^n. For quad-trees, $q = 4$ and scale n indexes a two-dimensional vector-valued field of size $2^n \times 2^n$. Extensions to higher dimensions ($q > 4$) are straightforward.

Our goal is to build MAR models for fine-scale random signals which we view as indexed by the leaf nodes of q-adic trees. In developing the theory and describing our algorithms, we frequently refer to other scales and other subsets of nodes. So, for simplicity of our subsequent presentation, we make the following definitions for frequently referred to subsets of the set of nodes of a q-adic tree:

$$\begin{aligned}
\mathcal{S}_s &\triangleq \{t \mid t = s \text{ or } t \text{ is a descendent of } s\} = \text{nodes in subtree rooted at } s,\\
\mathcal{S}_s^c &\triangleq \mathcal{S}_0 - \mathcal{S}_s = \text{nodes other than those in subtree rooted at } s,\\
\mathcal{T}_s(n) &\triangleq \{t \in \mathcal{S}_s \mid m(t) = n\} = \text{nodes at scale } n \text{ descending from } s,\\
\mathcal{T}_s^c(n) &\triangleq \mathcal{T}_0(n) - \mathcal{T}_s(n) = \text{nodes at scale } n \text{ not descending from } s.
\end{aligned}$$

Again, to simplify our development, we make the following definitions for frequently referred to sub-processes of a tree-indexed process $\{x(s)\}_{s \in \mathcal{S}_0}$:

$$\begin{aligned}
x_s^n &\triangleq \{x(t)\}_t \in \mathcal{T}_s(n) = \text{process at scale } n \text{ that descends from node } s,\\
x_{s^c}^n &\triangleq \{x(t)\}_t \in \mathcal{T}_s^c(n) = \text{process at scale } n \text{ that does not descend from node } s.
\end{aligned}$$

We often interpret these sub-processes as vectors.[4] Also, when referring to a sub-process at an entire scale we often drop the 0 subscript. For instance $x^M \equiv x_0^M$ is the finest-scale sub-process. Some of this notation is summarized in Figure 2.

A MAR process can be viewed as an implicit representation of a covariance matrix. The elements of P_{x^M}, the covariance matrix[5] for the leaf-node states of a MAR process $x(\cdot)$, are completely characterized by the MAR parameters $A(\cdot)$, $Q(\cdot)$, and the root-node covariance $P_{x(0)}$. As a specific example, the block-diagonal elements of P_{x^M} are obtained by recursively

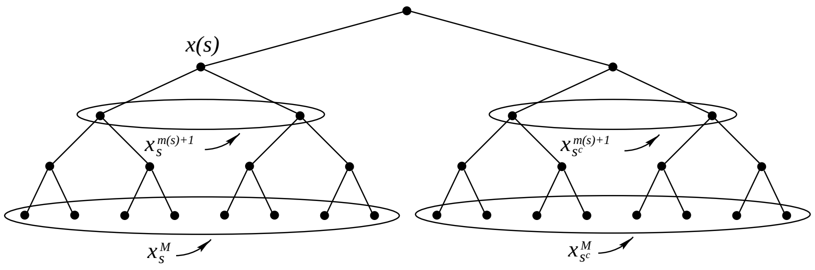


Figure 2. Notation for sub-processes of a tree-indexed process $x(\cdot)$.

solving the Lyapunov equation resulting from (1) using $P_{x(0)}$ as the initialization:

$$P_{x(s)} = A(s)P_{x(s\bar{\gamma})}A(s)^T + Q(s)\,. \tag{2}$$

We emphasize that, in general, $A(s)$ and $Q(s)$ vary as a function of node $s \in \mathcal{S}_0$. As mentioned in the Introduction, the concept of Markovianity is important for tree-indexed processes:

Definition 1 (Markov Property). A tree-indexed process $x(\cdot)$ has the *Markov property* if for all $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$, conditioned on $x(s)$, the sub-processes indexed by the sets of nodes in the sub-trees separated by s, namely, $\{x(t)\}_{t\in\mathcal{S}_{s\alpha_1}}$, $\{x(t)\}_{t\in\mathcal{S}_{s\alpha_2}}, \ldots, \{x(t)\}_{t\in\mathcal{S}_{s\alpha_q}}$ and $\{x(t)\}_{t\in\mathcal{S}_s^c}$, are conditionally uncorrelated.

That MAR processes have the Markov property is easily shown [8, Appendix A]. If a MAR process is Gaussian[6] then it has the (equivalent) properties of "pairwise Markovianity," "local Markovianity," and "global Markovianity" from the graphical modeling literature [34]. The Markov property of Definition 1 is a wide-sense equivalent to these notions of Markovianity.

The Markov property leads to fast statistical signal and image processing algorithms. Sample-path generation (with complexity quadratic in state dimension and linear in the number of finest-scale nodes) is accomplished using (1). Also, a linear least-squares estimator [5, 6] and likelihood calculator [38] have been developed based on a measurement model analogous to the classical state-space one:

$$y(s) = C(s)x(s) + v(s) \tag{3}$$

where $v(s)$ is white and uncorrelated with $x(\cdot)$ and $w(\cdot)$.The estimator and likelihood calculator have computational complexity cubic in state dimension and linear in the number of finest-scale nodes. This linear scaling with the number of leaf nodes is of great importance in two- and higher-dimensional problems, contexts in which the much higher complexity of other algorithms (e.g., those for Markov random field models) is often prohibitive. The cubic scaling in state dimension provides strong motivation for keeping state dimensions small. The MAR statistical inference algorithms just described are special cases of more general algorithms (like the junction-tree algorithm) that are familiar in the graphical modeling community [7].

3. Problem Statement and Overview

The ultimate objective of this paper and the motivation for the theoretical development of subsequent sections is the MAR stochastic realization problem, which we now describe in detail. Suppose we are given the covariance matrix P_{f^M} for the length N random vector f^M which may represent a one-dimensional signal or a multi-dimensional field, lexicographically ordered. For convenience only, suppose that f^M has length $N = dq^M$, for some positive integer d. Thus, we can view the elements of f^M as the fine-scale subprocess of a random process indexed by the nodes of an $(M+1)$-scale q-adic tree where each leaf-node indexes d consecutive elements of f^M. Our goal is to build an *internal* MAR process $x(\cdot)$ (where internality is to be defined) so that the finest-scale sub-process x^M is an exact or approximate model for the random signal f^M, i.e., so that $P_{x^M} \approx P_{f^M}$.

Building such a MAR model requires designing the states at all of the "hidden" coarse-scale nodes. Additionally, because an exact model (i.e., one for which $P_{x^M} \equiv P_{f^M}$) typically requires impractically large state dimension,[7] we seek a criterion for choosing the "best" state variables if a reduced-order, approximate model is desired. Our criterion, predictive efficiency, is developed in Section 6.

Before we develop this criterion and apply it to the realization problem, we need to take a step back to consider tree-indexed processes more generally. There are two basic concepts, the first of which is internality which we develop in Section 4. Internality has both intellectual and practical importance. First, as described in the Introduction, it is a natural extension of the well-studied time-series concept. Second, internal MAR models have coarse-scale states that include non-local linear functions of fine-scale states. This allows for efficient fusion of non-local and local measurements with no increase in computational complexity as compared to the case off using only fine-scale data [9]. Lastly, while non-internal MAR processes can be constructed, their states have exogenous random components, a property that is not suitable in many problems such as the fusion of multiresolution data [9]. For an internal model, all the statistical properties can be derived from the signal being modeled—there is no exogenous randomness.

The second basic concept associated with tree-indexed processes is Markovianity. As discussed in the Introduction, we develop a scale-recursive formulation of this concept for internal processes (Section 5) that leads to efficient model realization. Once we develop these two basic concepts for tree-indexed processes, we apply them to the MAR stochastic realization problem. In doing so, we deduce the structure of internal MAR models that both must be satisfied (and which is *not* satisfied by previous methods) *and* which reduces computational complexity.

4. Internal Processes

In this section, we first define internality for an arbitrary tree-indexed process and then seek to understand what structure must be imposed on the states of a MAR process to make it internal.

Definition 2 (Internal Tree-Indexed Process). A tree-indexed process $x(\cdot)$ is *internal* if for all $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$, $x(s)$ is a linear function of x_s^M, the process indexed by finest-scale nodes that descend from s. I.e., for some set of matrices $\{W_s\}$,

$$x(s) = W_s x_s^M. \tag{4}$$

The matrices $\{W_s\}$ are called *internal matrices.* If $x(\cdot)$ is also a MAR process then $x(s)$ defined by (4) is called an *internal state*. For a general tree-indexed process, internality places no restrictions on the internal matrices. However, we are interested in internal MAR processes which also obey (1). A consequence of this is that the internal matrices cannot be chosen independently. This can be seen intuitively because fine-scale states are derived from coarse scale ones by (1) and must be consistent with the information contained in coarse scale states (given by (4)). That is, (1) and (4) together place constraints on finer-scale states and, thereby, on the internal matrices associated with those states.

The constraints imposed by (4) are *not* enforced in previous systematic realization approaches [10, 22, 24, 26]. As a consequence the MAR models developed in these works are *not* internal [22, Section 3.7.2]. Although the authors of [10, 22, 24, 26] were aware of this, it has, until now, not been dealt with in a theoretically consistent and complete framework. We not only develop a framework that incorporates the constraints of (4) but we show how doing so vastly *simplifies* the construction of internal MAR models. The key is that (4) is not the right parameterization for internal MAR states. The right parameterization comes from the following.

Definition 3 (Locally Internal Tree-Indexed Process). A tree-indexed process $x(\cdot)$ is *locally internal if for all $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$, $x(s)$ is a linear function of $x_s^{m(s)+1}$, the process indexed by the children nodes of s. I.e., for some set of matrices $\{V_s\}$,*

$$x(s) = V_s x_s^{m(s)+1}. \tag{5}$$

Notice that (4) places all of the focus on the fine scale while (5) is scale-recursive. We call V_s a *local internal matrix*. The following proposition shows that the local internal matrices provide the right parameterization for an internal process.

PROPOSITION 1 *A MAR process, $x(\cdot)$, is internal if and only if it is locally internal.*

Proof: The "if" direction is trivial. If $x(\cdot)$ is locally internal then we may write $x(s)$ as a linear combination of its children. In turn, each $x(s\alpha_i)$ is a linear combination of *its* children and so on, scale-recursively down the tree. Therefore, $x(s)$ is a linear combination of its finest-scale descendents x_s^M. This holds for all $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$ so $x(\cdot)$ is internal.

For the "only if" direction, assume that $x(\cdot)$ is internal. For any $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$, we may write[8]

$$x(s) = \hat{\mathrm{E}}\left[x(s) \mid x_s^{m(s)+1}\right] + \tilde{x}(s) \tag{6}$$

where $\tilde{x}(s)$ is uncorrelated with $x_s^{m(s)+1}$. Since $x(s)$ and the $x(s\alpha_i)$ which comprise $x_s^{m(s)+1}$ are assumed to be internal states, $\tilde{x}(s)$ must be a linear function of x_s^M. We show that $\tilde{x}(s)$ must be zero (so that $x(s)$ is indeed a linear function of $x_s^{m(s)+1}$). By (6), $\tilde{x}(s)$ is a linear combination of $x(s)$ and $x_s^{m(s)+1}$ so we may write

$$\tilde{x}(s) = \hat{\mathrm{E}}\left[\tilde{x}(s) \mid x(s), x_s^{m(s)+1}\right] = \hat{\mathrm{E}}\left[\tilde{x}(s) \mid x_s^{m(s)+1}\right] = 0 \tag{7}$$

where the second equality follows from the fact that $\tilde{x}(s)$ must be a linear function of x_s^M and the fact that conditioned on $x_s^{m(s)+1}$, $x(s)$ and x_s^M are conditionally uncorrelated (by the Markov property). The third equality follows from the fact that$\tilde{x}(s)$ is uncorrelated with the $x(s\alpha_i)$. This completes the proof. ■

Note that, given the local internal matrices $\{V_s\}$ it is easy to derive the internal matrices $\{W_s\}$ recursively as follows:

$$W_s = \begin{cases} I_d & \text{if } m(s) = M, \\ V_s \begin{bmatrix} W_{s\alpha_1} \\ W_{s\alpha_2} \\ \vdots \\ W_{s\alpha_q} \end{bmatrix} & \text{otherwise} \end{cases} \tag{8}$$

where I_d is the $d \times d$ identity matrix and $d \triangleq \dim(x(s))$ is the state dimension[9] (i.e., the length) of state $x(s)$. Since an internal MAR process has states satisfying (5) as well as (1), we immediately have the following complete characterization of the parameters $A(s)$ and $Q(s)$ for such a process in terms of the local internal matrices and the covariance matrix for $x_s^{m(s)+1}$.

