
Exploiting Sparse Markov *and* Covariance Structure in Multiresolution Models

Myung Jin Choi
Venkat Chandrasekaran
Alan S. Willsky

MYUNGJIN@MIT.EDU
VENKATC@MIT.EDU
WILLSKY@MIT.EDU

Department of EECS, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139

Abstract

We consider Gaussian multiresolution (MR) models in which coarser, hidden variables serve to capture statistical dependencies among the finest scale variables. 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 that capture the residual correlations within each scale using *sparse covariance structure*. Our goal is to learn a tree-structured graphical model connecting variables across different scales, while at the same time learning sparse structure for the conditional covariance 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.

1. Introduction

Multiresolution (MR) models (Willsky, 2002) provide compact representations for encoding statistical dependencies among a large collection of random variables. In MR models, variables at coarser resolutions serve as common factors for explaining statistical dependencies among finer scale variables. For example, suppose that we would like to discover the dependency structure of the monthly returns of 100 different stocks by looking at pairwise correlations. It is likely that the covariance matrix will be full, i.e., the monthly return of one stock is correlated to all other 99 stocks, because stock prices tend to move together driven by the market situation. Therefore, it is more informative to introduce a hidden variable corresponding to the market and then model the residual covariance (after conditioning on the market) among the individual companies. This ap-

proach can be extended to multiple resolutions - representing the market, divisions, industries, and individual companies at each scale from the coarsest to the finest.

One approach in MR modeling is to use tree-structured graphical models in which nodes at any scale are connected to each other only through nodes at other scales (see Figure 1). While such tree models allow efficient inference and learning algorithms, they have a significant and well-known limitation that variables at any of the scales are *conditionally uncorrelated* when conditioned on neighboring scales. In our stock return example, the Standard Industrial Classification (SIC) system, a hierarchy widely-used in finance, places Microsoft and Apple in different branches of the tree because the former belongs to the business service industry in the services division while the latter belongs to the computer equipment industry in the manufacturing division. Tree-based modeling methods will assume that the monthly returns of Microsoft and Apple are uncorrelated conditioned on the market, which is likely not true.

A variety of methods (Bouman & Shapiro, 1994; Choi & Willsky, 2007) have been proposed to include additional edges - either inter-scale or within the same scale - to the MR tree model and to consider an overall sparse MR graphical model. We propose a different approach to address the limitation of MR tree models. Since the role of coarser scales in an MR model is to capture most of the correlations among the finer scale variables through coarser scales, shouldn't the *residual* correlation at each scale be (approximately) *sparse*? In other words, the residual correlation of any node (conditioned on coarser nodes) is concentrated completely on a small number of nodes at that scale. This suggests that the *conditional* correlations at each scale (when conditioned on the neighboring scales) should be sparse. Based on this idea, we can model that conditioned on the market and industries, Microsoft is correlated with Apple and possibly with a few other companies such as Google or IBM.

Such models lead to efficient inference algorithms that are fundamentally different from standard graphical model inference algorithms. We use the sparse tree structure *be-*

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

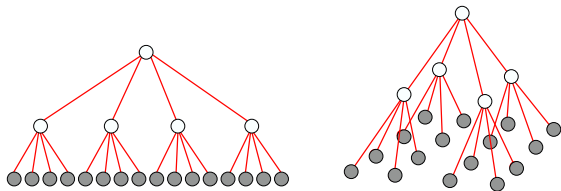


Figure 1. Examples of MR tree models for a one-dimensional process (left) and for a two-dimensional process (right). Shaded nodes represent original variables at the finest scale and white nodes represent hidden variables at coarser scales.

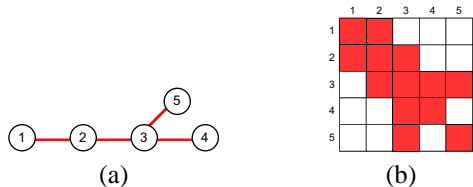


Figure 2. (a) A sparse graphical model and (b) the sparsity pattern of the corresponding information matrix.

tween scales, to propagate information from scale-to-scale, and then perform residual filtering *within* each scale using the sparse conditional covariance structure. In addition, we develop methods for *learning* such models given data at the finest scale. The structure optimization within each scale can be formulated as a convex optimization problem.

2. Preliminaries

Let $x \sim \mathcal{N}(\mu, \Sigma)$ be a jointly Gaussian random vector with a mean vector μ and a positive-definite covariance matrix Σ . 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 matrix) is sparse with respect to \mathcal{G} . That is, $J_{s,t} \neq 0$ if and only if $\{s, t\} \in \mathcal{E}$ (Lauritzen, 1996). Figure 2(a) shows one example of a sparse graph, and the sparsity pattern of the corresponding information matrix J is shown in Figure 2(b). The graph structure implies that x_1 is uncorrelated with x_5 *conditioned on* x_2 . For any subset $A \subset \mathcal{V}$, let $\setminus A \equiv \{s \in \mathcal{V}, s \notin A\}$ be its complement. The information matrix of the conditional distribution $p(x_A | x_{\setminus A})$ is the *submatrix* of J with rows and columns corresponding to elements in A . In Figure 2(b), the information matrix of the conditional distribution $p(x_1, x_2, x_3, x_4 | x_5)$ is the submatrix $J(1 : 4, 1 : 4)$, which is a tri-diagonal matrix.

