# Gaussian Multiresolution Models: Exploiting Sparse Markov and Covariance Structure

Myung Jin Choi, *Student Member, IEEE*, Venkat Chandrasekaran, *Student Member, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—In this paper, we consider the problem of learning Gaussian multiresolution (MR) models in which data are only available at the finest scale, and the coarser, hidden variables serve to capture long-distance dependencies. Tree-structured MR models have limited modeling capabilities, as variables at one scale are forced to be uncorrelated with each other conditioned on other scales. We propose a new class of Gaussian MR models in which variables at each scale have *sparse conditional covariance structure* conditioned on other scales. Our goal is to learn a tree-structured graphical model connecting variables across scales (which translates into sparsity in inverse covariance), while at the same time learning sparse structure for the conditional covariance (not its inverse) within each scale conditioned on other scales. This model leads to an efficient, new inference algorithm that is similar to multipole methods in computational physics. We demonstrate the modeling and inference advantages of our approach over methods that use MR tree models and single-scale approximation methods that do not use hidden variables.**

***Index Terms*—Gauss–Markov random fields, graphical models, hidden variables, multipole methods, multiresolution (MR) models.**

## I. Introduction

MULTIRESOLUTION (MR) methods have been widely used in large-scale signal processing applications due to their rich modeling power as well as computational efficiency [34]. Estimation algorithms based on MR representations are efficient since they perform global computations only at coarser scales in which the number of variables is significantly smaller than at finer scales. In addition, MR models provide compact representations for long-range statistical dependencies among far-apart variables by capturing such behavior at coarser resolutions. One of the most common settings [3], [7], [8], [11], [19], [23], [28], [34] for representing MR models is that of *graphical models*, in which the nodes of the graph index random variables and the edges encode conditional independence structure among the variables. Graphical models in which edges are undirected are also called Markov random fields (MRFs).

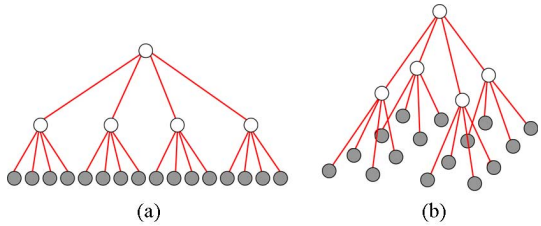


Fig. 1. Examples of MR tree models (a) for a 1-D process and (b) for a 2-D process. Shaded nodes represent original variables at the finest scale and white nodes represent hidden variables at coarser scales.

In many applied fields including communication [13], speech and image processing [32], and bioinformatics [27], statistical models have been represented with sparse graphical model structures in which each node in the graph is connected to only a few other nodes. For Gaussian phenomena, in which the variables being represented are jointly Gaussian, this corresponds to sparsity in the inverse covariance matrix. There are a variety of attractions of such sparse models, including parsimonious parameterization (with obvious advantages for learning such models and avoiding overfitting) and the potential for efficient inference algorithms (e.g., for computing posterior distributions given measurements or for parameter estimation).

The potential advantages of sparsity for efficient inference, however, depend very much on the structure of the resulting graph, with the greatest advantage for tree-structured graphs, i.e., graphs without cycles. Indeed, this advantage provided one of the major motivations for the substantial literature and application [10], [14], [22], [34] of models on MR trees (such as in Fig. 1) in which each level represents the phenomenon of interest at a corresponding scale or resolution. The coarser scales in these models are usually introduced solely or primarily[1] as *hidden variables*. That is, it is the finest scale of such a model that represents the phenomenon of interest, and coarser scales are introduced to capture long-range correlations in a manner that is graphically far more parsimonious than could be captured solely within a single, finest scale model. Indeed, a sparse single-scale graphical model is often poor at capturing long-range correlations, and even if it does, may result in the model being ill-conditioned.

A significant and well-known limitation of such MR tree models, however, is the set of statistical artifacts they can

Manuscript received April 24, 2009; accepted September 22, 2009. First published November 06, 2009; current version published February 10, 2010. This work was supported in part by AFOSR under Grant FA9550-08-1-1080, in part by MURI under AFOSR Grant FA9550-06-1-0324, and in part by Shell International Exploration and Production, Inc. The work of M. J. Choi was supported in part by a Samsung Scholarship. A preliminary version of this work appeared in the Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009), Montreal, QC, Canada. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mark J. Coates.

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2036042

[1]In some contexts, some of the variables at coarser scales represent nonlocal functionals of the finest scale phenomenon that are either measured or are to be estimated.

introduce. In an MR tree model, variables at one scale are *conditionally independent* when conditioned on neighboring scales, a direct consequence of the fact that nodes are connected to each other only through nodes at other scales. Thus, the correlation structure between variables at the finest scale can depend dramatically on exactly how the MR tree is arranged over these finest scale nodes. In particular, finest scale nodes that are the same "distance" from each other as measured solely within that finest scale can have very different distances along the MR tree due to the different lengths of fine-to-coarse-to-fine paths that connect them. While in some applications such fine-scale artifacts may have no significant effect on the particular estimation task of interest, there are many situations in which these artifacts are unacceptable. A variety of methods [3], [7], [8], [11], [19], [23], [28] have been proposed to overcome this limitation of tree models. These methods involve including additional edges—either interscale or within the same scale—to the MR tree model and considering an overall sparse MR graphical model.

In this work, we propose a different approach to address the limitation of MR tree models—one that has considerable intuitive appeal. Note that the role of coarser scales in an MR model is to capture most of the correlations among the finer scale variables through coarser scales. Then, should not the *residual* correlations at each scale that need to be captured be approximately *local*? In other words, conditioned on variables at other scales, the residual correlation of any node should be concentrated on a small number of neighboring nodes within the same scale. This suggests that instead of assuming that the conditional statistics at each scale (conditioned on the neighboring scales) have sparse graphical structure (i.e., sparse inverse covariance) as in the previous methods, we need to look for models in which the *conditional* statistics have *sparse covariance structure*.

MR models with the type of structure described above—tree-structure between scales and then sparse conditional covariance structure within each scale—have a special inverse covariance structure. As we describe later in the paper, the inverse covariance matrix of our MR model (denoted $J$) can be represented as a sum of the inverse covariance matrix of an MR tree (denoted $J^h$) and inverse of a conditional covariance matrix within each scale (denoted $\Sigma^c$), i.e., $J = J^h + (\Sigma^c)^{-1}$ where both $J^h$ and $\Sigma^c$ are sparse matrices. This structure leads to efficient estimation algorithms that are different in a fundamental way from standard graphical model estimation algorithms which exploit sparse graph structure. Indeed, as we describe in this paper, sparse in-scale conditional correlation structure generally corresponds to a *dense* graphical model within each scale, so that standard graphical model inference algorithms are not useful. However, estimation for phenomena that are only locally correlated requires local computations—essentially a generalization of finite impulse response (FIR) filtering within each scale—corresponding to multiplication involving the sparse conditional covariance matrix. Our approach can be viewed as a statistical counterpart to so-called *multipole methods* [20] for the rapid solution of elliptic partial differential equations (in particular those corresponding to evaluating electric fields given charge distributions); we use the sparse tree structure of *part* of the overall statistical structure, namely, that *between* scales, to propagate information from scale-to-scale (exploiting sparsity in $J^h$), and then perform local FIR-like residual filtering *within* each scale (exploiting sparsity in $\Sigma^c$).

In addition to developing efficient algorithms for inference given our MR model, we develop in detail methods for *learning* such models given data at the finest scale (or more precisely an empirical marginal covariance structure at the finest scale). Our modeling procedure proceeds as follows: given a collection of variables and a desired covariance among these variables, we construct an MR model by introducing hidden variables at coarser resolutions. Then, we optimize the structure of each scale in the MR model to approximate the given statistics with a *sparse conditional covariance structure* within each scale. This step can be formulated as a convex optimization problem involving the log-determinant of the conditional covariance matrix.

The rest of the paper is organized as follows. In the next section, we provide some background on graphical models and a sparse matrix approximation method using log-determinant maximization. In Section III, the desired structure of our MR model—sparse interscale graphical structure and sparse in-scale conditional covariance structure—is specified in detail. The special-purpose inference algorithm that exploits sparsity in both Markov and covariance structure is described in Section IV, while in Section V, we show how the log-det maximization problem can be used to learn our MR models. In Section VI, we illustrate the advantages of our framework in three modeling problems: dependencies in monthly stock returns, fractional Brownian motion [30], and a 2-D field with polynomially decaying correlations. We provide experimental evidence that our MR model captures long-range correlations well without blocky artifacts, while using many fewer parameters than single-scale approximations. We also demonstrate that our MR approach provides improved inference performance. Section VII concludes this paper, and in Appendixes I–III, we provide algorithmic details for our learning method.

## II. Preliminaries

### A. Gaussian Graphical Models

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a graph with a set of nodes $\mathcal{V}$ and (pairwise) edges $\mathcal{E}$. Two nodes $s$ and $t$ are said to be *neighbors* if there is an edge $\{s,t\}$ between them. A subset of nodes $S \subset \mathcal{V}$ is said to *separate* subsets $A, B \subset \mathcal{V}$ if every path in $\mathcal{G}$ between any node in $A$ and any node in $B$ passes through a node in $S$. A graphical model is a collection of random variables indexed by nodes of the graph: each node $s$ is associated with a random variable $x_s$,[2] and for any $A \in \mathcal{V}, x_A \equiv \{x_s | s \in A\}$. A probability distribution is said to be *Markov* with respect to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if for any subsets $A, B \subset \mathcal{V}$ that are separated by some $S \in \mathcal{V}, x_A$ and $x_B$ are independent conditioned on $x_S : p(x_A, x_B | x_S) = p(x_A | x_S) p(x_B | x_S)$. Specifically, if an edge is not present between two random variables, it indicates that the two variables are independent conditioned on all other variables in the graph.

[2]For simplicity, we assume that $x_s$ is a scalar variable, but any of the analysis in this paper can be easily generalized to the case when $x_s$ is a random vector.