Proposition 2 *Suppose $x(\cdot)$ is a MAR process. Let $J_{s\alpha_i}$ be the selection matrix such that $J_{s\alpha_i} x_s^{m(s)+1} = x(s\alpha_i)$. Then $x(\cdot)$ is a locally internal MAR process with $x(s) = V_s x_s^{m(s)+1}$ if and only if for all $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$,*

$$A(s\alpha_i) = J_{s\alpha_i} P_{x_s^{m(s)+1}} V_s^T \left(V_s P_{x_s^{m(s)+1}} V_s^T\right)^{-1}, \tag{9a}$$

$$Q(s\alpha_i) = J_{s\alpha_i} \left(P_{x_s^{m(s)+1}} - P_{x_s^{m(s)+1}} V_s^T \left(V_s P_{x_s^{m(s)+1}} V_s^T\right)^{-1} V_s P_{x_s^{m(s)+1}}\right) J_{s\alpha_i}^T. \tag{9b}$$

Proof: See Appendix A.1. ■

The relations of (9) may be written in another form which emphasizes that $A(s\alpha_i)$ and $Q(s\alpha_i)$ depend only on state covariances and parent-child cross covariances:

$$A(s\alpha_i) = P_{x(s\alpha_i)x(s)} P_{x(s)}^{-1}, \tag{10a}$$

$$Q(s\alpha_i) = P_{x(s\alpha_i)} - P_{x(s\alpha_i)x(s)} P_{x(s)}^{-1} P_{x(s\alpha_i)x(s)}^T. \tag{10b}$$

Together, Propositions 1 and 2 provide the necessary and sufficient conditions for a MAR process to be internal.

5. Notions of Markovianity

For internal tree-indexed processes, the Markov property of Definition 1 is equivalent to two other notions of Markovianity. These notions, which we develop in this section, are much simpler to work with and lead to a scale-recursive realization algorithm, presented in Section 7, with substantially reduced computational complexity as compared to previous methods. The first alternate notion of Markovianity is the *fine-scale Markov property*, which is the focus of the approaches in [10, 22, 24, 26].

Definition 4 (Fine-Scale Markov Property). A tree-indexed process $x(\cdot)$ has the *fine-scale Markov property* if conditioned on $x(s)$ for any $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$ the $q+1$ vectors in the set $\{x^M_{s\alpha_i}\}_{i=1}^q \cup \{x^M_{s^c}\}$ are conditionally uncorrelated.

Figure 3 provides some intuition about the relationship between the Markov property and the fine-scale Markov property. The Markov property focuses on the conditional decorrelation of the states indexed by the nodes in the subtrees extending from s (the three sets of nodes enclosed by solid lines and labeled "A","B", and "C" in Figure 3). The fine-scale Markov property, in contrast, places its attention on the conditional decorrelation of finest-scale sub-processes (dotted-lined regions in Figure 3). While there are fewer leaf nodes than nodes in the tree, this does not provide a substantial amount of simplification because the number of tree nodes and the number of leaf nodes are of the same order. This is the key to why previous realization methods [10, 22, 24, 26], which make extensive use of the fine-scale Markov property, scale poorly with problem size. We now show that, for internal processes, the fine-scale Markov property is equivalent to the Markov property.

PROPOSITION 3 *Assume that $x(\cdot)$ is an internal tree-indexed process. Then it has the fine-scale Markov property if and only if it has the Markov property.*

Proof: First, if $x(\cdot)$ has the Markov property it clearly has the fine-scale Markov property since the Markov property subsumes the fine-scale Markov property. Assume, then,

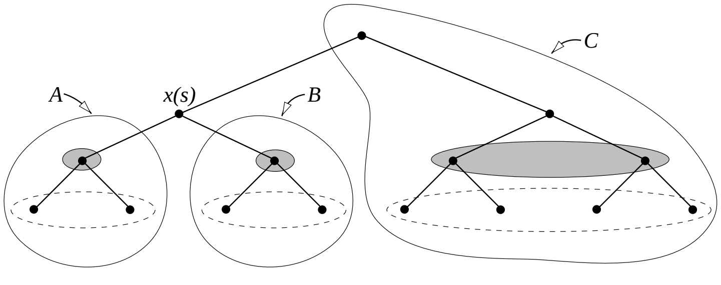


Figure 3. For an internal process $x(\cdot)$ indexed by a dyadic tree, the following are equivalent: $x(s)$ conditionally decorrelates the sub-processes indexed by the three sets of nodes (i) in the solid-lined regions labeled "A", "B", and "C" (Markov property), (ii) in the dashed-lined regions (fine-scale Markov property), (iii) in the shaded regions (scale-recursive Markov property).

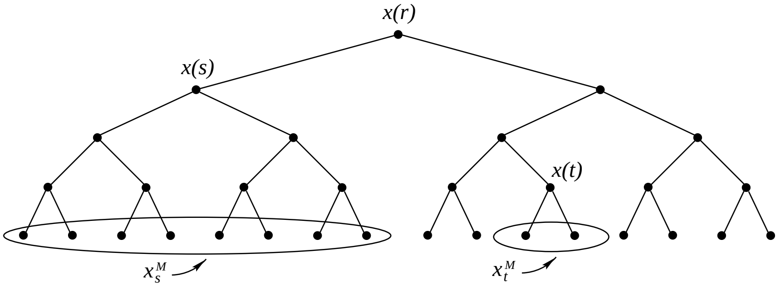


Figure 4. Proof of Proposition 3, case 1: t is not an ancestor of s and s is not an ancestor of t.

that $x(\cdot)$ has the fine-scale Markov property and let $s, r, t \in \mathcal{S}_0$ such that the unique shortest path from s to t goes through r. Subject only to this condition, s, r, t are arbitrary so we need to show that $x(s)$ and $x(t)$ are conditionally uncorrelated when conditioned on $x(r)$. There are two cases.

Case 1. First consider the case where t is not an ancestor of s nor is s an ancestor of t (see Figure 4). It is clear from Figure 4, that the index sets for the vectors x_t^M and x_s^M do not overlap. Therefore, by assumption, $x(r)$ conditionally decorrelates x_s^M and x_t^M. It follows that $x(r)$ conditionally decorrelates $x(s) = W_s x_s^M$ and $x(t) = W_t x_t^M$. The algebraic details are omitted and may be found in [19].

Case 2. Next consider the case where $s = t\bar{\gamma}^n$ for some n (see Figure 5). Therefore s is also an ancestor of r and consequently x_t^M is contained in x_r^M. It follows that $x(t)$ is a linear function of x_r^M so it suffices to show that $x(s)$ and x_r^M are conditionally decorrelated by $x(r)$. Because by assumption $x(\cdot)$ is an internal process, we may write $x(s)$ as

$$x(s) = W_s x_s^M = L_{sr} x_r^M + L x_{r^c}^M, \tag{11}$$

for some matrices L_{sr} and L where $L_{sr} = D_{sr} W_r$, for some matrix D_{sr}. Therefore,

$$\text{rowspace}(L_{sr}) \subseteq \text{rowspace}(W_r). \tag{12}$$

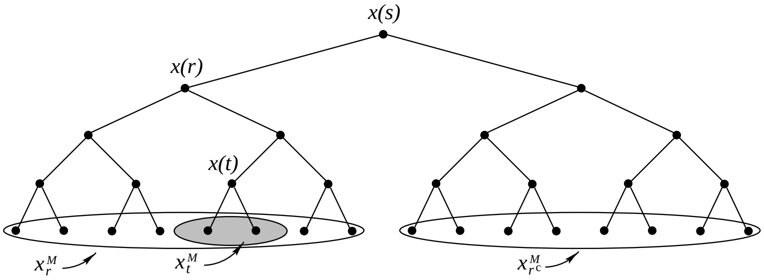


Figure 5. Proof of Proposition 3, case 2: $s = t\bar{\gamma}^n$. The shaded region indicates x_t^M.

Thus, $L_{sr}x_r^M$ can be linearly estimated from $x(r) = W_r x_r^M$ without error. Also, conditioned on $x(r)$, $x_{r^c}^M$ and x_r^M are conditionally uncorrelated by hypothesis. Therefore, $x(r)$ conditionally decorrelates $x(s)$ and x_r^M. ■

As we now develop, another notion of Markovianity which is equivalent to the Markov property is the *scale-recursive Markov property*. The development of this scale-recursive formulation of the Markov property is one of the major contributions of this paper. It is significant because it permits us to view the stochastic realization problem scale-recursively and thereby, develop efficient algorithms.

Definition 5 (Scale-Recursive Markov Property). A tree-indexed process $x(\cdot)$ has the *scale-recursive Markov property* if conditioned $x(s)$ for any $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$ the $q+1$ vectors in the set $\{x(s\alpha_i)\}_{i=1}^q \cup \{x_{s^c}^{m(s)+1}\}$ are conditionally uncorrelated.

Referring to Figure 3, we see that the scale-recursive Markov property is similar to the fine-scale Markov property except that rather than focusing on the leaf-node states, it focuses on those at the preceding finer scale (in the shaded regions). Since the sets of nodes associated with the scale-recursive Markov property are asymptotically of strictly smaller order than those associated with the Markov property (solid-lined regions labeled "A", "B", and "C") or the fine-scale Markov property (dotted-lined regions), our realization algorithm based on scale-recursive Markovianity is orders of magnitude more efficient than previous approaches. Specifically, at scale $M-1$, the total number of variables considered is the same for both the fine-scale Markov property and the scale-recursive Markov property. However, at coarser scales, the sets involved in the scale-recursive Markov property are smaller than those involved in the fine-scale Markov property. Indeed, at each successive coarser scale, the total number of variables considered in the scale-recursive Markov property is reduced by a factor of q. We now show that the scale-recursive Markov property is equivalent to the Markov property for internal processes.

PROPOSITION 4 *Assume $x(\cdot)$ is an internal tree-indexed process. Then $x(\cdot)$ has the scale-recursive Markov property if and only if it has the Markov property.*

Proof: First, if $x(\cdot)$ has the Markov property then, by definition, it has the scale-recursive Markov property. Next, assume that $x(\cdot)$ has the scale-recursive Markov property. We show in Appendix A.2 that for an arbitrary s in $\mathcal{S}_0 - \mathcal{T}_0(M)$, $x(s)$ conditionally decorrelates the vectors in the set $\{x_{s\alpha_i}^M\}_{i=1}^q \cup \{x_{s^c}^M\}$ (see Figure 6). Having shown this, then $x(\cdot)$ has the fine-scale Markov property and thus, by Proposition 3, it has the Markov property. ■

The computational complexity of previously developed realization algorithms stems from several sources, two of which are exposed here. First, because they are based on the fine-scale Markov property and use canonical correlations, they must consider the statistics (covariance matrix) for a large number of variables (those indexed by the fine-scale nodes). Second, they must do this at *every* node. In contrast,we focus on the (equivalent) scale-recursive Markov property which alleviates the latter source of complexity as

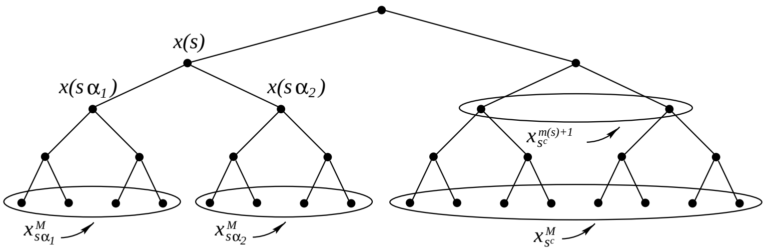


Figure 6. As shown in the proof of Proposition 4, if $x(s)$ conditionally decorrelates the vectors in the set $\{x(s\alpha_i)\}_{i=1}^q \cup \{x_{s^c}^{m(s)+1}\}$ then it conditionally decorrelates the vectors in the set $\{x_{s\alpha_i}^M\}_{i=1}^q \cup \{x_{s^c}^M\}$. This figure illustrates the case for $q = 2$.

described previously. We also use predictive efficiency and not canonical correlations which alleviates the former source of complexity as described in the next section.

6. Predictive Efficiency and Decorrelating Random Vectors

At the heart of our realization approach we must deduce the information necessary to conditionally decorrelate random vectors. The problem of conditionally decorrelating random vectors has been considered before. One well-known approach is through the computation of canonical correlations. The canonical correlations approach provides a method of achieving exact or approximate conditional decorrelation. In the latter case, it can be viewed as providing an absolute measure of the difference between an approximate model and an exact model that is related to the Kullback-Leibler distance [1, 2].

Rather than using canonical correlations, our approach is based on the estimation-theoretic concept of predictive efficiency, which we review in this section. As will be shown, predictive efficiency can be viewed as providing a measure of the difference between an approximate and an exact model directly related to the intended use of a MAR model for estimation. Further, use of predictive efficiency also leads to significant savings in complexity of the resulting realization algorithm. Predictive efficiency has been introduced elsewhere [3, 40] and the relevant aspects of it are well known in the statistics and control communities. Therefore, our treatment is brief and we refer the reader to [19] for details.

To begin, we define $\varepsilon(z_2 \mid z_1)$ to be the mean-square error in the linear least-squares estimate of the length-n_2 vector z_2 based on the length-n_1 vector z_1:

$$\varepsilon(z_2 \mid z_1) \triangleq \mathrm{E}\Big(\|z_2 - \hat{\mathrm{E}}[z_2 \mid z_1]\|^2\Big) \tag{13a}$$

$$= \mathrm{trace}\big(P_2 - P_{12}^T P_1^{-1} P_{12}\big) \tag{13b}$$

where P_i is the positive-definite covariance matrix for z_i and P_{12} is the cross-covariance matrix for z_1, z_2. (We remind the reader that $\hat{\mathrm{E}}[\cdot \mid \cdot\cdot]$ denotes the linear least-squares estimator (see the proof of Proposition 1).)

Consider now the problem of estimating z_2 not from z_1 but from no more than r linear functionals of z_1 given by Vz_1 where $V \in \mathcal{M}_r$ which is the set of all matrices of size $\ell \times n_1$ with $\ell \leq r$. We can measure the quality of the estimate based on Vz_1 relative to that which can be obtained from z_1 by

$$\bar{\varepsilon}(z_2 \mid Vz_1) \triangleq \varepsilon(z_2 \mid Vz_1) - \varepsilon(z_2 \mid z_1) \tag{14a}$$

$$= \operatorname{trace}\left(P_{12}^T P_1^{-1} P_{12}\right) - \operatorname{trace}\left(P_{12}^T V^T (V P_1 V^T)^{-1} V P_{12}\right). \tag{14b}$$

The idea of predictive efficiency is to minimize $\bar{\varepsilon}(z_2 \mid Vz_1)$ over $\mathcal{M}_r$. Let

$$\widehat{V} \triangleq \arg \min_{V \in \mathcal{M}_r} \bar{\varepsilon}(z_2 \mid Vz_1)\,. \tag{15}$$

Notice that the minimum is lower bounded by zero and equality obtains if and only if Vz_1 conditionally decorrelates z_1 and z_2. Therefore, we can interpret $\bar{\varepsilon}(\cdot \mid \cdot\cdot)$ as a measure of distance from Markovianity although it is not a true distance because it is not symmetric. The optimal $V \in \mathcal{M}_r$ according to the predictive efficiency measure is provided in the following proposition, which is proved in [19].

PROPOSITION 5 *Let $U\Lambda U^T$ be the eigen-decomposition of $P_1^{-1/2} P_{12} P_{12}^T P_1^{-T/2}$ with the eigen-value matrix $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{n_1})$ and $\lambda_i \geq \lambda_j$ for $i \leq j$. Let $r \leq n_1$. Then $\widehat{V} \triangleq \arg\min_{V \in \mathcal{M}_r} \bar{\varepsilon}(z_2 \mid Vz_1)$ is given by the first r rows of $U^T P_1^{-1/2}$.*

In the sequel, we call the pair of matrices (U, Λ) of Proposition 5 the *predictive efficiency matrices*. The computational complexity of computing these matrices is $O(n_1^2 n_2 + n_1^3)$. The $n_1^2 n_2$ term comes from the formation of the matrix $P_1^{-1/2} P_{12} P_{12}^T P_1^{-T/2}$. The n_1^3 term comes from the fact that we must invert the matrix square root of P_1 and compute an eigen-decomposition of an $n_1 \times n_1$ matrix. The inversion of P_2 is not required because the predictive efficiency method is asymmetric. In fact, P_2 plays no role in the computation of the predictive efficiency matrices. In contrast, the canonical correlations method of [22, 24, 26] requires the inversion of *both* P_1 and P_2 because it is symmetric. It is precisely this difference in symmetry that accounts for the efficiency of our approach as compared to previous methods. In the context of the MAR stochastic realization problem, n_2 is related to problem size while n_1 is related to state dimension and can be chosen to be independent of n_2. Thus, the asymptotic complexity of the predictive efficiency method is $O(n_2)$ whereas that of the canonical correlations approach is $O(n_2^3)$.

The main message of Proposition 5 is that $\widehat{V} z_1$ does the best job (in the sense of (15)) of conditionally decorrelating z_1 and z_2 subject to the constraint that $\widehat{V}$ may have no more than r rows. We have a need to generalize this idea to consider the problem of (approximately) conditionally decorrelating more than two random vectors. For this purpose, we define (with an abuse of notation) the following generalization of $\bar{\varepsilon}(\cdot \mid \cdot\cdot)$:

$$\bar{\varepsilon}(z_1, z_2, \ldots, z_{q+1} \mid Vz_0) \triangleq \max_i \{\bar{\varepsilon}(z_i \mid Vz_0)\}. \tag{16}$$

The corresponding predictive efficiency problem is

$$\widehat{V} = \arg \min_{V \in \mathcal{M}_d} \bar{\varepsilon}(z_1, z_2, \ldots, z_{q+1} \mid Vz_0). \tag{17}$$

A special case of (17) is at the heart of our approach to the MAR realization problem as we show in the next section. This special case is the one for which $z_0 = [z_1^T \; z_2^T \ldots z_q^T]^T$. In this case, we can view $\bar{\varepsilon}$ as a measure of Markovianity since Vz_0 conditionally decorrelates $\{z_i\}_{i=1}^{q+1}$ if and only if $\bar{\varepsilon}(z_1, z_2, \ldots, z_{q+1} \mid Vz_0) = 0$. Unfortunately, unlike the case for a pair-wise predictive efficiency problem (cf., Proposition 5), to our knowledge, a procedure for solving this higher order predictive efficiency problem is not known. However, by considering q pair-wise predictive efficiency problems instead of (17), we can obtain a good sub-optimal solution.

Rather than attempt to conditionally decorrelate all $q+1$ random vectors in the set $\{z_i\}_{i=1}^{q+1}$ at once, we instead consider each one in turn. That is, for each i, we seek a linear function of z_i that (approximately) conditionally decorrelates it from the others. Using the predictive efficiency criterion, this becomes formally

$$V_{i,r_i} = \arg \min_{V \in \mathcal{M}_{r_i}} \bar{\varepsilon}(z_i^c \mid Vz_i) \tag{18}$$

where the r_i satisfy $\sum_{i=1}^{q} r_i \le d$ and $z_i^c = [z_1^T, z_2^T, \ldots, z_{i-1}^T, z_{i+1}^T, \ldots, z_{q+1}^T]^T$, a vector consisting of z_j for $j \neq i$. This pair-wise problem is solved by computing the predictive efficiency matrices (U_i, Λ_i) as explained in Proposition 5.

Having solved these q pair-wise problems, we concatenate the resulting matrices V_{i,r_i} to form $\bar{V}$, our sub-optimal solution to (17):

$$\bar{V} \triangleq \mathrm{diag}(V_{1,r_1}, V_{2,r_2}, \ldots, V_{q,r_q}). \tag{19}$$

To completely define our sub-optimal solution $\bar{V}$, we need to specify exactly how the r_i are chosen. Our approach is to first compute all of the q sets of predictive efficiency matrices$\{(U_i, \Lambda_i)\}_{i=1}^{q}$. Then, we create one ordered list consisting of all of the eigenvalues and select the largest d eigenvalues from our list, thereby determining the number, r_i, of rows taken from each U_i. To be sure, one can consider other ways of specifying the r_i. Some of these are discussed in Section 9.

7. Stochastic Realization

Recall from Section 3 that our objective is to build an internal MAR model $x(\cdot)$ such that its fine-scale sub-process x^M has a covariance matrix P_{x^M} that closely approximates the (given) $N \times N$ covariance matrix P_{f^M} of the process f^M. In Sub-section 7.1 we provide an $O(N^2)$ algorithm to address this problem and in Sub-section 7.2 we introduce an approximation that results in $O(N)$ algorithm. The algorithms we describe compute MAR models assuming a given fixed target value for the state dimension, d. Additionally, while our algorithms possess some properties that are similar to those possessed by model selection algorithms in the graphical modeling literature, our approach is substantially different and deals with issues that do not commonly arise in the graphical modeling literature. We discuss this point more thoroughly in the concluding section of the paper.

7.1. $O(N^2)$ Algorithm

In constructing an internal MAR model $x(\cdot)$ for P_{f^M} we define another locally internal tree-indexed (and not necessarily MAR) process as an intermediate step. This intermediate process $f(\cdot)$ has as its finest-scale sub-process f^M, the signal to be modeled. At any node s not at the finest scale, we define the value of $f(\cdot)$ at node s scale-recursively as $f(s) \triangleq V_s f_s^{m(s)+1}$ where each local internal matrix V_s is derived based on a predictive efficiency criterion (detailed shortly).

From the set of local internal matrices $\{V_s\}$ and the given fine-scale covariance P_{f^M}, the statistics $P_{f(s)}$ and$P_{f(s\alpha_i)f(s)}$ are easily computed. In turn, these may be used to define the dynamical model for $f(\cdot)$:

$$f(s\alpha_i) = A(s\alpha_i)f(s) + \mu(s\alpha_i), \tag{20a}$$

$$Q(s\alpha_i) \triangleq \mathrm{E}[\mu(s\alpha_i)\mu(s\alpha_i)^T] \tag{20b}$$

where $A(s\alpha_i)$ and $Q(s\alpha_i)$ are computed from $P_{f(s)}$ and $P_{f(s\alpha_i)f(s)}$ as described in (10) (in which $x(s\alpha_i)$ and $x(s)$ are replaced by $f(s\alpha_i)$ and $f(s)$, respectively).

If the process $f(\cdot)$ has the scale-recursive Markov property then $\mu(\cdot)$ is a white noise process, uncorrelated with $f(0)$. Hence, (20) is an exact MAR model for f^M [26]. As we will explain, this will occur if no approximation is made in the predictive efficiency step that defines the local internal matrices. If, on the other hand, we *do* make an approximation in the predictive efficiency step, then $\mu(\cdot)$ will not be a white noise process, uncorrelated with $f(0)$. However, in this case, we can *define* an *approximate* model by assuming that $\mu(\cdot)$ is white and uncorrelated with $f(0)$. That is, we define the internal MAR process $x(\cdot)$ to approximate $f(\cdot)$ as

$$x(s\alpha_i) = A(s\alpha_i)x(s) + w(s\alpha_i), \tag{21a}$$

$$\mathrm{E}[w(s\alpha_i)w(s\alpha_i)^T] = Q(s\alpha_i) \tag{21b}$$

where $A(\cdot)$ and $Q(\cdot)$ are the same as in (20) and $w(\cdot)$ is white, uncorrelated with $x(0)$. Note that, while this results in an approximate model ($P_{x^M} \neq P_{f^M}$), the state covariances $P_{x(s)}$ at each node s and the child-parent cross-covariances $P_{x(s\alpha_i)x(s)}$ for each child-parent pair of nodes exactly match $P_{f(s)}$ and $P_{f(s\alpha_i)f(s)}$, respectively. Consequently, the $d \times d$ diagonal blocks of P_{x^M} exactly match those of P_{f^M}.

It remains only to specify the predictive efficiency step in which we define the local internal matrices $\{V_s\}$. To obtain an exact model requires that $f(\cdot)$ have the scale-recursive Markov property which it does (by definition) if $f(s) = V_s f_s^{m(s)+1}$ conditionally decorrelates the set of vectors $\{f(s\alpha_i)\}_{i=1}^q \cup f_{s^c}^{m(s)+1}$ for all $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$. This occurs exactly when

$$\bar{\varepsilon}(f(s\alpha_1), f(s\alpha_2), \ldots, f(s\alpha_q), f_{s^c}^{m(s)+1} \mid V_s f_s^{m(s)+1}) = 0. \tag{22}$$

Typically, any V_s satisfying (22) has too many rows, leading to models with impractically high state dimensions. Therefore, to obtain lower dimensional states, we may apply the procedure described in Section 6 to find a sub-optimal solution to the predictive efficiency problem

$$\arg \min_{V \in \mathcal{M}_d} \bar{\varepsilon}\Big(f(s\alpha_1), f(s\alpha_2), \ldots, f(s\alpha_q), f_{s^c}^{m(s)+1} \mid V f_s^{m(s)+1}\Big) \tag{23}$$

thereby constraining V_s to have no more than d rows. This predictive efficiency step provides useful information on whether the state dimension d is large enough or too large. Specifically, the predictive efficiency matrices provide a rank-ordered quantification of the value of keeping each successive state variable. If d is too small, there will be significant remaining value in keeping more than d variables. If d is too large, we could have kept fewer variables with negligible loss of performance.