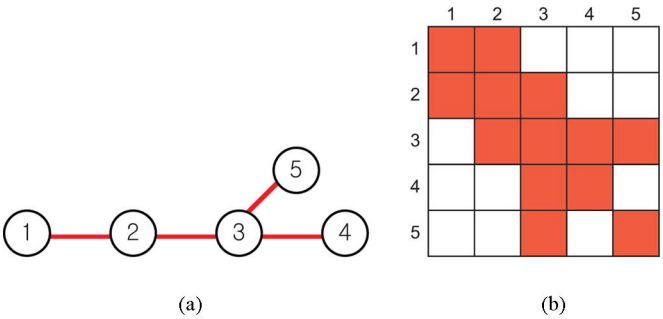


Fig. 2. (a) Sparse graphical model. (b) Sparsity pattern of the corresponding information matrix.

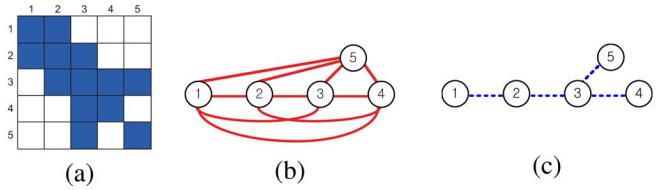


Fig. 3. Conjugate graph. (a) Sparsity pattern of a covariance matrix. (b) Corresponding graphical model. (c) Conjugate graph encoding the sparsity structure of the covariance matrix in (a).

Let $x \sim \mathcal{N}(\mu, \Sigma)$ be a jointly Gaussian random vector with a mean vector $\mu$ and a positive–definite covariance matrix $\Sigma$. If the variables $x$ are Markov with respect to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the inverse of the covariance matrix $J = \Sigma^{-1}$ (also called the information, or precision, or concentration matrix) is sparse with respect to $\mathcal{G}$ [26]. That is, $J_{s,t} \neq 0$ if and only if $\{s,t\} \in \mathcal{E}$. We use $\mathcal{N}(J^{-1}h, J^{-1})$ to denote a Gaussian distribution with an information matrix $J$ and a potential vector $h \equiv \Sigma^{-1}\mu$; the distribution has the form $p(x) \propto \exp(-(1/2)x^T Jx + hx)$. Fig. 2(a) shows one example of a sparse graph, and the sparsity pattern of the corresponding information matrix $J$ is shown in Fig. 2(b). The graph structure implies that $x_1$ is uncorrelated with $x_5$ *conditioned* on $x_2$. Note that this does not indicate that $x_1$ is uncorrelated with $x_5$. In fact, the covariance matrix (the inverse of the information matrix) will, in general, be a full matrix.

For any subset $A \subset \mathcal{V}$, let $\backslash A \equiv \{s \in \mathcal{V}, s \notin A\}$ be its complement. Then, the conditional distribution $p(x_A|x_{\backslash A})$ is Markov with respect to the induced subgraph of $\mathcal{G}$ with nodes $A$ and edges $\mathcal{E}_A = \{\{s,t\}|\{s,t\} \in \mathcal{E}, s,t \in A\}$. The corresponding information matrix of the conditional model is the *submatrix* of $J$ with rows and columns corresponding to elements in $A$. For example, in Fig. 2, $p(x_1, x_2, x_3, x_4|x_5)$ is a chain model connecting variables $x_1$ through $x_4$, and the information matrix of this conditional distribution is the submatrix $J(1:4, 1:4)$, which is a tri-diagonal matrix.

### B. Conjugate Graphs

While Gaussian graphical models provide a compact representation for distributions with a sparse information matrix, in general, a sparse graphical model cannot represent distributions with a sparse *covariance* matrix. Consider a distribution with the sparsity pattern of the *covariance matrix* given as in Fig. 3(a). Its information matrix will, in general, be a full matrix, and the corresponding graphical model will be fully connected as shown in Fig. 3(b). Therefore, we introduce *conjugate graphs* to illustrate the sparsity structure of a covariance matrix. Specifically, in the conjugate graph, when two nodes are not connected with a *conjugate* edge, they are *uncorrelated* with each other.[3] We use solid lines to display graphical model edges, and dotted lines to represent conjugate edges. Fig. 3(c) shows the corresponding conjugate graph for a distribution with covariance structure as in Fig. 3(a). From the conjugate edge structure, we can identify that $x_1$ is uncorrelated with $x_3, x_4$, and $x_5$.

The term conjugate graph is motivated by the notion of conjugate processes [25]—two random processes that have covariances that are inverses of one another.[4] Our concept of a conjugate graph that represents marginal independence structure is also called a *covariance graph* or a *bi-directed graph* [12], [16], [24].

### C. Log-Determinant Maximization

In this section, we introduce the log-determinant maximization problem to obtain a positive–definite matrix that approximates a given target matrix and has a sparse inverse. This technique will be used in Section V to learn a sparse graphical model approximation or a sparse covariance matrix approximation. Suppose that we are given a target matrix $A^*$, and we wish to learn an approximation $\hat{A}$ that is positive–definite and has a sparse inverse. Thresholding the elements of $(A^*)^{-1}$ can be ineffective as the resulting matrix may not be positive–definite. One alternative is to solve the following convex optimization problem of maximizing the log-determinant of $\hat{A}$ subject to elementwise constraints with respect to the target matrix:

$$\begin{aligned} \hat{A} = \underset{A \succ 0}{\operatorname{argmax}} \quad & \log\det A \\ \text{s.t.} \quad & d(A_{i,j}, A^*_{i,j}) \leq \gamma_{ij} \qquad \forall i,j \end{aligned} \tag{1}$$

where $\gamma_{ij}$ is a nonnegative regularization parameter and $d(\cdot,\cdot)$ is a convex distance function. In Section V, we use the absolute difference between the two values as the distance function: $d(A_{i,j}, A^*_{i,j}) = |A_{i,j} - A^*_{i,j}|$. Note that this optimization problem is convex in $A$. In the following proposition, we show that when $\gamma$ is large enough, a set of elements of the inverse of $\hat{A}$ are forced to be zero.

*Proposition 1:* Assume that $\gamma_{ij} > 0$ for all $i,j$ and that the feasible set of (1) is nonempty. Then, for each $i,j$ such that the inequality constraint is not tight [i.e., $d(\hat{A}_{i,j}, A^*_{i,j}) < \gamma_{ij}$], the corresponding element of $\hat{A}^{-1}$ is zero [i.e., $(\hat{A}^{-1})_{i,j} = 0$].

*Proof:* From the Karush–Kuhn–Tucker (KKT) conditions [4], there exists $\lambda_{ij} \geq 0$ for all $i,j$ such that the following equations are satisfied:

$$\begin{aligned} \lambda_{ij}(d(\hat{A}_{i,j}, A^*_{i,j}) - \gamma_{ij}) &= 0 \\ -\hat{A}^{-1} + W &= 0 \end{aligned}$$

where $W$ is a matrix with its elements $W_{i,j} = \lambda_{ij}\nabla d(\hat{A}_{i,j}, A^*_{i,j})$. The first equation is also called the complementary slackness condition. The second equation is obtained using $(\partial \log\det A)/(\partial A) = A^{-1}$. For all $\{i,j\}$ such that $d(\hat{A}_{i,j}, A^*_{i,j}) < \gamma_{ij}$, we get $\lambda_{ij} = 0$ from the first equation. Since $\hat{A}^{-1} = W$ from the second equation, for each $\{i,j\}$ that the equality constraint is not tight, $(\hat{A}^{-1})_{i,j} = 0$. □

[3]Since we consider jointly Gaussian variables, uncorrelated variables are independent.

[4]This is different from the widely known conjugate priors [2].

This optimization problem is commonly used in Gaussian modeling to learn a sparse graphical model approximation given the target covariance [1] as we describe in Section V-A. We also use the same framework to learn a *sparse covariance matrix approximation* given the target information matrix as described in Section V-B.

## III. Multiresolution Models With Sparse In-Scale Conditional Covariance

We propose a class of MR models with tree-structured connections between different scales and sparse conditional covariance structure at each scale. Specifically, within each scale, a variable is correlated with only a few other variables in the same scale *conditioned* on variables at scales above and below. We illustrate the sparsity of the *in-scale conditional covariance* using the conjugate graph. Thus, our model has a sparse graphical model for interscale structure and a sparse conjugate graph for in-scale structure. In the rest of the paper, we refer to such an MR model as a sparse in-scale conditional covariance multiresolution (SIM) model.

We would like to emphasize the difference between the concept of *in-scale conditional covariance* with the more commonly used concepts of *marginal covariance* and *pairwise conditional covariance*. Specifically, marginal covariance between two variables is the covariance without conditioning on any other variables. Pairwise conditional covariance refers to the conditional covariance between two variables when conditioned on *all other variables*, including the variables within the same scale. In-scale conditional covariance is the conditional covariance between two variables (in the same scale) when conditioned on *variables at other scales* (or equivalently, variables at scales above and below, but not the variables at the same scale).

As we illustrate subsequently in this section, the distinction between SIM models and the class of MR models with sparse pairwise conditional covariance structure is significant in terms of both covariance/information matrix structure and graphical model representation. The latter, which has been the subject of study in previous work of several authors, has sparse information matrix structure and, corresponding to this, sparse structure as a graphical model, including within each scale. In contrast, our SIM models have sparse graphical model structure *between* scales but generally have *dense conditional information matrices* within each scale. At first this might seem to be undesirable, but the key is that the conditional *covariance* matrices within each scale *are* sparse—something we display graphically using conjugate graphs. As we show in subsequent sections, this leads both to advantages in modeling power and efficient inference.