The asymptotic computational complexity of our realization approach stems from two sources. The first is the computation of the local internal matrices. If d is chosen independent of problem size N then, as described in Section 6, the complexity of finding our sub-optimal solution to (23) is $O(q^n)$ because $f_{s\alpha_i^c}^{m(s)+1}$ (which plays the role of $z_i^c$of Section 6), is length $O(q^n)$. Summing this up over all nodes we arrive at an $O(q^{2M})$ complexity which is equivalent to $O(N^2)$ because $N \propto q^M$. While our focus is on asymptotic complexity (just described) it is worth mentioning that for a fixed problem size, complexity is proportional to d^3 because the computation of the predictive efficiency matrices involves an eigen-decomposition of a $d \times d$ matrix.

The second source of complexity is the computation of P_{f^n}, the statistics of $f(\cdot)$ at scale n which are needed to compute the local internal matrices at scale $n-1$. We have that

$$P_{f^n} = \mathcal{V}_n P_{f^{n+1}} \mathcal{V}_n^T \tag{24}$$

where $\mathcal{V}_n$ is a block diagonal matrix whose diagonal blocks are V_s for $s \in \mathcal{T}_0(n)$, lexicographically ordered. By construction, each row of $\mathcal{V}_n$ has at most d non-zero elements. Taking advantage of this sparsity, we can compute $\{P_{f^n}\}_{n=0}^{M-1}$ with complexity $O(N^2)$.

7.2. Boundary Approximations

The predictive efficiency-based MAR realization method proposed in the previous section has complexity proportional to N^2 (where the signal or (lexicographically ordered) image to be modeled has total size N). While this is relatively efficient as compared to other approaches [10, 22, 24, 26], it is still too burdensome for some problems, particularly those arising in image processing. The source of this complexity stems from the fact that, in computing the predictive efficiency matrices, we focus on estimating *every* element of a *large* random vector, z_2, from a small one, z_1. (We use the notation z_1 and z_2 as shorthand for vectors that arise in the computation of a sub-optimal solution to the predictive efficiency problem posed in (23).) In this section, we propose the *boundary approximation* which focuses on estimating only a small number of elements of z_2 which are temporally or spatially close to z_1. As we will show, this boundary approximation leads to a realization algorithm that has complexity proportional to N. We note that a similar approximation is employed in conjunction with canonical correlations in [22, 26]. However, in some sense it is more severe because, due to the previously discussed symmetry of canonical correlations, it requires truncating *both* z_2 *and* z_1. Since predictive efficiency is not symmetric, we need only truncate z_2 to obtain an $O(N)$ algorithm.

Intuitively, the boundary approximation should not be a severe one for processes that are Markov (or nearly so) or have quickly decaying long-range correlations. In the former case, the boundaries of z_1 contain all the relevant information for estimating the more distant random variables. Therefore, a summary of z_1 (i.e., Vz_1) that does a good job of estimating these local variables ought to be sufficient for estimating the distant ones. In the latter case of quickly decaying long-range correlations, distant random variables are negligibly correlated with z_1 and, therefore, do not substantially contribute to the mean-square estimation error. In our examples, we will show that the class of processes for which the boundary approximation results in small modeling error is, in fact, considerably *richer* than Markov and fast-decorrelating processes. A theoretical understanding of the foregoing points is a topic of current research as we will discuss in Section 9.

Let us begin by examining the two sources of the N^2 complexity of our realization approach in the context of building a MAR model for a one-dimensional random signal (as opposed to a two-dimensional random field, to which we return later). The first source comes from finding the local internal matrices which are sub-optimal solutions to (23). The second is the computation of P_{f^n} for all scales $n \in \{0, 1, \dots, M-1\}$. With respect to the former, we noted in Section 6 that summarizing a length-n_1 vector z_1 for the purposes of estimating a length-n_2 vector z_2 has complexity $O(n_2)$. In the context of the stochastic realization problem, this translated into a complexity of $O(N)$ per node which, when summed over all $O(N)$ nodes, lead to an overall $O(N^2)$ complexity for computing the internal matrices. This suggests that we can reduce the overall complexity to $O(N)$ by somehow ignoring all but a small portion of z_2 (whose size is independent of N).

To this end, let k be an integer chosen independent of n_2 and H_k be a selection matrix such that, when post-multiplied by z_2, selects the kd elements of z_2 that are temporally closest to z_1 as illustrated in Figure 7. We will call the kd elements selected by H_k the size-k boundary of z_1. With this notation, consider

$$\widehat{V}_k \triangleq \arg\min_{V \in \mathcal{M}_r} \bar{\varepsilon}(H_k z_2 \mid V z_1)\,. \tag{25}$$

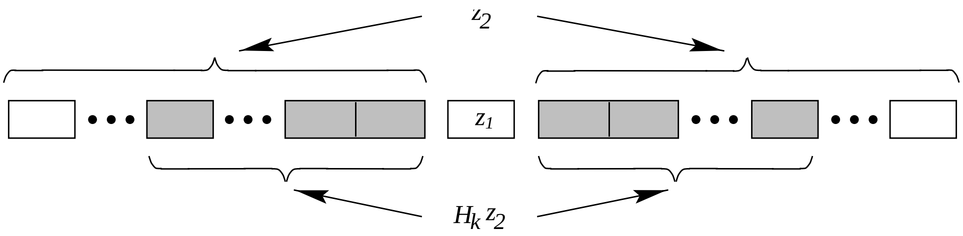


Figure 7. $H_k z_2$ selects the kd elements of z_2 (shaded) that are temporally closest to z_1.

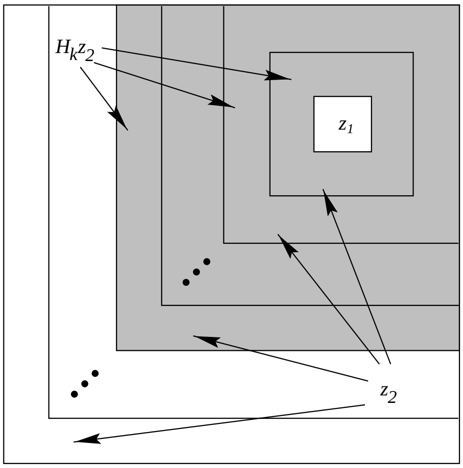


Figure 8. $H_k z_2$ selects the kd elements of z_2 (shaded) that are spatially closest to z_1.

Since the complexity of computing $H_k P_{z_2} H_k^T$ (a quantity needed to solve (25)) is independent n_2, the solution of (25) can be computed with complexity that is also independent of n_2. We then view $\widehat{V}_k$ as a sub-optimal solution to (15). Using this idea in our stochastic realization approach, we arrive at a complexity of $O(N)$ for computing the internal matrices.

We now turn to the second source of the N^2 complexity of the MAR stochastic realization approach of Section 7.1—the computation of the P_{f^n}. The boundary approximation reduces this source of complexity as well since using (25) implies that we need not compute all of P_{f^n}. Rather, only a diagonal band of size that is a function of k is needed because we never consider cross-correlations involving elements that are further than kd away from the node at which the current predictive efficiency matrices are being computed. It is not hard to show that the total complexity of computing the required diagonal bands of the P_{f^n} matrices for all $n \in \{0, 1, \ldots, M-1\}$ is $O(N)$. Hence, the overall asymptotic complexity of the MAR realization algorithm with the boundary approximation is $O(N)$.

We now discuss the boundary approximation for modeling two-dimensional random fields. In this case, the vector z_1 represents a pixel of a random field and z_2 the rest of the random field as illustrated in Figure 8. The matrix H_k selects the elements of z_2 that lie in the k concentric square annuli, each of which is one pixel wide, that surround z_1 (with the obvious modifications for boundary effects as illustrated). All of the complexity analysis provided previously for the one-dimensional case is identical for the two-dimensional case.

8. Examples

In this section we provide several examples illustrating the performance of the $O(N^2)$ realization algorithm of Sub-section 7.1 as well as the boundary approximation discussed in Sub-section 7.2. In our one-dimensional examples we always use dyadic trees ($q = 2$) and in our two-dimensional examples we always use quad-trees ($q = 4$).

8.1. One-Dimensional Realization and Estimation Examples

Our one-dimensional examples are intended to illustrate the application of our realization algorithms and, in particular, the power of the boundary approximation. Additionally, we demonstrate that an approximate model can achieve estimation results that are statistically indistinguishable from results based on an exact model. Our first example is the realization and estimation of fractional Brownian motion with Hurst parameter $H = 0.7$ (denoted fBm(0.7)). The correlation function for fBm(H) is [39]

$$r_H(t_1, t_2) = \frac{1}{2}\left(|t_1|^{2H} + |t_2|^{2H} - |t_1 - t_2|^{2H}\right). \tag{26}$$

The true fBm(0.7) covariance matrix, P_{f^M}, associated with 128 samples of fBm(0.7) on $(0, 1]$ is illustrated in Figure 9(a). The realized covariance matrix, P_{x^M}, associated with a MAR model with state dimension $d = 4$ and based on our full $O(N^2)$ algorithm is illustrated in Figure 9(b). In Figure 9(c) we have plotted $|P_{f^M} - P_{x^M}|$ where $|\cdot|$ is element-wise. Notice that even for this relatively low dimensional model, the approximation

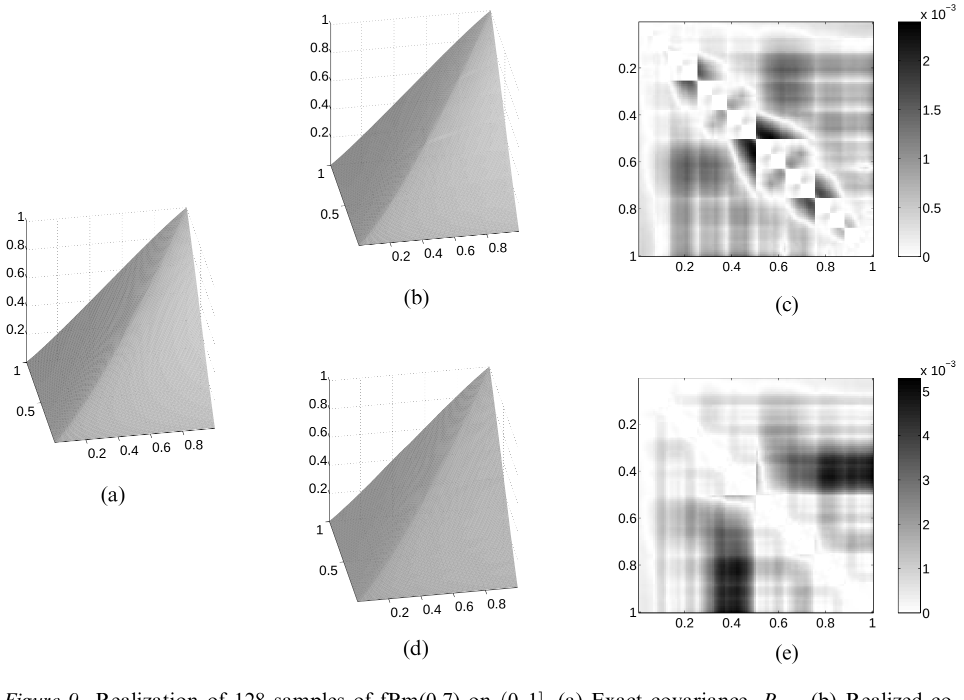


Figure 9. Realization of 128 samples of fBm(0.7) on $(0, 1]$. (a) Exact covariance, P_{f^M}. (b) Realized covariance, P_{x^M} ($d = 4$). (c) $|P_{f^M} - P_{x^M}|$ where P_{x^M} is from (b). (d) Realized covariance, P_{x^M}, using the boundary approximation ($d = 4$, $k = 1$). (e) $|P_{f^M} - P_{x^M}|$ where P_{x^M} is from (d).

is quite good, with the largest element-wise error on the order of 10^{-3}. In addition, that the 4×4 diagonal blocks of $|P_{f^M} - P_{x^M}|$ are zero can be plainly seen in Figure 9(c). Notice also that some of the largest errors correspond to correlations between elements that are spatially close. This is due to the fact that spatially close elements (like those at sample numbers 64 and 65) can be quite far apart in tree distance and the correlation between them suffers from errors induced by the approximation made at all the tree nodes between them.