Fig. 4(b) shows an example of our SIM model. We denote the coarsest resolution as *scale 1* and increase the scale number as we go to finer scales. In the model illustrated in Fig. 4(b), *conditioned* on scale 1 (variable $x_1$) and scale 3 (variables $x_5$ through $x_{10}$), $x_2$ is *uncorrelated* with $x_4$. Note that this is different from $x_2$ and $x_4$ being uncorrelated without conditioning on other scales (the marginal covariance is nonzero), and also different from the corresponding element in the information matrix being zero (the pairwise conditional covariance is nonzero). In fact, the corresponding graphical model representation of the model in Fig. 4(b) consists of a densely connected graphical structure within each scale as shown in Fig. 4(c).

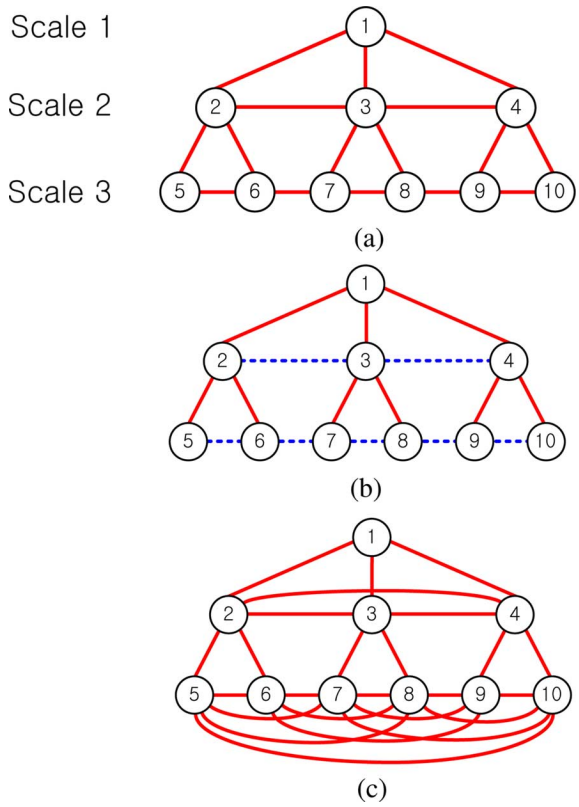


Fig. 4. Examples of MR models. (a) MR model with a sparse graphical structure. (b) SIM model with sparse conjugate graph within each scale. (c) Graphical model corresponding to the model in (b).

In contrast, an MR model with a sparse graphical model structure within each scale is shown in Fig. 4(a).[5] Such a model does not enforce sparse covariance structure within each scale conditioned on other scales: conditioned on scales above and below, $x_2$ and $x_4$ are correlated unless we condition on the other variables at the same scale (namely variable $x_3$). In Section VI, we demonstrate that SIM models lead to better modeling capabilities and faster inference than MR models with sparse graphical structure.

The SIM model, to our best knowledge, is the first approach to enforce sparse conditional covariance at each scale explicitly in MR modeling. A majority of the previous approaches to overcoming the limitations of tree models [7], [8], [11], [23], [28] focus on constructing an overall sparse graphical model structure [as in Fig. 4(a)] to enable an efficient inference procedure. A different approach based on a directed hierarchy of densely connected graphical models is proposed in [32], but it does not have a sparse conjugate graph at each layer and requires mean-field approximations unlike our SIM model.

### A. Desired Structure of the Information Matrix

A SIM model consists of a sparse interscale graphical model connecting different scales and a sparse in-scale conditional co-

[5]Throughout this paper, we use the term "sparse" loosely for coarser scales with just a few nodes. For these coarse scales, we have a small enough number of variables so that computation is not a problem even if the structure is not sparse.

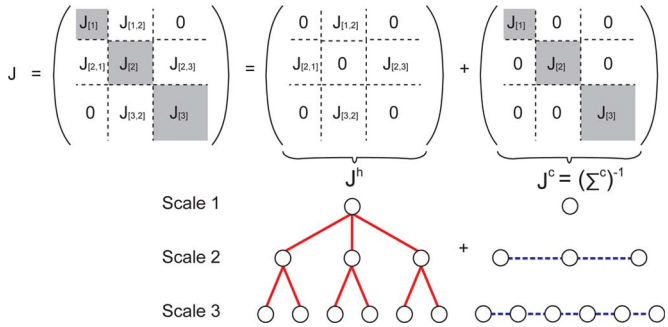


Fig. 5. Decomposition of a SIM model into a sparse hierarchical structure connecting different scales and a sparse conjugate graph at each scale. Shaded matrices are dense and nonshaded matrices are sparse.

variance matrix at each scale. Here, we specify the desired sparsity structure for each submatrix of the information matrix of a SIM model. First, we partition the information matrix $J$ of a SIM model by scale as shown in Fig. 5 (corresponding to a model with three scales). The submatrix $J_{[m]}$ of $J$ corresponds to the information matrix of the *conditional* distribution at scale $m$ conditioned on other scales (see Section II-A). As illustrated in Fig. 4(c), a SIM model has a densely connected graphical model within each scale, so $J_{[m]}$ in general is not a sparse matrix. The inverse of $J_{[m]}$, however, is sparse since we have a sparse conditional covariance matrix within each scale. The submatrix $J_{[m_1,m_2]}$ is sparse with respect to the graphical model structure connecting scale $m_1$ and $m_2$. We consider hierarchical models in which only successive neighboring scales are connected. Hence, $J_{[m_1,m_2]}$ is a zero matrix if $|m_1 - m_2| > 1$. By the modeling assumption that the interscale graphical model connecting different scales is sparse, $J_{[m,m+1]} = J^T_{[m+1,m]}$ is a sparse matrix. In Fig. 5, shaded matrices are dense and nonshaded matrices are sparse.

The matrix $J$ can be decomposed as a sum of $J^h$, corresponding to the hierarchical interscale tree structure, and $J^c$, corresponding to the conditional in-scale structure. Let $\Sigma^c \equiv (J^c)^{-1}$. Since $J^c$ is a block-diagonal matrix, its inverse $\Sigma^c$ is also block-diagonal with each diagonal block equal to $(J_{[m]})^{-1}$. Hence, $\Sigma^c$ is a sparse matrix, whereas $J^c$ is not sparse in general. Therefore, the information matrix $J$ of a SIM model can be decomposed as a sum of a sparse matrix and the inverse of a sparse block-diagonal matrix

$$J = J^h + (\Sigma^c)^{-1}. \tag{2}$$

Each nonzero entry in $J^h$ corresponds to an interscale edge connecting a pair of variables at different scales. The block diagonal matrix $\Sigma^c$ has nonzero entries corresponding to *conjugate* edges within each scale. One simple example is demonstrated in Fig. 5. In Section IV, we take advantage of sparsity in *both* $J^h$ and $\Sigma^c$ for efficient inference.

## IV. Inference Exploiting Sparsity in Markov and Covariance Structure

Let $x$ be a collection of random variables with a prior distribution: $x \sim \mathcal{N}(0, J^{-1})$. Suppose that we have a set of noisy measurements at a subset of the variables: $y = Cx + v$ where $C$ is a selection matrix that only selects variables at which measurements are specified, and $v$ is a zero-mean Gaussian noise vector with covariance $R$. The maximum *a posteriori* (MAP) estimate $\hat{x}$ is equivalent to the mean of the posterior distribution

$$\hat{x} = \underset{x}{\operatorname{argmax}}\, p(x|y) = \mathbb{E}[x|y] = (J + J^p)^{-1}h \tag{3}$$

where $J^p \equiv C^T R^{-1} C$, and $h \equiv C^T R^{-1} y$. The posterior information matrix $(J + J^p)$ has the same sparsity structure as $J$ if we assume that the noise covariance matrix $R$ is diagonal. If $J$ corresponds to a tree-structured model, (3) can be solved with linear complexity. If the prior model is not a tree, solving this equation directly by matrix inversion requires $\mathcal{O}(N^3)$ computations where $N$ is the number of variables. We review a class of iterative algorithms that solve linear systems using the idea of a *matrix splitting* in Section IV-A. Based on the specific splitting of the information matrix of our SIM model as in (2), we propose a new and efficient inference algorithm in Section IV-B.

### A. Iterative Algorithms Based on a Matrix Splitting

As described above, computing the optimal estimates in Gaussian models is equivalent to solving a linear equation $A\hat{x} = h$ where $A \equiv (J + J^p)$ is a posterior information matrix. Many iterative linear system solvers are based on the idea of a matrix splitting: $A = M - K$. Let us rewrite the original equation as $M\hat{x} = h + K\hat{x}$. Assuming that $M$ is invertible, we obtain the following iterative update equations:

$$\hat{x}^{\text{new}} = M^{-1}(h + K\hat{x}^{\text{old}}) \tag{4}$$

where $\hat{x}^{\text{old}}$ is the value of $\hat{x}$ at the previous iteration, and $\hat{x}^{\text{new}}$ is the updated value at the current iteration. The matrix $M$ is called a *preconditioner*, and (4) corresponds to the preconditioned Richardson iterations [18]. If solving the equation $M\hat{x} = z$ for a fixed vector $z$ is easy due to a special structure of $M$, each iteration can be performed efficiently.[6] There are a variety of ways in which splittings can be defined [15]. For example, Gauss–Jacobi iterations set the preconditioner $M$ as a diagonal matrix with diagonal elements of $A$, and embedded tree (ET) algorithms [33] split the matrix so that $M$ has a tree structure.

### B. Efficient Inference in SIM Models

We use the matrix splitting idea in developing an efficient inference method for our SIM model. Recall that the information matrix of the SIM model can be decomposed as in (2). Our goal is to solve the equation $(J^h + (\Sigma^c)^{-1} + J^p)\hat{x} = h$ where $J^h, \Sigma^c$, and $J^p$ are all sparse matrices. We alternate between two inference steps corresponding to *interscale* computation and *in-scale* computation in the MR model. Our interscale computation, called the *tree inference step* exploits sparse Markov structure connecting different scales, while our *in-scale inference step* exploits sparse in-scale conditional covariance structure within each scale.

*1) Tree Inference:* In the tree-inference step, we select the interscale tree structure as the preconditioner in (4) by setting

[6]We may use different preconditioners for each iteration, resulting in nonstationary Richardson iterations [6].

$M = J^h + J^p + D$, where $D$ is a diagonal matrix added to ensure that $M$ is positive–definite[7]

$$(J^h + J^p + D)\hat{x}^{\text{new}} = h - (\Sigma^c)^{-1}\hat{x}^{\text{old}} + D\hat{x}^{\text{old}}. \quad (5)$$

With the right-hand side vector fixed, solving the above equation is efficient since $M$ corresponds to a tree-structured graphical model.[8] On the right-hand side, $D\hat{x}$ can be evaluated easily since $D$ is diagonal, but computing $z \equiv (\Sigma^c)^{-1}\hat{x}$ directly is not efficient because $(\Sigma^c)^{-1}$ is a dense matrix. Instead, we evaluate $z$ by *solving the matrix equation* $\Sigma^c z = \hat{x}$. The matrix $\Sigma^c$ (in-scale conditional covariance) is sparse and well-conditioned in general; hence the equation can be solved efficiently. In our experiments, we use just a few Gauss–Jacobi iterations (see Section IV-A) to compute $z$.

*2) In-scale Inference:* In this step, we select the in-scale structure to perform computations within each scale by setting $M = (\Sigma^c)^{-1}$. Then, we obtain the following update equation:

$$\hat{x}^{\text{new}} = \Sigma^c(h - J^h\hat{x}^{\text{old}} - J^p\hat{x}^{\text{old}}). \quad (6)$$

Evaluating the right-hand side only involves multiplications of a sparse matrix $(\Sigma^c)$ and a vector, so $\hat{x}^{\text{new}}$ can be computed efficiently. Note that although we use a similar method of splitting the information matrix and iteratively updating $\hat{x}$ as in the Richardson iteration (4), our algorithm is efficient due to a fundamentally different reason. In the Richardson iteration (specifically, the ET algorithm) and in the tree-inference step, solving the matrix equation is efficient because it is equivalent to solving an inference problem on a tree model. In our in-scale inference step, the preconditioner selected actually corresponds to a densely connected graphical model, but since it has a sparse conjugate graph, the update equation reduces to a sparse matrix multiplication. Thus, our in-scale inference step requires only *local* computations, which is in the same spirit as multipole methods [20] or FIR filtering methods.

After each iteration, the algorithm checks whether the procedure has converged by computing the relative residual error: $\epsilon \equiv (\| h - A\hat{x} \|_2)/(\| h \|_2)$ where $\| \cdot \|_2$ is the $\ell_2$ norm and $A = J^h + (\Sigma^c)^{-1} + J^p$. The term $A\hat{x}$ can be evaluated efficiently even though $A$ is not a sparse matrix. Since $A\hat{x} = J^h\hat{x} + J^p\hat{x} + (\Sigma^c)^{-1}\hat{x}$, the value of $z = (\Sigma^c)^{-1}\hat{x}$ computed from the tree-inference step can be used to evaluate the residual error as well, and since $J^h$ and $J^p$ are sparse matrices, the first two terms can be computed efficiently.

The concept of performing local in-scale computations can be found in algorithms that use multiple scales to solve partial differential equations, such as multipole methods [20] or multigrid methods [5]. The efficiency of these approaches comes from the assumption that after a solution is computed at coarser resolutions, only *local* terms need to be modified at finer resolutions. However, these approaches do not have any statistical basis or interpretation. The models and methods presented in this paper are aimed at providing a precise statistical framework leading to inference algorithms with very solid advantages analogous to those of multipole and multigrid methods.

[7]In (4), $M$ needs to be invertible, but $(J^h + J^p)$ is singular since the diagonal elements at coarser scales (without measurements) are zero. In our experiments, we use $D = (\text{diag}(\Sigma^c))^{-1}$ where $\text{diag}(\Sigma^c)$ is a diagonal matrix with diagonal elements of $\Sigma^c$.

[8]This step is efficient for a more general model as well in which the interscale structure is sparse but not a tree.

## V. Learning MR Models With Sparse In-Scale Conditional Covariance

In this section, we describe the procedure of learning a SIM model approximation to a given target covariance. As has been well-developed in the literature and reviewed in Section V-A, optimization of the log-determinant of a covariance matrix leads to sparse inverse covariances and hence sparse graphical models. In Section V-B, we turn the tables—optimizing the log-determinant of the *inverse* covariance to yield a sparse covariance. We learn SIM models with sparse hierarchical graphical structure and sparse in-scale conditional covariance structure by combining these two methods as described in Section V-C.

### A. Sparse Graphical Model Approximation

Suppose that we are given a target covariance $\Sigma^*$ and wish to learn a sparse graphical model that best approximates the covariance. The target covariance matrix may be specified exactly when the desired statistics of the random process are known, or may be the empirical covariance computed from samples. One possible solution for selecting a graphical model is to use the inverse of the target covariance matrix, $(\Sigma^*)^{-1}$. However, whether $\Sigma^*$ is exact or empirical, its inverse will, in general, be a full matrix, resulting in a fully connected graphical model. One may threshold each element of $(\Sigma^*)^{-1}$ so that small values are forced to zero, but often, this results in an invalid covariance matrix that is not positive–definite.

Therefore, standard approaches in Gaussian graphical model selection [1], [17], [21] use the log-determinant problem in (1) to find an approximate covariance matrix

$$\hat{\Sigma} = \underset{\Sigma \succ 0}{\text{argmax}} \quad \log\det\Sigma \qquad \text{s.t.} \quad \left|\Sigma_{i,j} - \Sigma^*_{i,j}\right| \le \gamma \quad \forall i,j. \quad (7)$$

From Proposition 1, the solution of the above problem has a sparse inverse, which is a sparse graphical model approximation. The entropy of a Gaussian distribution is proportional to the log-determinant of its covariance matrix. Hence, this learning approach is also called maximum-entropy modeling [21].

It can be shown that the dual problem of (7) is given as follows [1]:

$$\hat{J} = \underset{J \succ 0}{\text{argmin}} \quad D(p(x)\|p^*(x)) + \frac{\gamma}{2}\sum_{i,j}|J_{i,j}| \quad (8)$$

where $p(x) \sim \mathcal{N}(0, J^{-1}), p^*(x) \sim \mathcal{N}(0, \Sigma^*)$, and $D(p(x)\|p^*(x)) \equiv \mathbb{E}_p[\log(p(x)/p^*(x))]$ is the divergence between the two distributions. This problem minimizes the divergence between the approximate and the original distribution with an $\ell_1$ penalty on the elements of $J$ to obtain a sparse graphical model approximation. Both the primal (7) and the dual (8) optimization problems are convex and can be solved efficiently using interior-point methods [21], block coordinate descent methods [1], or the so-called graphical lasso [17].

### B. Sparse Covariance Approximation

We now consider the problem of approximating a target distribution with a distribution that has a sparse *covariance* matrix (as opposed to a sparse information matrix as in the previous section). That is, we wish to approximate a target Gaussian distribution with information matrix $J^*$ by a distribution in which many pairs of the variables are uncorrelated. We again use the log-determinant problem in (1), but now in the information matrix domain

$$\begin{aligned} \hat{J} = \underset{J \succ 0}{\operatorname{argmax}} \quad & \log\det J \\ \text{s.t.} \quad & |J_{i,j} - J^*_{i,j}| \leq \gamma \qquad \forall i,j. \end{aligned} \tag{9}$$

The solution $\hat{J}$ has a sparse inverse, leading to a sparse covariance approximation. Note the symmetry between (7) and (9). In a Gaussian model, the *log-partition function* [7] is proportional to the negative of the log-determinant of the information matrix. Thus, the problem in (9) can be interpreted as *minimizing the log-partition function*.

In our MR modeling approach, we apply this sparse covariance approximation method to model distributions at each scale conditioned on other scales. Thus, the conditional distribution at each scale is modeled as a Gaussian distribution with a sparse covariance matrix.

### C. Learning a SIM Model

In this section, we discuss a method to learn a SIM model to approximate a specified MR model that has some complex structure (e.g., without the local in-scale conditional covariance structure). When a target covariance (or graphical model) is specified only for the finest scale variables, we first need to construct a full MR model that serves as the target model for the SIM approximation algorithm; such an "exact" target MR model must have the property that the marginal covariance at the finest scale equals the specified covariance for the finest scale variables.

Appendix I describes in detail the algorithm that we use to produce a target MR information matrix $J^*$ if we are only provided with a target covariance at the finest scale. The basic idea behind this approach is relatively simple. First, we use an EM algorithm to fit an MR tree model so that the *marginal* covariance at each finest scale node in this model matches those of the provided finest scale target covariance. As is well known, because of the tree structure of this MR model, there are often artifacts across finest scale tree boundaries, a manifestation of the fact that such a model does not generally match the *joint statistics*, i.e., the cross covariances, across different finest scale nodes. Thus, we must correct the statistics at each scale of our MR model in order to achieve this finest scale matching. Therefore, in our second step, we introduce correlations within each scale resulting in a full target $J^*$ whose finest scale marginal covariance matches the originally given covariance. Referring to Fig. 5, what the first tree construction does is to build the tree-structured information matrix $J^{h*}$, capturing interscale connections, as well as a first approximation to the *diagonal* of the in-scale conditional covariance $J^{c*}$. What the second step does is to fill in the remainder of the shaded blocks in $J^{c*}$ and modify the diagonals in order to match the finest scale marginal statistics. In so doing, this target covariance does not, in general, have sparse in-scale conditional covariance (i.e., $\Sigma^{c*} \equiv (J^{c*})^{-1}$ is *not* sparse), and the procedure we now describe (with many more details in Appendixes II and III) takes the target $J^* = J^{h*} + J^{c*}$ and produces an approximation that has our desired SIM structure.