In Figure 9(d), we have plotted the realized covariance, P_{x^M}, based on a MAR model for fBm(0.7), again with state dimension $d = 4$, but using the $O(N)$ boundary approximation algorithm. The boundary size is $k = 1$ which corresponds to designing local internal matrices to (approximately) conditionally decorrelate MAR variables at a given node from those indexed by the two nearest nodes at the same scale (or one nearest node if the given node is on the boundary). The modeling error $|P_{f^M} - P_{x^M}|$ is illustrated in Figure 9(e) and should be compared with Figure 9(c). Notice that the errors, while different, are of the same order, 10^{-3}. Since fBm(0.7) is not Markov and has slowly (polynomially) decaying correlations [4], this illustrates that the boundary approximation is effective for a *broader* class of processes than those that motivated it.

Next, we apply the MAR model for fBm(0.7) associated with Figure 9(b) to an estimation problem based on incomplete measurements corrupted by non-stationary noise. We emphasize that this is a problem that *cannot* be handled with fast transform techniques due to the non-stationarity of the process to be estimated and the process noise and the fact that the measurements are incomplete. Figure 10(a) is a sample path of fBm(0.7). Figure 10(b) illustrates noisy incomplete measurements of Figure 10(a). Measurements are taken over the first and last third of the process. No measurements are available over the middle third. The white measurement noise has variance 0.3 over the first third sub-interval and 0.5 over the last third sub-interval. Figure 10(c) shows the output of the MAR estimator [5] based on the model associated with Figure 9(b) (solid line) with one-standard-deviation error bars (dotted lines). The optimal estimate based on the exact fBm(0.7) statistics (rather than our approximate model of them) is also plotted (dashed line) in Figure 10(c). However it is not easily distinguishable from the MAR estimate since the two nearly coincide. Moreover, the difference between the two is well within the one-standard-deviation error bars. This demonstrates that the degree to which our MAR model deviates from the exact model is statistically irrelevant. Note that the optimal estimate requires $O(N^3)$ computations while the MAR estimator is $O(N)$. The MAR estimator also produces estimation error statistics with no additional computations beyond what are needed to compute the estimates themselves. In Figure 10(d) we have plotted the MAR error standard-deviations (solid line) and the optimal error standard-deviations (dashed line). The two nearly coincide, again illustrating that the degree to which our model deviates from an exact one is not relevant to this estimation problem.

Next, we illustrate MAR realizations using our $O(N^2)$ algorithm of a 12-th order stationary Markov process. The purpose of these examples is to show that, while fBm can be well modeled with state dimension $d = 4$, some processes require a higher state dimension. In Figure 11(a) we illustrate the true covariance matrix, P_{f^M}. Figure 11(b) is the realized covariance matrix, P_{x^M}, associated with a MAR model with state dimension $d = 4$. Notice that the errors $|P_{f^M} - P_{x^M}|$, which are plotted in Figure 11(c), are *much*

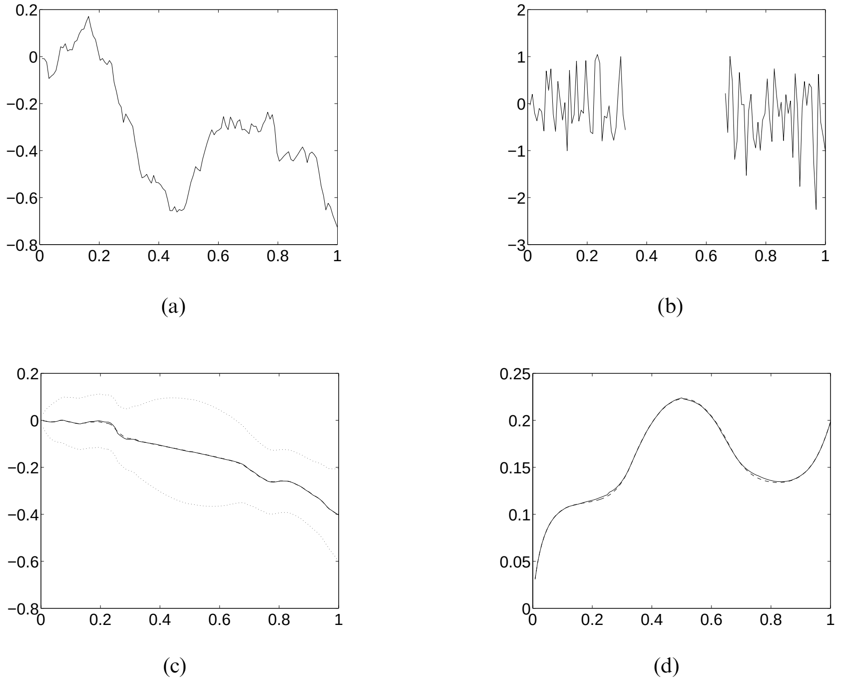
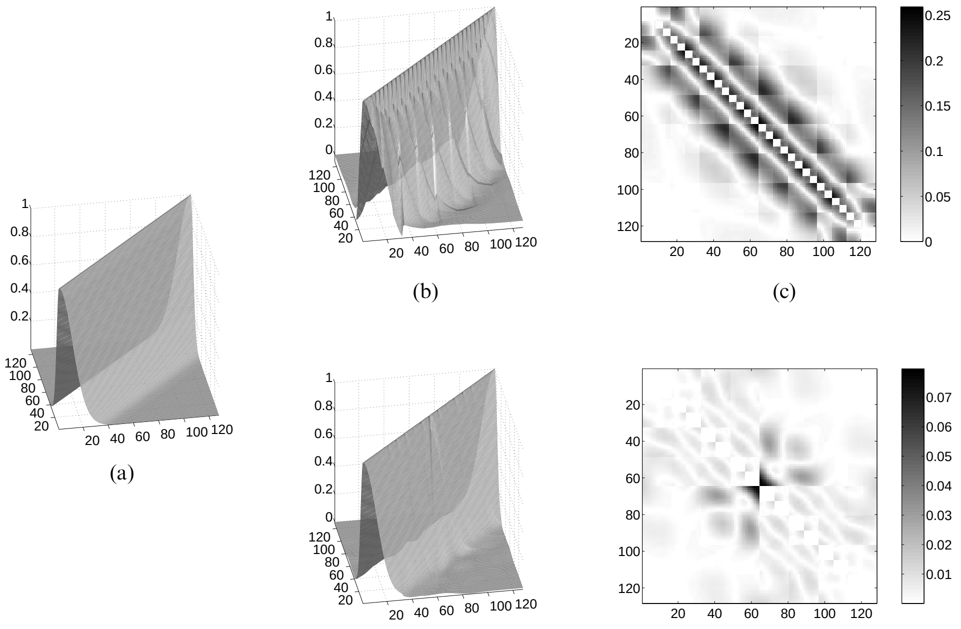


Figure 10. Estimation of fBm(0.7) using the model of Figure 9(b). (a) Sample-path using exact statistics. (b) Noisy, incomplete observations of (a). (c) MAR estimates (solid line), optimal estimates based on the exact statistics (dashed line), and plus/minus one standard deviation error bars (dotted lines). (d) Error standard deviation given by the MAR estimator (solid line) and based on the exact statistics (dashed line).

larger (25% of the process variance) than those associated with the fBm(0.7) model of Figure 9(c) which also has state dimension $d = 4$. If, however, we increase the state dimension to $d = 8$, we achieve a MAR realization with errors on the order of 7% of the process variance. This is illustrated in Figure 11(d) which shows P_{χ^M} and Figure 11(e) which shows $|P_{f^M} - P_{\chi^M}|$ for this higher state dimension model. A more accurate model of the 12-th order stationary Markov process than the one associated with Figure 11(d) requires maximum state dimension larger than $d = 8$.

To achieve modeling errors on the order of those depicted in Figure 11(e), one need not use a model with state dimension $d = 8$ at all nodes. It is possible to achieve similar performance with state dimensions that decrease at coarser scales. We illustrate this point in Figure 12. Figure 12(a) is the realized covariance matrix, P_{χ^M}, associated with a four-scale MAR model with state dimension 8 at scales 3 (the finest) and 2, state dimension 6 at scale 1, and state dimension 4 at scale 0 (the coarsest). The error $|P_{f^M} - P_{\chi^M}|$ is plotted in Figure 12(b) and is on the order of 8% of the process variance, comparable to that achieved with the $d = 8$ (at all nodes) model of Figure 11(e).

Figure 12(c) illustrates the realized covariance for another MAR model of the 12-th order stationary Markov process with state dimensions that vary with scale as described. However, in this case, the boundary approximation was used with boundary size $k = 3$.

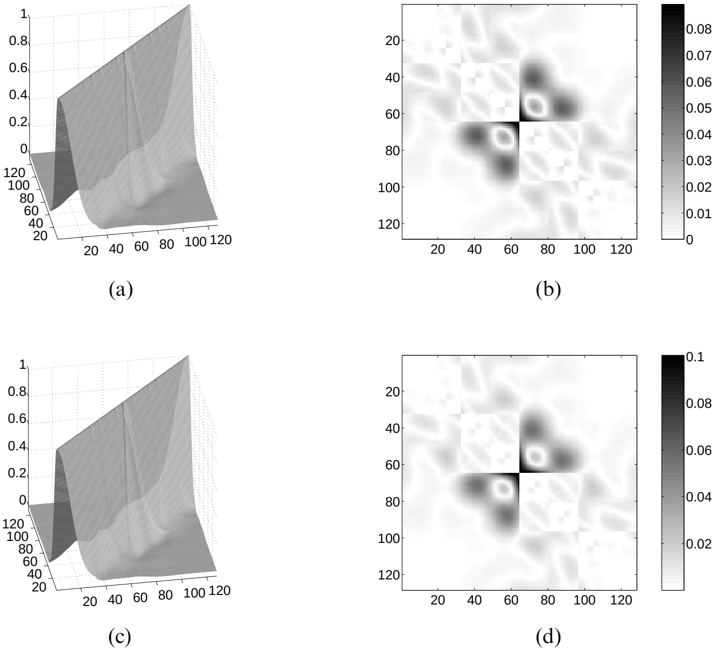


(a) (b) (c) (d) (e)

Figure 11. Realization of a 12-th order stationary Markov process. (a) Exact covariance, P_{f^M}. (b) Realized covariance, P_{x^M} $(d = 4)$. (c) $|P_{f^M} - P_{x^M}|$ where P_{x^M} is from (b). (d) Realized covariance, P_{x^M}, $(d = 8)$. (e) $|P_{f^M} - P_{x^M}|$ where P_{x^M} is from (d).

Errors are plotted in Figure 12(d) and should be compared with Figure 12(b). Notice that the errors,while slightly different, are on the same order (roughly 10% of the process variance). This illustrates that little modeling fidelity is lost in making the boundary approximation. In this case, this result is consistent with our intuition because the underlying process is 12-th order Markov and a boundary size $k = 3$ corresponds to keeping kd state elements on either side of the node being designed. In this example d varies from 4 to 8 so the number of boundary elements is always at least as large as the Markov order. Despite this fact, the results depicted in Figure 12(b) and Figure 12(d) are different because we are designing internal matrices to do different jobs. In the former case, we are attempting to conditionally decorrelate MAR variables at a given node from *all* other variables at the same scale. In the latter case, we are only considering the nearby variables at the same scale. Naturally, these two criteria lead to a different emphasis and different linear functionals that comprise the internal matrices.

Next we illustrate model fidelity as a function of boundary size. We again consider MAR models for the 12-th order stationary Markov process where the state dimension varies with scale as described previously. For different boundary sizes $k \in \{1, 2, 3, 4, 5, 6\}$ we computed a realization. We then compared the realized covariance to the true one with three different norms $||P_{f^M} - P_{x^M}||$: the Frobenius norm, induced 2-norm (maximum singular value) and maximum absolute value of the difference $|P_{f^M} - P_{x^M}|$. We point out



(a) (b)

(c) (d)

Figure 12. Realization for 128 samples of a 12-th order stationary Markov process. (a) Exact covariance, P_{f^M}. (b) Realized covariance, P_{x^M} (state dimension varies with scale (see text)). (c) $|P_{f^M} - P_{x^M}|$ where P_{x^M} is from (b). (c) Realized covariance using boundary approximation ($k = 3$). (d) $|P_{f^M} - P_{x^M}|$ where P_{x^M} is from (d).

that, in our realization procedure, we are not explicitly minimizing any of these norms. Figure 13 illustrates the value of these three norms as a function of boundary size. As expected, modeling fidelity improves as boundary size increases. Notice that boundary size $k = 3$ seems to be the appropriate choice under these norms since negligible improvement can be expected for larger sizes and substantial degradation obtains for smaller sizes.