Suppose that the target MR model is specified in information form with information matrix $J^*$. We can find a SIM model that approximates $J^*$ by solving the following optimization problem:

$$\begin{aligned} \hat{J} = \underset{J \succ 0}{\operatorname{argmax}} \quad & \sum_m \log\det J_{[m]} + \sum_{\{i,j\} \in \mathcal{E}_{\text{inter}}} |J_{i,j}| \\ \text{s.t.} \quad & |J_{i,j} - J^*_{i,j}| \leq \gamma \qquad \forall i,j \end{aligned} \tag{10}$$

where $J_{[m]}$ is the in-scale information matrix at scale $m$ and $\mathcal{E}_{\text{inter}}$ is the set of all possible interscale edges connecting successive neighboring scales. Note that except for the positive–definiteness condition $J \succ 0$, the objective function as well as the constraints can be decomposed into an interscale component and in-scale components. If we only look at the terms involving the parameters at scale $m$ (i.e., elements of the matrix $J_{[m]}$), the above problem maximizes the log-determinant of the information matrix $J_{[m]}$ subject to elementwise constraints. Therefore, from the arguments in Section V-B, the log-det terms ensure that each $\hat{J}_{[m]}$ has a sparse inverse, which leads to a sparse in-scale conditional covariance, and thus a sparse conjugate graph. The $\ell_1$-norm on the interscale edges penalizes nonzero elements [performing the same role as in the second term of (8)] and thus encourages the interscale structure connecting different scales to be sparse. Often, the specified target information matrix $J^*$ of the MR model already has a sparse interscale graphical structure, such as an MR tree structure (see Appendix I, for example). In such a scenario, the $\ell_1$-norm can be dropped from the objective function.

The problem in (10) is convex and can be efficiently solved using general techniques for convex optimization [4], [29]. In Appendixes II and III, we provide a simplified version of the problem in (10) to further reduce the computational complexity in solving the optimization problem. This can be achieved by interleaving the procedure of constructing the target MR model and the optimization procedure at each scale to obtain a sparse conjugate graph structure scale-by-scale. The regularization parameter $\gamma$ in the constraints of (10) provides a tradeoff between sparsity of the in-scale conjugate graphs and data-fidelity (i.e., how close the approximation $\hat{J}$ is to the target information matrix $J^*$). In practice, we allow two different regularization parameters for each scale: one for all node constraints and one for all edge constraints. For our experimental results, we selected these regularization parameters using a heuristic method described in Appendix III.

## VI. Experimental Results

Modeling of complex phenomena is typically done with an eye to at least two key objectives: 1) model accuracy; and 2) tractability of the resulting model in terms of its use for various statistical inference tasks.

In this section, we compare the performance of our SIM model to four other modeling approaches. First, we consider

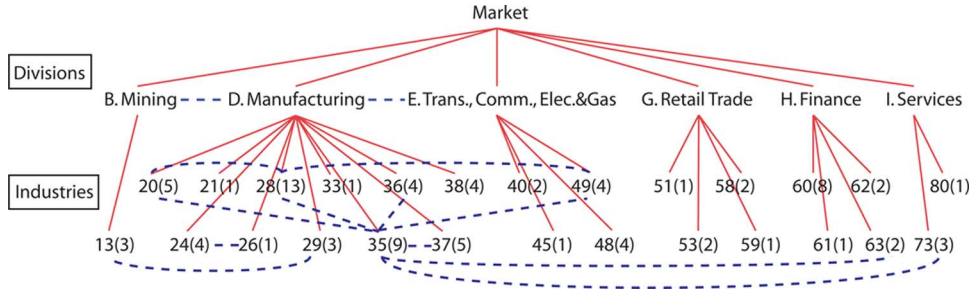


Fig. 6. Structure of the SIM model approximation for stock data.

TABLE I
TOP FOUR STRONGEST CONJUGATE EDGES AT SCALE 3 OF FIG. 6

| Sign | SIC code | Industry Group | Representative Companies |
|---|---|---|---|
| + | 13 | Oil and Gas Extraction | Schlumberger |
| | 29 | Petroleum Refining | Exxon Mobile, Chevron |
| + | 35 | Machinery And Computer Equipment | Dell, Apple, IBM, Xerox |
| | 36 | Other Electrical Equipment Except Computer Equipment | TI, Intel, GE |
| + | 20 | Food And Kindred Products | Coca Cola, Heinz |
| | 28 | Chemicals And Allied Products | Dow Chemical, Johnson & Johnson |
| + | 35 | Machinery And Computer Equipment | Dell, Apple, IBM, Xerox |
| | 73 | Business Services | Microsoft, Oracle |

a single-scale approximate model where we learn a sparse graphical model using (7) without introducing hidden variables. This is motivated by the fact that one of the dominant themes in statistical modeling is to encourage a sparse graphical model structure to approximate given statistics. Another widely used modeling method is a tree-structured MR model. Such tree models are the absolute winner in terms of computational tractability, but they are not nearly as good in terms of modeling accuracy. Third, we consider a sparse MR model in the form introduced in [7], which aims to overcome the limitations of the tree. Note that unlike a SIM model, a sparse MR model has a sparse information matrix but *not* sparse in-scale conditional covariance. Finally, for each of our examples, we have the original model defined by the exact given statistics. They serve as target statistics for each approximate modeling method, but they do not have a sparse structure that makes inference computationally tractable in larger examples.

We measure the modeling accuracy of approximate models by computing the divergence between the exact distribution and the approximate distribution.[9] The tractability of each model can be evaluated either by measuring computation time for a specific inference task or by counting the number of parameters. An important point here is that all of the methods to which we compare, as well as our SIM model, are general-purpose modeling frameworks that are not tailored or tuned to any specific application.

### A. Stock Returns

Our first experiment is modeling the dependency structure of monthly stock returns of 84 companies in the S&P 100 stock index.[10] We use the hierarchy defined by the Standard Industrial Classification (SIC) system,[11] which is widely used in finance, and compute the empirical covariance using the monthly returns from 1990 to 2007. Our MR models have four scales, representing the market, six divisions, 26 industries, and 84 individual companies, respectively, at scales from the coarsest to the finest.

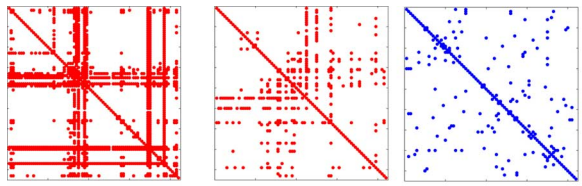


Fig. 7. Stock returns modeling example. Sparsity pattern of the information matrix of (a) the single-scale (122.48), and (b) the sparse MR approximation (28.34). (c) Sparsity pattern of the in-scale conditional covariance of the SIM approximation (16.36). All at the finest scale. We provide the divergence between the approximate and the empirical distribution in the parenthesis. The tree approximation has divergence 38.22.

Fig. 6 shows the first three scales of the SIM model approximation. At scale 3, we show the SIC code for each industry (represented by two digits) and in the parenthesis denote the number of individual companies that belong to that industry (i.e., number of children). We show the finest scale of the SIM model using the sparsity pattern of the *in-scale conditional covariance* in Fig. 7(c). Often, industries or companies that are closely related have a conjugate edge between them. For example, the strongest conjugate edge at scale 3 is the one between the oil and gas extraction industry (SIC code 13) and the petroleum refining industry (SIC code 29). Table I shows four conjugate edges at scale 3 in the order of their absolute magnitude (i.e., the top four strongest in-scale conditional covariance).

Fig. 7(a) shows the sparsity pattern of the *information matrix* of a single-scale approximation. Note that the corresponding graphical model has densely connected edges among companies

[9]For multiscale models, we marginalize out coarser scale variables and use the marginal covariance at the finest scale to compute this divergence.

[10]We disregard 16 companies that have been listed on S&P 100 only after 1990.

[11]http://www.osha.gov/pls/imis/sic_manual.html

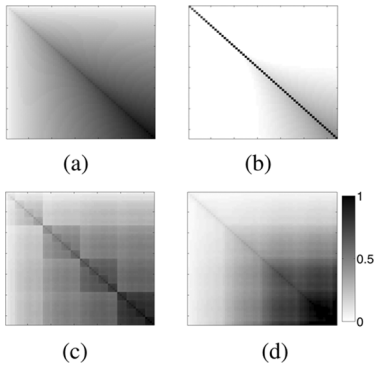


Fig. 8. Covariance approximation for fBm-64. (a) Original model. (b) Single-scale approximation. (c) Tree approximation. (d) SIM model.

that belong to the same industry, because there is no hidden variable to capture the correlations at a coarser resolution. Fig. 7(b) shows the information matrix at the finest scale of a sparse MR model approximation [8]. Although the graphical model is sparser than the single-scale approximation, some of the companies still have densely connected edges. As shown in the caption of Fig. 7, the SIM model approximation provides the smallest divergence of all approximations.

### B. Fractional Brownian Motion

We consider fractional Brownian motion (fBm) [30] with Hurst parameter $H = 0.3$ defined on the time interval $(0, 1]$ with the covariance function: $\Sigma(t_1, t_2) = (1/2)(|t_1|^{2H} + |t_2|^{2H} - |t_1 - t_2|^{2H})$. Note that this is a nonstationary process. Fig. 8 shows the covariance realized by each model using 64 time samples. For the tree model and the SIM model, we only show the marginal covariance of the finest scale variables. Our SIM approximation in Fig. 8(d) is close to the original covariance in Fig. 8(a), while the single-scale approximation in Fig. 8(b) fails to capture long-range correlations and the tree model covariance in Fig. 8(c) appears blocky.

A similar covariance realization without severe blocky artifacts can also be obtained by the sparse MR model of [7]. However, we observe that a SIM model can achieve a smaller divergence with respect to the true model with a smaller number of parameters than the counterpart sparse MR model. Fig. 9(a) shows the sparsity pattern of the conjugate graph (i.e., the conditional covariance) of the finest scale of the SIM model and Fig. 9(b) shows the sparsity pattern of the graphical model (i.e., the information matrix) of the finest scale of the sparse MR model. The SIM model has 134 conjugate edges at the finest scale and the sparse MR model has 209 edges. The divergence with respect to the true distribution is 1.62 for the SIM model and 2.40 for the sparse MR model. Moreover, note that the structure of the conjugate graph in Fig. 9(a) is mostly local, but in the sparse MR model in Fig. 9(b), some nodes are connected to many other nodes. This suggests that the conditional covariance structure is a more natural representation for capturing in-scale statistics.

Fig. 10(a) displays a 256-point sample path using the exact statistics and Fig. 10(b) displays sparse and noisy observations of Fig. 10(a). Observations are only available on $(0, 1/3]$ (over

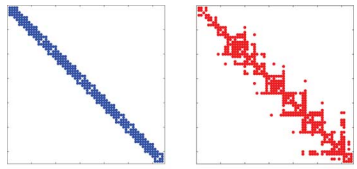


Fig. 9. Sparsity pattern of (a) the in-scale conditional covariance of the finest scale of the SIM model and (b) the information matrix of the finest scale of the sparse MR model for the fBm-64 example.

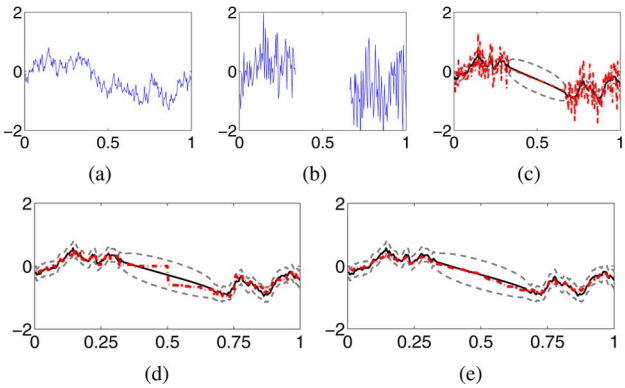


Fig. 10. Estimation for fBm-256. (a) Sample-path using exact statistics. (b) Noisy and sparse observations of (a). Estimates using (c) single-scale approximation, (d) tree model, and (e) SIM model are shown in the dashed–dotted lines. In each figure, the solid black line indicates the optimal estimate based on exact statistics, and the dashed gray lines show plus/minus one standard deviation error bars of the optimal estimate.

TABLE II
FBM-256 APPROXIMATION

| | #Variables | #Parameters[a] | Divergence[b] | RMS[c] |
|---|---|---|---|---|
| Original | 256 | 32896 | 0 | 0 |
| Single | 256 | 20204 | 3073 | 0.2738 |
| Tree | 341 | 681 | 80.4 | 0.1134 |
| Sparse MR | 341 | 1699 | 15.68 | 0.1963 |
| **SIM** | **341** | **1401** | **8.56** | **0.0672** |

[a] #variables + #edges (or conjugate edges)
[b] Divergence between original and approximation
[c] Root-mean-square error w.r.t. the optimal estimate

which the noise variance is 0.3) and $(2/3, 1]$ (with noise variance 0.5). Fig. 10(c)–(e) shows the estimates (in dashed–dotted line) based on the approximate single-scale model, the tree, and the SIM model, respectively, together with the optimal estimate based on the exact statistics (in solid black). The dashed gray lines in Fig. 10(c)–(e) indicate plus/minus one standard deviation error bars of the optimal estimate. We see that the single-scale estimate differs from the optimal estimate by a significant amount (exceeding the error bars around the optimal estimate), while both the tree estimate and the SIM estimate are close to the optimal estimate (i.e., well within the error bars around the optimal). In addition, the estimate based on our SIM model does not have blocky artifacts as in the estimate based on the tree.

The performance of each model is summarized in Table II. Note that the number of parameters (number of nodes plus the number of (conjugate) edges) in the SIM model is much smaller than the original or the single-scale approximate model. Specifically, the number of interscale edges and conjugate in-scale edges in the SIM model is $\mathcal{O}(N)$ while the number of edges in

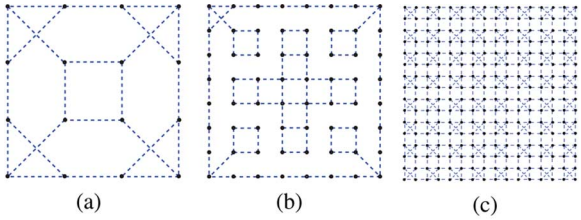

Fig. 11. Conjugate graph at each scale of the SIM model for polynomially decaying covariance approximation. (a) Scale 2 ($4 \times 4$). (b) Scale 3 ($8 \times 8$). (c) Scale 4 ($16 \times 16$).

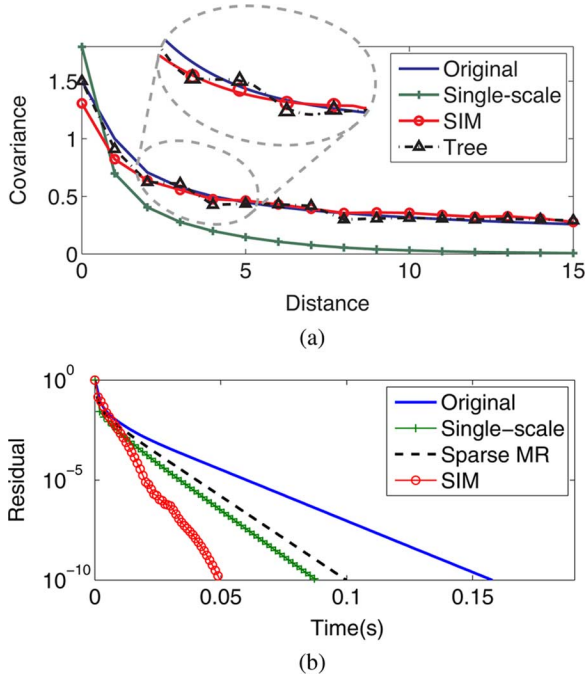


Fig. 12. (a) Covariance behavior of various models. (b) Comparison of inference performance for polynomially decaying covariance experiments.

the original and the single-scale approximation is $\mathcal{O}(N^2)$ where $N$ is the number of variables.

### C. Polynomially Decaying Covariance for A 2-D Gaussian Field

We consider a collection of 256 Gaussian random variables arranged spatially on a $16 \times 16$ grid. The variance of each variable is given by $\Sigma_{x_s} = 1.5$ and the covariance between each pair of variables is given by $\Sigma_{x_s,x_t} = d(s,t)^{-(1/2)}$, where $d(s,t)$ is the spatial distance between nodes $s$ and $t$. The original graphical structure (corresponding to the inverse of the specified covariance matrix) is fully connected, and the single-scale approximation of it is still densely connected with each node connected to at least 31 neighbors.

Fig. 11 shows the *conjugate* graph of the SIM model approximation within each scale, i.e., the sparsity of the conditional covariance at that scale. We emphasize that these conjugate edges encode the in-scale conditional correlation structure among the variables directly, so each node is only *locally* correlated when conditioned on other scales. Fig. 12(a) displays the covariance as a function of the distance between a pair of nodes. The covariance of the single-scale approximation falls off much more rapidly than that of the original model, and the magnified portion of the plot emphasizes the blocky artifacts of the tree model.

TABLE III
POLYNOMIALLY DECAYING COVARIANCE APPROXIMATION

| | #Variables | #Parameters | Divergence | Time(s)[a] |
|---|---|---|---|---|
| Original | 256 | 32896 | 0 | 0.1511 |
| Single | 256 | 7513 | 64.5 | 0.0942 |
| Tree | 341 | 681 | 34.3 | 0.0013 |
| Sparse MR | 341 | 973 | 8.38 | 0.1035 |
| **SIM** | **341** | **1396** | **6.87** | **0.0773** |

[a]Average computation time in seconds for solving the inference problem

We conclude that our SIM model provides good modeling capabilities for processes with long-range correlation.

To compare the inference performance, we generate random noisy measurements using the specified statistics and compare the computation time to solve the inference problem for the SIM model (using the inference algorithm in Section IV-B), the original and the single-scale approximate model (using the ET algorithm described in Section IV-A), and the sparse MR model (using the algorithm in [8]). Table III shows the average time until convergence (the relative residual error $\epsilon$ reaches $10^{-10}$) averaged over 100 experiments, and Fig. 12(b) shows the residual error versus computation time for one set of random measurements.[12] The SIM modeling approach provides a significant gain in convergence rate over the other models. Note that the sparse MR model has a smaller number of parameters, but the divergence and the average time until convergence are larger. Hence, even though sparse MR models have advantages over single-scale approximations, SIM models provide more accurate approximations of the underlying process and enable more efficient inference procedures.

## VII. Conclusion and Future Work

In this paper, we have introduced a new class of Gaussian MR models with sparse in-scale conditional covariance structure at each scale and tree-structured connections across scales. In our SIM model, each variable is correlated with only a few other variables in the same scale when conditioned on other scales. Our approach overcomes the limitations of tree-structured MR models and provides good modeling performance especially in capturing long-range covariance behavior without blocky artifacts. In addition, by decomposing the information matrix of the resulting MR model into the sum of a sparse matrix (the information matrix corresponding to interscale graphical structure) and an information matrix that has a sparse inverse (the in-scale conditional covariance), we develop an efficient inference algorithm utilizing the sparsity in both Markov and covariance structure. Our algorithm alternates computations across scales using the sparse interscale graphical structure, and in-scale computations that reduce to sparse matrix multiplications.

We also describe a method for learning models with this structure, i.e., for building SIM models that provide a good approximation to a target covariance. Given a target covariance at the finest scale, our learning algorithm first constructs an exact MR model for the target covariance, and then optimizes the structure of each scale using log-determinant maximization to obtain

[12]The computation time was measured at AMD Opteron 270 Dual Core Processor using Matlab 7.4.0 code.

a sparse conjugate graph approximation. In Appendix I, we introduce one method to construct an exact MR model, which first learns a good MR tree model and then augments each scale in a coarse-to-fine way. An important and interesting extension of our learning method would be to alternatively optimize the tree and the in-scale models in a computationally tractable way. Although for simplicity we assumed that the interscale structure of SIM models is a tree, our inference procedure is efficient for the more general case of having a sparse interscale structure (but not necessarily a tree) as well.