As pointed out previously, the most significant modeling errors occur for samples that are close spatially but distant on the tree. In our next example, we explore the impact of

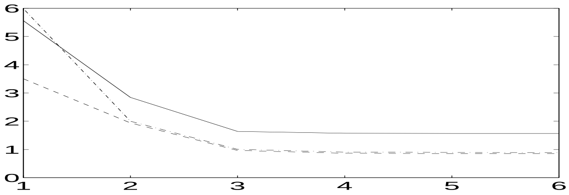
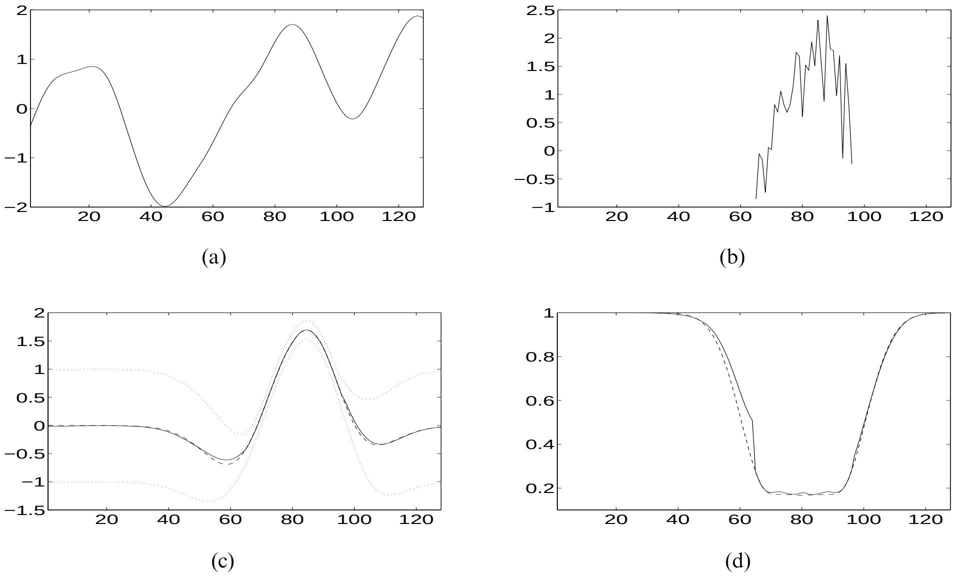


Figure 13. Boundary approximations for 12-th order stationary Markov process. $||P_{f^M} - P_{x^M}||$ is plotted as a function of boundary size k for three different norms: Frobenius (solid line), maximum singular value (dashed line), maximum element-wise absolute difference (dash-dot line). The last of these is multiplied by 10 so that it is on the same scale as the first two.



(a) (b)

(c) (d)

Figure 14. Estimation of a 12-th order stationary Markov process using the model of Figure 12(a). (a) Sample-path using exact statistics. (b) Noisy, observations of (a) over [65, 96]. (c) MAR estimates (solid line), optimal estimates based on the exact statistics (dashed line), and plus/minus one standard deviation error bars (dotted lines). (d) Error standard deviation given by the MAR estimator (solid line) and based on the exact statistics (dashed line).

this phenomenon on an estimation problem that is, in some sense, most likely to test this modeling weakness. Figure 14(a) is a sample path of a 12-th order stationary Markov process. Figure 14(b) illustrates noisy and incomplete measurements of Figure 14(a). Measurements are taken only over the interval [65, 96] which is just to the right of the point of greatest modeling error. The white measurement noise has variance 0.3. Figure 14(c) shows the output of the MAR estimator based on the model associated with Figure 12(a) (solidline) with one-standard-deviation error bars (dotted lines). The optimal estimate based on the exact statistics is also plotted (dashed line) in Figure 14(c). We can see that the largest estimation error due to modeling occurs just to the left of sample 64 as expected given the pattern of modeling error in Figure 12(b) and our measurement locations. Nevertheless, the difference between the optimal and the MAR estimates are well within the one-standard-deviation error bars and, therefore, are not particularly significant statistically. In Figure 14(d) we have plotted the MAR error standard-deviations (solid line) and the optimal error standard-deviations (dashed line). Again, the most significant errors are just to the left of sample 64 as expected and are small.

8.2. Two-Dimensional Sample-Path Generation Examples

We now turn to some image processing examples. These examples are intended to illustrate the application of our algorithms to modeling larger processes that arise in two-dimensions.

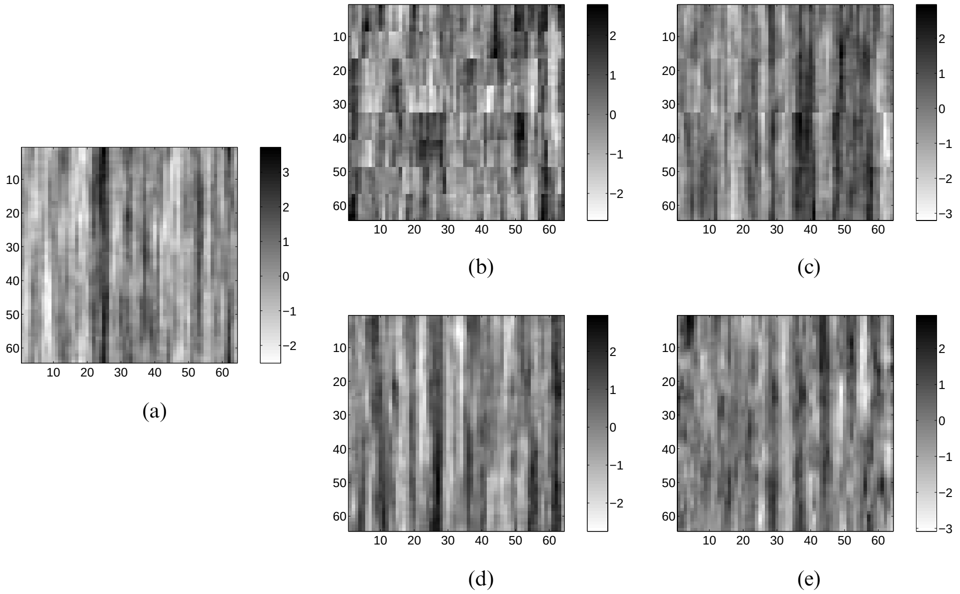


Figure 15. Sample-paths for wood texture of [32]. (a) Exact. (b) MAR model ($d = 16$). (c) MAR model ($d = 64$). (d) MAR model and the overlapping framework of [23] ($d = 16$). (e) MAR model and the overlapping framework with boundary approximation ($d = 16$, $k = 1$).

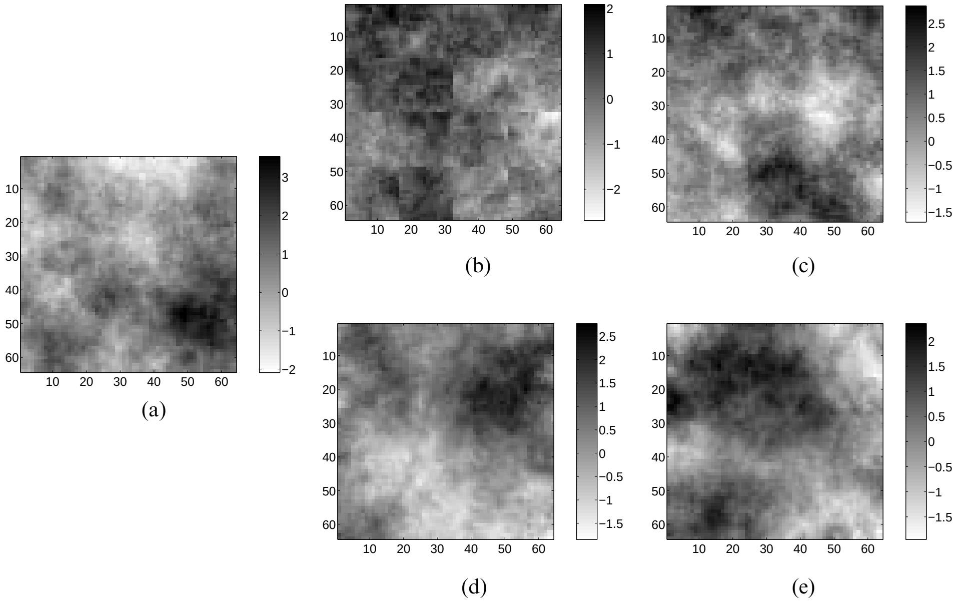
As will be shown, the boundary approximation is as effective in two-dimensions as it has already been shown to be in one-dimension. First, we consider building a MAR model for a Markov random field that mimics the texture of wood [32]. An exact 64 × 64 sample-path[10] is illustrated in Figure 15(a). Notice that this wood texture is highly correlated vertically and less-so horizontally. Figure 15(b) is a sample-path generated by a MAR model with state dimension $d = 16$. A distracting blockiness is apparent in this figure and is due to the quad-tree structure of our model and the small state dimension. Additionally, the extreme directionality of the wood texture makes this blockiness particularly easy to see. In some applications such blockiness is of no practical significance while in others, such as surface reconstruction where gradients must be taken [17], smoothness is required.

There are two techniques for reducing blockiness. One is to increase the state dimension. This is illustrated in Figure 15(c) which is a sample path based on a MAR model with state dimension $d = 64$. Unfortunately, increasing the state dimension leads to less efficient image processing algorithms. However, there is another approach: the overlapping framework of [23] in which the original field is oversampled to form a redundant field. A MAR model for this redundant field has multiple leaf nodes corresponding to each image pixel. The mapping back from MAR leaf-node variables to the image domain consists of simply averaging leaf-node variables corresponding to each individual pixel. Thus, this averaging of MAR variables does *not* introduce spatial averaging in the image domain. However, because multiple leaf nodes correspond to each pixel, the maximum tree distance between different pixels is significantly reduced as compared to a non-redundant model. This results in smaller modeling errors and the elimination of blocky artifacts. We refer the reader to [23] for details. A sample path for a MAR model based on this overlapping framework with state dimension $d = 16$ is illustrated in Figure 15(d). Finally, Figure 15(e) represents a sample image from a model constructed again using the overlapping framework but in this case also employing the boundary approximation. The boundary size is $k = 1$ which corresponds to conditionally decorrelating MAR variables with those residing at nodes one pixel away. Notice that there are no blocky artifacts in either Figure 15(d) or Figure 15(e), and both models produce wood textures comparable to that in Figure 15(a).

Next we consider sample-path generation of a two-dimensional, isotropic random field of interest in the geological sciences [27, 43]. The correlation function is

$$r_\ell(\rho) = \begin{cases} 1 - \dfrac{3\rho}{2\ell} + \dfrac{\rho^3}{2\ell^3} & \text{if } 0 \le \rho \le \ell, \\ 0 & \text{if } \rho > \ell \end{cases} \tag{27}$$

where $\rho = \sqrt{i^2 + j^2}$ and i, j are indices into a two-dimensional grid. An exact sample-path for $\ell = 40$ is illustrated in Figure 16(a). In Figure 16(b) and Figure 16(c) we provide a sample-path associated with a MAR model with state dimension $d = 16$ and $d = 64$, respectively. In Figure 16(d) the overlapping framework is used with $d = 16$. Finally, in Figure 16(e), the boundary approximation is employed with boundary size $k = 1$ in conjunction with the overlap framework ($d = 16$). As in the previous example, little degradation is evident when the boundary approximation is used.



(a)

(b)

(c)

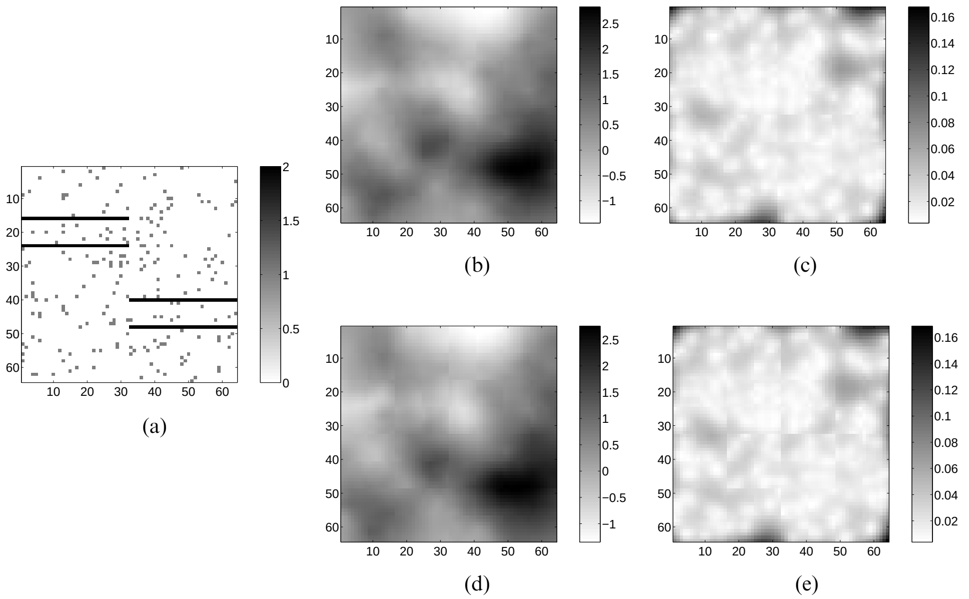
(d)

(e)

Figure 16. Sample-paths for the isotropic random field of (27). (a) Exact. (b) MAR model ($d = 16$). (c) MAR model ($d = 64$). (d) MAR model and overlapping framework of [23] ($d = 16$). (e) MAR model and overlapping framework with boundary approximation ($d = 16$, $k = 1$).

8.3. Two-Dimensional Multiresolution Data-Fusion Example

We conclude this section with a random field data-fusion problem based on irregular local and non-local measurements. In particular, we estimate the field illustrated in Figure 16(a) based on measurements whose locations are indicated in Figure 17(a). Each grey point in Figure 17(a) corresponds to a point measurement. Only about 20% of the fine-scale pixels are measured and, as can be seen, the point measurements are scattered irregularly. The four horizontal black lines correspond to non-local averages. That is, we have four line-integral measurements taken over the regions indicated with black lines. This example is included to show that a simple addition to our realization methodology provides a significant extension. In particular, this example illustrates the power of the MAR framework by allowing us to fuse heterogeneous and multiresolution measurements (such as the point and line-integral measurements of Figure 17(a)). An important point is that such data-fusion problems can be approached with the *same* MAR estimation algorithm and the same computational efficiency as for the case of estimation based only on point measurements. The key here is to design MAR models that explicitly include specific linear functionals as state variables at specific nodes on the tree. We note that while here we include specific linear functionals that represent measurements in our model exactly, other approaches have been developed that include measurements (as well as variables to estimate) approximately [8, 9, 19].



(a)

(b)

(c)

(d)

(e)

Figure 17. Random field multiresolution data-fusion example. (a) Measurement geometry: point measurements (grey), line-integral measurements (black). (b) Estimates based on the measurement geometry of (b) using a MAR model designed as discussed in the text ($d = 64$). (c) Error variances. (d) and (e) represent similar processing as (b) and (c) but with a MAR model with state dimension $d = 32$.

MAR models that accommodate non-local variables exactly have been considered before in [9]. This previous work required a MAR model to be provided and then *augmented* it to include the specific desired linear functionals. In contrast, our approach is to tailor our model to the non-local variables by *first* incorporating them and then building a MAR model around them, conditioning on the information they carry. More specifically, we pre-define some of the state variables in our MAR model so that (i) each line-integral appears as a state element (so we can incorporate a measurement associated with it, cf. (3)) and (ii) the resulting model is internal.

The first point (i) requires pre-defining four state elements. We incorporate each of the four line-integral measurements in a model by placing one measurement at each of the four first-scale nodes of a quad-tree (i.e., at each of the four nodes that are children of the root node). The second point (ii) requires propagating the information contained in these four state elements down the tree to maintain an internal model [9, 19]. After this is done, our realization algorithm is applied with the additional step that, when computing a given local internal matrix at node s, we first *condition* (by computing linear least-squares residuals) the relevant random quantities on the information already present at node s due to the prior incorporation of the non-local linear functionals associated with the four line-integral measurements. In other words, the criterion we use in rank-ordering the additional (unconstrained) linear functionals that define the state is in terms of the additional predictive efficiency they provide given that we have already committed to keeping the specific linear functional corresponding to a measurement. Our construction necessarily produces an internal model since we ensure that each linear functional at each node is, indeed, a linear functional of the state variables at the immediate descendent nodes (the computation of linear least-squares residuals is a linear operation).

Applying a MAR model so designed (and with $d = 64$) to the data-fusion problem at hand, we obtain the MAR estimates and estimation error variances shown in Figure 17(b) and Figure 17(c), respectively. Figure 17(d) and Figure 17(e) represent the same processing as Figure 17(b) and Figure 17(c), respectively, but with a MAR model with state dimension $d = 32$. Notice that the estimates of Figure 17(d) are nearly identical to those of Figure 17(b). Additionally, the estimation error variances of Figure 17(e) are on the order as those of Figure 17(c) and some blocky artifacts can be seen in the former due to the lower dimensionality of the model.

9. Conclusion

This paper represents a substantial contribution to stochastic realization theory for internal MAR processes. In contrast to previous approaches, ours leads to fast algorithms for internal models. First, our detailed analysis of internality provides a parameterization of internal states in terms of local internal matrices. This new parameterization leads to a scale-recursive characterization of Markovianity for tree-indexed processes. Based on the principles of internality, scale-recursive Markovianity, and predictive efficiency, we developed a general-purpose stochastic realization algorithm with complexity quadratic in problem size. In contrast, the most efficient general purpose algorithm previously developed is quartic in problem size [22, 24, 26] and does not provide internal models. Finally, with an approximation, we obtained a realization algorithm with complexity linear in problem size and demonstrated its effectiveness with a variety of examples.

One of our examples (the data-fusion example of Section 8.3) showed that if we know the form of linear non-local measurements, we can tailor a model to them. Since our model construction procedure scales well with problem size, we have, in essence, provided a "just-in-time" modeling methodology: Once the form of the measurements is known, we construct a model that incorporates these non-local functionals as state variables using one of our efficient realization algorithms. Then we fuse the (local and non-local) data using the efficient optimal estimation algorithm for MAR models. This is a non-trivial idea for adaptive data-fusion.

There are several fundamental open questions in MAR stochastic realization theory, some of which are suggested by our work. The first question, raised in Section 6, is how to solve the higher-order predictive efficiency problem (17). Our sub-optimal solution of solving several pair-wise problems also raises issues. One issue is how to choose the r_i, the number of linear functionals of z_i to keep. Our approach of choosing the r_i implicitly by keeping the linear functionals corresponding to the d highest eigenvalues has one unfortunate consequence; the collection of linear functionals may contain redundant information. A way to avoid this redundancy is to consider adding linear functionals sequentially where at each sequential step we add one or more linear functionals that have the highest predictive efficiency conditioned on the linear functionals that have been chosen during previous steps. One simple way to do this is to first incorporate linear functionals from z_1, then from z_2, etc., an approach that requires specifying the r_i sequentially rather than collectively. This will produce models that depend on the order in which the z_i appear in the sequence. More complex alternatives (e.g., cycling through the z_i several times, incorporating smaller numbers of linear functionals at each step) can potentially achieve greater statistical fidelity with an increase in computational load. Assessing the tradeoffs in model accuracy versus complexity of the realization procedure represents one direction for enhancing the procedure described in this paper.

Another issue raised by our work concerns the boundary approximation. While we have provided some intuition for why the boundary approximation works for approximately Markov process and processes with fast-decaying long-range correlations, we have shown that it is also remarkably effective for processes, like fBm(0.7), which do not have these characteristics. Characterizing the range of applicability of the boundary approximation is a topic of current study. Another related topic is generalizing the boundary approximation to consider summarizing distant variables rather than discarding them (e.g., perhaps in a multipole- or mean-field-like fashion in which the effects of distant variables are treated in aggregate).

Finally, there are several issues which we have not mentioned in the body of this paper but are important. One issue concerns our reliance on complete knowledge of the covariance matrix P_{f^M}. In many real-world problems, it is unlikely one will have precise knowledge of the entire covariance matrix of the process to be modeled. Moreover, in large image processing (and higher-dimensional) problems, such precise knowledge is impractical due to the large amount of memory it would require. We are currently working on two approaches that do not require complete knowledge of the underlying covariance. For problems in which partial covariance information is available, we are developing a covariance extension technique that builds a MAR model for an extension of the known covariance information without explicitly finding the full extension itself. For problems in which only sample-path data are available, we are working on techniques to estimate the MAR parameters $A(\cdot)$ and $Q(\cdot)$ directly.

A second open issue is the specification and application of a global criterion for MAR model identification. All previous approaches, as well as our own, are based on local criteria, involving the construction of states independently rather than jointly. The use of a global criterion for joint state construction is elusive and difficult. Further work is required to develop efficient algorithms based on a global criterion or to assess the effect on the global criterion of locally-chosen variables.

A third open issue is the joint impact of the choice of the branching factor q and the state dimension d. The impact of q on complexity is rather weak and not of real significance. Its impact on accuracy, however, deserves some study. Whenever we introduce several state vectors (corresponding to several tree branches) that are to be decorrelated by a single, common, parent, the larger the number of states vectors (branches) the poorer the job we will do for a fixed state dimension d. How one trades off d and q is an open issue.

Another set of issues arise when contrasting our work with other well-known results from graphical modeling. In the context of graphical models there is a significant literature for model construction [20]. While these methods are applicable to tree models, our methods are more powerful for our purposes. Specifically, most of the existing literature deals with finite-state models and thus does not address issues such as keeping sets of variables to form state vectors. Furthermore, while the idea of hidden variables is well-known in the graphical model literature, the concept of internality is a new one for that field. The fact that we can deal with such variables exactly rather than approximately is another distinguishing factor. Also, the existing literature does not, in general, worry much about the cardinality of the hidden variables, while the dimensionality of our state vectors is a major concern. Of course, we deal with all of these issues in the context of tree models and linear models with second-order statistics. This begs the question of extensions to domains closer to those considered in most of the graphical model literature.

MAR models offer substantial computational advantages for statistical inference as long as they have state dimension that is small relative to the problem size, N. This raises the question: what class of processes can be captured with a MAR model with state dimension that is independent of (or a slowly growing function of) N? This is, in some sense, the deepest question concerning MAR processes and the work presented here represents only one step toward its resolution.

Appendix A
Proofs

A.1. Proof of Proposition 2

Proof of Proposition 2: We begin with the "only if" direction. Given that $x(\cdot)$ is a locally internal MAR process, it has dynamics of the form (1). Thus, for all $s \in \mathcal{S}_0 - \mathcal{T}_0(M)$

$$x(s\alpha_i) = A(s\alpha_i)x(s) + w(s\alpha_i)\,. \tag{A.1}$$

Since $w(\cdot)$ is white and uncorrelated with $x(0)$, it follows that $w(s\alpha_i)$ is uncorrelated with $x(s)$. Therefore, (A.1) represents the linear least-squares estimate of $x(s\alpha_i)$ from $x(s)$ plus the estimation error $w(s\alpha_i)$. Then, (9) follows from (5) together with standard linear least-squares formulae.

To show the "if" direction, notice that (9) implies that by the MAR dynamics

$$x_s^{m(s)+1} = \underbrace{\begin{bmatrix} J_{s\alpha_1} \\ J_{s\alpha_2} \\ \vdots \\ J_{s\alpha_q} \end{bmatrix}}_{I} P_{x_s^{m(s)+1}} V_s^T \left(V_s P_{x_s^{m(s)+1}} V_s^T\right)^{-1} x(s) + \underbrace{\begin{bmatrix} w(s\alpha_1) \\ w(s\alpha_2) \\ \vdots \\ w(s\alpha_q) \end{bmatrix}}_{\triangleq\, w}. \tag{A.2}$$

Pre-multiplying (A.2) by V_s results in$V_s x_s^{m(s)+1} = x(s) + V_s w$. To conclude the proof we now show that the second term $V_s w$ is zero. For notational simplicity, let us define

$$R \triangleq P_{x_s^{m(s)+1}} - P_{x_s^{m(s)+1}} V_s^T \left(V_s P_{x_s^{m(s)+1}} V_s^T\right)^{-1} V_s P_{x_s^{m(s)+1}}. \tag{A.3}$$

The covariance matrix for $V_s w$ is

$$\begin{aligned} \mathrm{E}[V_s w w^T V_s^T] &= V_s \operatorname{diag}\left[Q(s\alpha_1), Q(s\alpha_2), \ldots, Q(s\alpha_q)\right] V_s^T && \text{(A.4a)} \\ &= V_s \operatorname{diag}\left[J_{s\alpha_1} R J_{s\alpha_1}^T, J_{s\alpha_2} R J_{s\alpha_2}^T, \ldots, J_{s\alpha_q} R J_{s\alpha_q}^T\right] V_s^T && \text{(A.4b)} \\ &= V_s R V_s^T && \text{(A.4c)} \\ &= 0 && \text{(A.4d)} \end{aligned}$$

where the first equality follows from the definition of w in (A.2) and the second equality follows from the definition of $Q(s\alpha_i)$ given in (9) and of R given above. The third equality follows from the fact that R is block diagonal because it is the estimation error covariance matrix in estimating $x_s^{m(s)+1}$ from $x(s)$ and $x(s)$ conditionally decorrelates $\{x(s\alpha_i)\}_{i=1}^q$ by the Markov property. The fourth equality follows from the definition of R. Since $V_s w$ has zero-mean and zero covariance it is deterministically zero. This completes the proof. ■