SIM models are of most value when there are long-distance correlations, which are most prominent in multidimensional data such as in geophysical fields, and the application of our methods in such areas is a promising line of work. While our focus in this paper is on the Gaussian model, applying similar principles to discrete or other more general models is also of interest. Although the sparse matrix multiplication and the log-det optimization framework for Gaussian models are not directly applicable to the discrete case, we expect that having a sparse in-scale dependency structure at each scale conditioned on other scales may still result in efficient inference and learning algorithms.

## Appendix I
## Computing the Target Information Matrix of an MR Model

Suppose that we are given a target covariance $\Sigma_F^*$ of the variables at the finest scale. In this section, we discuss a method to introduce hidden variables at coarser scales and build an *exact* MR model, so that when we marginalize out all coarser scale variables, the marginal covariance at the finest scale is exactly equal to $\Sigma_F^*$. The information matrix of this exact MR model can be used as the target information matrix $J^*$ in (10) to obtain a SIM model approximation.

To begin with, we learn an interscale model by selecting a tree structure (without any in-scale connections) with *additional hidden* variables at coarser scales and the original variables at the finest scale. Selecting a good tree structure is important, but this structure does not need to be perfect since we later augment the interscale model with in-scale structures. For some processes, there exists a natural hierarchical structure: for example, for regular 1-D or 2-D processes, the MR tree models in Fig. 1 can be used. For other problems in which the spatial relation among the variables is not clearly defined, we can group variables that are highly correlated and insert one coarser scale variable for each group. Once the structure is fixed, the EM algorithm [5] can be applied to choose the parameters that best match the given target covariance $\Sigma_F^*$ for the finest scale variables. This procedure is efficient for a tree-structured model.

After the parameter fitting, we have an information matrix $J_{\text{tree}}$ corresponding to an MR tree model. Although the EM algorithm will adjust the elements of $J_{\text{tree}}$ so that the marginal covariance at the finest scale is close to $\Sigma_F^*$, it will in general not match the cross-correlation between variables at different finest scale nodes. As mentioned in Section V-C, if we view $J_{\text{tree}}$ as a first approximation to $J^*$, it has a structure as in Fig. 5 except that the in-scale conditional structure that we have learned (the shaded blocks in $J^c$ in the figure) is *diagonal* rather than full, resulting in artifacts that correspond to inaccurate matching of finest scale cross covariances. As a result, the basic idea of our construction is to recursively modify our approximation to $J^*$, from coarse-to-fine scales to get full matching of marginal statistics at the finest scale.

In an MR tree model, the covariance matrix at each scale can be represented in terms of the covariance at the next finer scale

$$\Sigma_{[m]} = A_m \Sigma_{[m+1]} A_m^T + Q_m \tag{11}$$

where $A_m$ and $Q_m$ are determined by $J_{\text{tree}}$.[13] Since we wish to modify the tree model so that the covariance matrix at the finest scale becomes $\Sigma_F^*$, we set $\Sigma_{[M]} = \Sigma_F^*$ for the finest scale $M$ and compute a target marginal covariance for each scale in a *fine-to-coarse* way using (11). These target marginal covariances at each scale can be used to modify $J^*$. Specifically, the diagonal matrix $J_{[m]}^*$ of the tree model is replaced with a nondiagonal matrix so that the *marginal* covariance at scale $m$ is equal to $\Sigma_{[m]}$, the target marginal covariance at that scale computed using (11). In modifying $J^*$, we proceed in a *coarse-to-fine* way. Suppose that we have replaced $J_{[1]}^*$ through $J_{[m-1]}^*$, and let us consider computing $J_{[m]}^*$. We partition the information matrix of the resulting MR model into nine submatrices with the in-scale information matrix at scale $m$ at the center[14]:

$$J^* = \begin{pmatrix} J_c^* & J_{c,[m]}^* & 0 \\ J_{[m],c}^* & J_{[m]}^* & J_{[m],f}^* \\ 0 & J_{f,[m]}^* & J_f^* \end{pmatrix}. \tag{12}$$

Note that except for $J_c^*$, all submatrices are equivalent to the corresponding components in $J_{\text{tree}}$ because we have only replaced coarser in-scale blocks.

From (12), the marginal covariance at scale $m$ is $(J_{[m]}^* - J_{[m],c}^*(J_{[c]}^*)^{-1}J_{[c],m}^* - J_{[m],f}^*(J_{[f]}^*)^{-1}J_{f,[m]}^*)^{-1}$. By setting this equal to the target covariance matrix $\Sigma_{[m]}$ in (11), the *target information matrix at scale* $m$ can be computed as follows:

$$J_{[m]}^* = (\Sigma_{[m]})^{-1} + J_{[m],c}^*(J_c^*)^{-1}J_{c,[m]}^* + J_{[m],f}^*(J_f^*)^{-1}J_{f,[m]}^* \tag{13}$$

which we replace with $J_{[m]}^*$ in (12) and proceeds to the next finer scale until we reach the finest scale. The matrix inversion in the above equation requires computation that is cubic in the number of variables $N$. Learning a graphical model structure typically involves at least $\mathcal{O}(N^4)$ computation [1], so computing $J_{[m]}^*$ is not a bottleneck of the learning process.

After the algorithm augments in-scale structures for all scales, the resulting information matrix $J^*$ has the marginal covariance at the finest scale exactly equal to the target covariance matrix $\Sigma_F^*$. In addition, $J^*$ has dense in-scale structure both as a graphical model and in terms of the corresponding conjugate graph (since in general the matrix $J_{[m]}^*$ is not sparse and does not have a sparse inverse), and a sparse interscale graphical structure. Hence, the information matrix $J^*$ can be used as the target

[13]Let $B_m = (J_{\text{tree}})_{[m-1,m]}^{-1}$ and $D_m = (J_{\text{tree}})_{[m]}^{-1}$. Then, $A_m = B_m D_m^{-1}$ and $Q_m = D_{m-1} - B_m D_m^{-1} B_m^T$.

[14]For $m = 1$ (the coarsest scale) and $m = M$ (the finest scale), the partition consists of only four submatrices. Also, the 0-blocks in (12) are immediate because of the MR structure, which does not have edges directly between scales $m-1$ and $m+1$.

TABLE IV
LEARNING ALGORITHM IN DETAIL

1) Select an MR tree structure that best describes the given collection of variables at the finest scale.
2) Fit the parameters of the MR tree using the EM algorithm.
3) Compute the target covariance at each scale using (11).
4) At the coarsest scale ($m = 1$), compute the target information matrix at scale $m$ using (13).
5) Solve the problem in (15) to obtain $\hat{J}_{[m]}$ with a sparse inverse.
6) Replace the information matrix at scale $m$ of the MR model with $\hat{J}_{[m]}$. If the resulting information matrix $\hat{J}$ is not positive definite, increase the node parameter $\gamma_s$ and repeat Step 5.
7) Go to the next finer scale and repeat steps 4 through 6 until we reach the finest scale.

information matrix of the MR model in (10) with the $\ell_1$-norm dropped from the objective function to learn a SIM model approximation.

## APPENDIX II
## SEQUENTIAL STRUCTURE OPTIMIZATION

In Appendix I, we constructed an exact MR model such that the marginal covariance at the finest scale matches the specified target covariance exactly. The information matrix of the exact MR model can be used as the target information matrix in (10) to learn a SIM model approximation. In this section, we introduce an alternative approach to learn a SIM model; instead of first constructing an exact MR model across all scales and then optimizing the structure of all scales in parallel by solving (10), one can interleave the procedure of finding a target information matrix at scale $m$ and optimizing its structure to have a sparse conjugate graph.

After computing the target information matrix $J^*_{[m]}$ at scale $m$ using (13) (before proceeding to compute $J^*_{[m+1]}$ at the next finer scale), we perform structure optimization at scale $m$ to obtain a sparse in-scale conditional covariance approximation (i.e., a sparse conjugate graph). This in-scale structure optimization can be performed by solving a simplified version of the log-det problem in (10). Since the interscale edges of $J^*$ are sparse by our construction, the $\ell_1$-norm can be dropped from the objective function of (10). In addition, the parameters at all scales other than scale $m$ are fixed. Thus, the optimization problem reduces to the following:

$$\begin{aligned} \hat{J} = \underset{J \succ 0}{\operatorname{argmax}} \quad & \log\det J_{[m]} \\ \text{s.t.} \quad & |J_{i,j} - J^*_{i,j}| \le \gamma \qquad \forall i,j \in V_{[m]} \end{aligned} \tag{14}$$

where $V_{[m]}$ is the set of nodes at scale $m$. Using the approximation techniques described in Appendix III, the above problem can be solved more efficiently than the problem in (10) that does not use the sequential approach.

## APPENDIX III
## COMPUTATIONAL SIMPLIFICATIONS IN SOLVING THE LOG-DET PROBLEM

In this section, we introduce some techniques to obtain an approximate solution of the log-determinant problem in (14) efficiently, and provide a method for choosing the regularization parameters. The problems in (10) and (14) are both convex and can be solved using standard convex optimization techniques [4]. In order to further reduce the computational complexity, we ignore the positive–definiteness condition $J \succ 0$ until we find a solution that maximizes the log-determinant with the element-wise constraints satisfied. Then, the problem reduces to (9) that involves only the information matrix at scale $m$, $J_{[m]}$, which can be efficiently solved using the techniques in [1], [17], and [21]. If, after replacing $J_{[m]}$ with the solution $\hat{J}_{[m]}$, the entire information matrix $\hat{J}$ is positive–definite, then $\hat{J}_{[m]}$ is indeed the optimal solution. If $\hat{J}$ is not positive–definite, then we adjust the regularization parameter, and for this purpose, we allow two regularization parameters: one for all nodes and one for all edges

$$\begin{aligned} \hat{J}_{[m]} = \operatorname{argmax} \quad & \log\det J_{[m]} \\ \text{s.t.} \quad & |J_{i,j} - J^*_{i,j}| \le \gamma_E \qquad \forall i \neq j \in V_m \\ & |J_{i,i} - J^*_{i,i}| \le \gamma_s \qquad \forall i \in V_m \end{aligned} \tag{15}$$

where $\gamma_E$ and $\gamma_s$ are parameters for edges and nodes, respectively. Note that the KKT conditions of the above problem are exactly the same as those in Proposition 1, and the inverse of $\hat{J}_{[m]}$ (the conjugate graph at scale $m$) is sparse.