A.2. Completion of Proof of Proposition 4

To complete the proof we need to show that for an arbitrary s in $\mathcal{S}_0 - \mathcal{T}_0(M)$, $x(s)$, which has the scale-recursive Markov property, conditionally decorrelates the vectors in the set $\{x_{s\alpha_i}^M\}_{i=1}^q \cup \{x_{s^c}^M\}$. This is trivially true for $m(s) = M - 1$ since the two sets $\{x(s\alpha_i)\}_{i=1}^q \cup \{x_{s^c}^{m(s)+1}\}$ and $\{x_{s\alpha_i}^M\}_{i=1}^q \cup \{x_{s^c}^M\}$ coincide. Suppose that for all $s \in \mathcal{T}_0(n)$, $x(s)$ conditionally decorrelates $\{x_{s\alpha_i}^M\}_{i=1}^q \cup \{x_{s^c}^M\}$ and consider the case for which $s \in \mathcal{T}_0(n-1)$. For an arbitrary node s at scale $n-1$ and for an arbitrary child of s we have that

$$\begin{aligned} x_{s\alpha_i}^M &= \hat{\mathrm{E}}\left[x_{s\alpha_i}^M \mid x^n\right] + \tilde{x}_{s\alpha_i}^M && \text{(A.5a)} \\ &= \hat{\mathrm{E}}\left[x_{s\alpha_i}^M \mid x(s\alpha_i)\right] + \tilde{x}_{s\alpha_i}^M && \text{(A.5b)} \end{aligned}$$

and

$$x^M_{s\alpha_i^c} = \hat{\mathrm{E}}\left[x^M_{s\alpha_i^c} \mid x^n\right] + \tilde{x}^M_{s\alpha_i^c} \tag{A.6a}$$

$$= \hat{\mathrm{E}}\left[x^M_{s\alpha_i^c} \mid x^n_{s\alpha_i^c}\right] + \tilde{x}^M_{s\alpha_i^c} \tag{A.6b}$$

where in these identities we've used the induction hypothesis. It follows that the errors $\tilde{x}^M_{s\alpha_i}$ and $\tilde{x}^M_{s\alpha_i^c}$ are uncorrelated with each other (due to the induction hypothesis) and with x^n (due to the orthogonality property of linear least-squares estimation). By assumption, $x(s)$ is an internal state and so it is a linear combination of its children. That is, we have that for some V_s, $x(s) = V_s x^n_s$.

We now use these facts to show that $x^M_{s\alpha_i}$ and $x^M_{s\alpha_i^c}$ are uncorrelated when conditioned on $x(s)$. By assumption, $x(s)$ conditionally decorrelates $x^n_{s\alpha_i^c}$ from $x(s\alpha_i)$. Therefore, referring to (A.5b) and (A.6b), the two terms $\hat{\mathrm{E}}[x^M_{s\alpha_i} \mid x(s\alpha_i)]$ and $\hat{\mathrm{E}}[x^M_{s\alpha_i^c} \mid x^n_{s\alpha_i^c}]$ are conditionally uncorrelated when conditioned on $x(s)$. As mentioned, the terms $\tilde{x}^M_{s\alpha_i}$ and $\tilde{x}^M_{s\alpha_i^c}$ are uncorrelated with each other and with x^n and therefore with $x(s) = V_s x^n_s$. Hence, it follows that $x^M_{s\alpha_i}$ and $x^M_{s\alpha_i^c}$ are uncorrelated when conditioned on $x(s)$. Since $s\alpha_i$ was an arbitrary child of s, this holds for all children and the proposition is proved.

Notes

1. While the class of internal models is rich enough to include minimal models in the state-space case, the same is not true for MAR models [22]. Nevertheless, for reasons discussed in the text, we focus on internal MAR models and seek the minimal model within this class.
2. The approach in [10] is only applicable to self-similar processes with stationary increments while that of [22, 24, 26] as well as our approach is completely general.
3. In this paper, we assume without loss of generality that all random quantities are zero-mean.
4. Using, for example, lexicographic ordering of the nodes comprising $\mathcal{T}_s(n)$ or $\mathcal{T}_s^c(n)$ in order to construct a large vector from the component vectors $x(t)$.
5. P_z is our notation for the covariance matrix for random vector z. Similarly, P_{uv} denotes the cross-covariance matrix for random vectors u and v.
6. We may assume Gaussianity without loss of generality because our interest is only in second-order statistics and linear processing.
7. Recall that the MAR estimator has a cubic dependence on state dimension.
8. We use the notation $\hat{\mathrm{E}}[u \mid v]$ to denote the linear least-squares estimation of u given v ($\hat{\mathrm{E}}[\cdot \mid \cdot\cdot]$ is linear in its second argument) while $\mathrm{E}[\cdot \mid \cdot\cdot]$ is the Bayes least-squares estimator which is, in general, non-linear. These, of course, coincide in the jointly Gaussian case. ($\mathrm{E}[z]$ is defined as the expectation of z with respect to the probability measure on z.)
9. For clarity of presentation, we restrict attention to processes with constant state dimension, d. However, our results are applicable and easily generalizable to cases in which the state varies with $s \in \mathcal{S}_0$, as we will show by example.
10. To compute exact sample-paths for random fields we use the FFT techniques described in [13]. Note that this requires $O(N \log N)$ computations while MAR sample-path generation is $O(N)$.

References

1. H. Akaike. Stochastic theory of minimal realizations. *IEEE Transactions on Automatic Control*, AC–19(6):667–674, December 1974.

2. H. Akaike. Markovian representation of stochastic processes by canonical variables. *SIAM Journal of Control*, 13(1):162–173, January 1975.
3. K. Arun and S. Kung. Balanced approximation of stochastic systems. *SIAM Journal of Matrix Analysis and Applications*, 11(1):42–68, January 1990.
4. J. Beran. *Statistics for Long-Memory Processes*. Chapman and Hall, New York, 1994.
5. K. Chou, A. Willsky, and A. Benveniste. Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control*, 39(3):464–478, March 1994.
6. K. Chou, A. Willsky, and R. Nikoukhah. Multiscale systems, Kalman filters, and Riccati equations. *IEEE Transactions on Automatic Control*, 39(3):479–492, March 1994.
7. R. Cowell, P. Dawid, S. Lauritzen, and D. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, NY, 1999.
8. M. Daniel. *Multiresolution statistical modeling with application to modeling groundwater flow*. PhD thesis, Massachusetts Institute of Technology, February 1997.
9. M. Daniel and A. Willsky. A multiresolution methodology for signal-level fusion and data assimilation with applications to remote sensing. *Proceedings of the IEEE*, 85(1):164–180, January 1997.
10. M. Daniel and A. Willsky. The modeling and estimation of statistically self-similar processes in a multiresolution framework. *IEEE Transactions on Information Theory*, 45(3):955–970, April 1999.
11. M. Daniel, A. Willsky, and D. McLaughlin. Travel time estimation using a multiscale stochastic framework. *Advances in Water Resources*, 23(6):571–665, May 2000.
12. K. Daoudi, A. Frakt, and A. Willsky. Multiscale autoregressive models and wavelets. *IEEE Transactions on Information Theory*, 45(3):828–845, April 1999.
13. C. Dietrich and G. Newsam. Fast and exact simulation of stationary Gaussian processes through the circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, July 1997.
14. P. Fieguth, W. Karl, A Willsky, and C. Wunsch. Multiresolution optimal interpolation and statistical analysis of TOPEX/POSEIDON satellite altimetry. *IEEE Transactions on Geoscience and Remote Sensing*, 33(2):280–292, March 1995.
15. P. Fieguth, D. Menemenlis, T. Ho, A. Willsky, and C. Wunsch. Mapping Mediterranean altimeter data with a multiresolution optimal interpolation algorithm. *Journal of Atmospheric and Ocean Technology*, 15:535–546, April 1998.
16. P. Fieguth and A. Willsky. Fractal estimation using models on multiscale trees. *IEEE Transactions on Signal Processing*, 44(5):1297–1300, May 1996.
17. P. Fieguth, A. Willsky, and W. Karl. Efficient multiresolution counterparts to variational methods for surface reconstruction. *Computer Vision and Image Understanding*, 70(2):157–176, May 1998.
18. C. Fosgate, H. Krim, W. Irving, and A. Willsky. Multiscale segmentation and anomaly enhancement of SAR imagery. *IEEE Transactions on Image Processing*, 6(1):7–20, January 1997.
19. A. Frakt. *Internal multiscale autoregressive processes, stochastic realization, and covariance extension*. PhD thesis, Massachusetts Institute of Technology, August 1999.
20. D. Heckerman. *Learning in graphical models*, chapter A tutorial on learning with Bayesian networks. MIT Press, 1999. M. Jordan (Ed).
21. T. Ho, P. Fieguth, and A. Willsky. Multiresolution stochastic models for the efficient solution of large-scale space-time estimation problems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3097–3100, Atlanta, GA, May 1996.
22. W. Irving. *Multiscale stochastic realization and model identification with applications to large-scale estimation problems*. PhD thesis, Massachusetts Institute of Technology, September 1995.
23. W. Irving, P. Fieguth, and A. Willsky. An overlapping tree approach to multiscale stochastic modeling and estimation. *IEEE Transactions on Image Processing*, 6(11), November 1997.
24. W. Irving, W. Karl, and A. Willsky. A theory for multiscale stochastic realization. In *Proceedings of the 33rd IEEE Conference on Decision and Control*, volume 1, pages 655–62, Lake Buena Vista, FL, December 1994.
25. W. Irving, L. Novak, and A. Willsky. A multiresolution approach to discriminating targets from clutter in SAR imagery. *IEEE Transactions on Aerospace and Electronic Systems*, 33(4):1157–1169, October 1997.

26. W. Irving and A. Willsky. A canonical correlations approach to multiscale stochastic realization. *IEEE Transactions on Automatic Control*. Submitted. Preprint available at http://vougeot.mit.edu/ssg.cgi/pubs/pubs.mpl.
27. A. Journel and C. Huijbregts. *Mining Geostatistics*. Academic Press, New York, 1978.
28. A. Kannan. *Adaptation of spectral trajectory models for LVCSR*. PhD thesis, Boston University, 1997.
29. A. Kannan and S. Khudanpur. Tree-structured models of parameter dependence for rapid adaptation in large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, Arizona, 1999.
30. A. Kannan and M. Ostendorf. Modeling dependence in adaptation of acoustic models using multiscale tree processes. In *Proceedings of EUROSPEECH*, pages 1863–1866, 1997.
31. A. Kim and H. Krim. Hierarchical stochastic modeling of SAR imagery for segmentation/compression. *IEEE Transactions on Signal Processing*, 47(2):458–468, February 1999.
32. B. Kosko. *Neural Networks for Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1992.
33. P. Kumar. A multiple scale state-space model for characterizing subgrid scale variability of near-surface soil moisture. *IEEE Transactions on Geoscience and Remote Sensing*, 37(1):182–197, January 1999.
34. S. Lauritzen. *Graphical Models*. Oxford University Press, 1996.
35. A. Lindquist and G. Picci. On the stochastic realization problem. *SIAM Journal on Control and Optimization*, 17(3):365–389, May 1979.
36. M. Luettgen, W. Karl, and A. Willsky. Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Transactions on Image Processing*, 3(1):41–64, January 1994.
37. M. Luettgren, W. Karl, A. Willsky, and R. Tenney. Multiscale representations of Markov random fields. *IEEE Transactions on Signal Processing*, 41(12):3377–3396, December 1993.
38. M. Luettgen and A. Willsky. Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination. *IEEE Transactions on Image Processing*, 4(2):194–207, February 1995.
39. B. Mandelbrot and J. Van Ness. Fractional Brownian motions, fractional noises, and applications. *SIAM Review*, 10:422–437, 1968.
40. R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya, series A*, 26:329–358, 1964.
41. M. Schneider, P. Fieguth, W. Karl, and A. Willsky. Multiscale methods for the segmentation of images. *IEEE Transactions on Image Processing*. To appear.
42. J. Schroeder and D. Howard. Multiscale modeling for target detection in complex synthetic aperture radar. In *Proceedings of the Asilomar Conference on Signals, Systems, and Signal Processing*, November 1998.
43. M. Sironvalle. The random coin method: solution of the problem of simulation of a random function in the plane. *Mathematical Geology*, 12(1):29–36, January 1980.