It is straightforward to show that the optimal solution of (15) has the diagonal elements equal to $\operatorname{diag}(J^*_{[m]}) + \gamma_s I$, so for large enough value of $\gamma_s$, $\hat{J}$ becomes positive–definite. Therefore, if the resulting $\hat{J}$ is not positive–definite, we can increase the value of $\gamma_s$. In practice, we set $\gamma_E$ equal to $0.5 \times \xi$ where $\xi$ is the maximum value of the off-diagonal elements of $J^*_{[m]}$, and set the initial value of $\gamma_s = 2\gamma_E$ for all coarser scales. For the finest scale, we use $\gamma_E = 0.25 \times \xi$ and adjust $\gamma_s$ so that the divergence between the approximate and target distribution is minimized.

After every scale in the MR model is augmented with a sparse conjugate graph, the resulting SIM model has a sparse interscale structure, and a sparse conjugate graph at each scale. Table IV summarizes the algorithm for learning a SIM model given the target covariance at the finest scale.

## ACKNOWLEDGMENT

The authors would like to thank Prof. H. Chen for helpful discussions about the stock returns example.

## REFERENCES

[1] O. Banerjee, L. E. Ghaoui, A. D'Aspremont, and G. Natsoulis, "Convex optimization techniques for fitting sparse Gaussian graphical models," in *Proc. 23rd Annu. Int. Conf. Mach. Learn.*, 2006, pp. 89–96.
[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.
[3] C. A. Bouman and M. Shapiro, "A multiscale random field model for Bayesian image segmentation," *IEEE Trans. Image Process.*, vol. 3, no. 2, pp. 162–177, Mar. 1994.
[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[5] W. L. Briggs, *A Multigrid Tutorial*. Philadelphia, PA: SIAM, 1987.
[6] V. Chandrasekaran, J. K. Johnson, and A. S. Willsky, "Estimation in Gaussian graphical models using tractable subgraphs: A walk-sum analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1916–1930, May 2008.
[7] M. J. Choi, V. Chandrasekaran, and A. S. Willsky, "Maximum entropy relaxation for mutiscale graphical model selection," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Apr. 2008, pp. 1889–1892.
[8] M. J. Choi and A. S. Willsky, "Multiscale Gaussian graphical models and algorithms for large-scale inference," in *Proc. IEEE Statist. Signal Process. Workshop*, Aug. 2007, pp. 229–233.

[9] M. J. Choi, V. Chandrasekaran, and A. S. Willsky, "Exploiting sparse Markov and covariance structure in multiresolution models," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 177–184.
[10] K. C. Chou, A. S. Willsky, and A. Benveniste, "Multiscale recursive estimation, data fusion, and regularization," *IEEE Trans. Autom. Control*, vol. 39, no. 3, pp. 464–478, Mar. 1994.
[11] M. L. Comer and E. J. Delp, "Segmentation of textured images using a multiresolution Gaussian autoregressive model," *IEEE Trans. Image Process.*, vol. 8, no. 3, pp. 408–420, Mar. 1999.
[12] D. R. Cox and N. Wermuth, *Multivariate Dependencies: Models, Analysis and Interpretation*. London, U.K.: Chapman & Hall/CRC, 1996.
[13] C. Crick and A. Pfeffer, "Loppy belief propagation as a basis for communication in sensor networks," in *Proc. 19th Conf. Uncertainty Artif. Intell.*, 2003, pp. 159–166.
[14] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.
[15] V. Delouille, R. Neelamani, and R. Baraniuk, "Robust distributed estimation using the embedded subgraphs algorithm," *IEEE Trans. Signal Process.*, vol. 54, no. 8, pp. 2998–3010, Aug. 2006.
[16] M. Drton and M. D. Perlman, "Model selection for Gaussian concentration graphs," *Biometrika*, vol. 91, no. 3, pp. 591–602, 2004.
[17] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.
[18] G. H. Golub and C. H. Van Loan, *Matrix Computations*. Baltimore, MD: The Johns Hopkins Univ. Press, 1990.
[19] C. Graffigne, F. Heitz, P. Perez, F. Prêteux, M. Sigelle, and J. Zerubia, "Hierarchical Markov random field models applied to image analysis: A review," in *SPIE Conference Series*. Bellingham, WA: SPIE, 1995, vol. 2568, pp. 2–17.
[20] L. Greengard and V. Rokhlin, "A fast algorithm for particle simulations," *J. Comput. Phys.*, vol. 73, no. 2, pp. 325–348, 1987.
[21] J. K. Johnson, V. Chandrasekaran, and A. S. Willsky, "Learning Markov structure by maximum entropy relaxation," in *Proc. 11th Int. Conf. Artif. Intell. Statist.*, Mar. 2007.
[22] A. Kannan, M. Ostendorf, W. C. Karl, D. A. Castanon, and R. K. Fish, "ML parameter estimation of multiscale stochastic processes using the EM algorithm," *IEEE Trans. Signal Process.*, vol. 48, no. 6, pp. 1836–1847, Jun. 2000.
[23] Z. Kato, M. Berthod, and J. Zerubia, "Multiscale Markov random field models for parallel image classification," in *Proc. Int. Conf. Comput. Vis.*, May 1993, pp. 253–257.
[24] G. Kauermann, "On a dualization of graphical Gaussian models," *Scandinavian J. Statist.*, vol. 23, pp. 105–116, 1996.
[25] A. J. Krener, R. Frezza, and B. C. Levy, "Gaussian reciprocal processes and self-adjoint stochastic differential equations of second order," *Stochastics Stochastics Rep.*, vol. 34, pp. 29–56, Jun. 1991.
[26] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Univ. Press, 1996.
[27] S. Lee, V. Ganapathi, and D. Koller, "Efficient structure learning of Markov networks using $l_1$ regularization," in *Advances in Neural Information Processing Systems (NIPS) 19*. Cambridge, MA: MIT Press, 2007.
[28] J. Li, R. M. Gray, and R. A. Olshen, "Multiresolution image classification by hierarchical modeling with two-dimensional hidden Markov models," *IEEE Trans. Inf. Theory*, vol. 46, no. 5, pp. 1826–1841, Aug. 2000.
[29] J. Löfberg, "Yalmip: A toolbox for modeling and optimization in MATLAB," in *Proc. Comput.-Aided Control Syst. Design Conf.*, 2004, pp. 284–289 [Online]. Available: http://control.ee.ethz.ch/joloef/yalmip.php
[30] B. B. Mandelbrot and J. W. Van Ness, "Fractional Brownian motions, fractional noises and applications," *SIAM Rev.*, vol. 10, pp. 422–437, 1968.
[31] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Cambridge, MA: MIT Press, 1999, pp. 355–368.
[32] S. Osindero and G. Hinton, "Modeling image patches with a directed hierarchy of Markov random fields," in *Advances in Neural Information Processing Systems (NIPS) 20*. Cambridge, MA: MIT Press, 2008.
[33] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky, "Embedded trees: Estimation of Gaussian processes on graphs with cycles," *IEEE Trans. Signal Process.*, vol. 52, no. 11, pp. 3136–3150, Nov. 2004.
[34] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, Aug. 2002.

**Myung Jin Choi** (S'06) received the B.S. degree in electrical engineering and computer science from Seoul National University, Seoul, Korea, in 2005 and the S.M. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 2007, where she is currently working towards the Ph.D. degree with the Stochastic Systems Group.

She is a Samsung scholarship recipient. Her research interests include statistical signal processing, graphical models, and multiresolution algorithms.

**Venkat Chandrasekaran** (S'03) received the B.S. degree in electrical engineering and the B.A. degree in mathematics from Rice University, Houston, TX, in 2005 and the S.M. degree in electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 2007, where he is currently working towards the Ph.D. degree with the Stochastic Systems Group.

His research interests include statistical signal processing, optimization methods, machine learning, and computational harmonic analysis.

**Alan S. Willsky** (S'70–M'73–SM'82–F'86) received the S.B. and Ph.D. degrees from the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology (MIT), Cambridge, in 1969 and 1973, respectively.

He joined the MIT faculty, in 1973 and is the Edwin Sibley Webster Professor of Electrical Engineering and Director of the Laboratory for Information and Decision Systems. He was a founder of Alphatech, Inc. and Chief Scientific Consultant, a role in which he continues at BAE Systems Advanced Information Technologies. From 1998 to 2002, he served on the U.S. Air Force Scientific Advisory Board. He has delivered numerous keynote addresses and is coauthor of the text *Signals and Systems* (Englewood Cliffs, NJ: Prentice-Hall, 1996). His research interests are in the development and application of advanced methods of estimation, machine learning, and statistical signal and image processing.

Dr. Willsky received several awards including the 1975 American Automatic Control Council Donald P. Eckman Award, the 1979 ASCE Alfred Noble Prize, the 1980 IEEE Browder J. Thompson Memorial Award, the IEEE Control Systems Society Distinguished Member Award in 1988, the 2004 IEEE Donald G. Fink Prize Paper Award, and Doctorat Honoris Causa from Universit de Rennes in 2005. He and his students, colleagues, and postdoctoral associates have also received a variety of Best Paper Awards at various conferences and for papers in journals, including the 2001 IEEE Conference on Computer Vision and Pattern Recognition, the 2003 Spring Meeting of the American Geophysical Union, the 2004 Neural Information Processing Symposium, Fusion 2005, and the 2008 award from the journal Signal Processing for the outstanding paper in the year 2007.