Conjugate Graphs Consider a distribution with the sparsity pattern of the *covariance matrix* given as in Figure 3(a). Its information matrix will, in general, be a full matrix, and the corresponding graphical model will be fully connected as shown in Figure 3(b). We introduce *conjugate graphs*¹ to illustrate the sparsity structure of a covari-

¹This term is motivated by conjugate processes - two processes with covariances that are inverses of one another. This

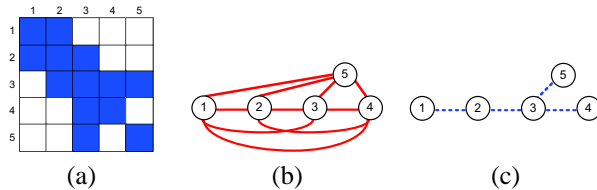


Figure 3. (a) Sparsity pattern of a covariance matrix and (b) the corresponding graphical model. (c) Conjugate graph encoding the sparsity structure of the covariance matrix in (a).

ance matrix. Specifically, in the conjugate graph, when two nodes are not connected with a *conjugate* edge, they are *uncorrelated* with each other. We use red solid lines to display graphical model edges, and blue dotted lines to represent conjugate edges. Figure 3(c) shows the corresponding conjugate graph for a distribution with covariance structure as in Figure 3(a). From the conjugate graph, we can identify that x_1 is uncorrelated with x_3, x_4 , and x_5 .

3. 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. We define *in-scale conditional covariance* as 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). Note that this is different from the more commonly used concept of *pairwise conditional covariance*, which refers to the conditional covariance between two variables when conditioned on *all other variables* (including other variables within the same scale). An information matrix (i.e., a graphical model) is sparse with respect to the pairwise conditional covariance structure. We illustrate the sparsity of the in-scale conditional covariance using the conjugate graph. Thus, our model has a sparse graphical model for inter-scale 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.

Figure 4(b) shows an example of a SIM model: *conditioned on* scale 1 (variable x_1) and scale 3 (variables x_5 through x_{10}), x_2 is *uncorrelated* with x_4 . 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). Indeed, the graphical model representation of the model in Figure 4(b) is a densely connected graphical structure within each scale as shown in Figure 4(c).

graph is also called a *covariance graph* (Cox & Wermuth, 1996).

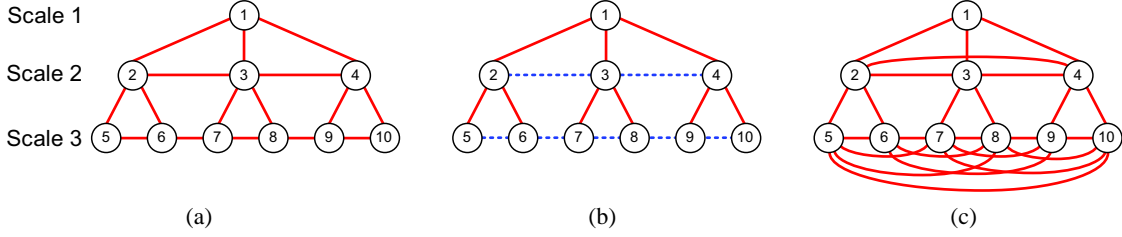


Figure 4. Examples of MR models. (a) An MR model with a sparse graphical structure. (b) A SIM model with sparse conjugate graph within each scale. (c) A graphical model corresponding to the model in (b).

$$J = \begin{pmatrix} J_{[1]} & J_{[1,2]} & 0 \\ J_{[2,1]} & J_{[2]} & J_{[2,3]} \\ 0 & J_{[3,2]} & J_{[3]} \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & J_{[1,2]} & 0 \\ J_{[2,1]} & 0 & J_{[2,3]} \\ 0 & J_{[3,2]} & 0 \end{pmatrix}}_{J^h} + \underbrace{\begin{pmatrix} J_{[1]} & 0 & 0 \\ 0 & J_{[2]} & 0 \\ 0 & 0 & J_{[3]} \end{pmatrix}}_{J^c = (\Sigma^c)^{-1}}$$

Scale 1: + +

Scale 2: + +

Scale 3: + +

Figure 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 non-shaded matrices are sparse.

In contrast, an MR model with a sparse graphical model structure within each scale is shown in Figure 4(a). 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 6, 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 (Bouman & Shapiro, 1994; Choi & Willsky, 2007) focus on constructing an overall sparse graphical model structure (as in Figure 4(a)). A different approach based on a directed hierarchy of densely connected graphical models is proposed in (Osindero & Hinton, 2007), but it does not have a sparse conjugate graph at each layer and requires mean-field approximations unlike our SIM model.

Desired Structure of the Information Matrix 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 Figure 5 (corresponding to a model with 3 scales). The submatrix $J_{[m_1, m_2]}$, corresponding to the graphical structure between scales m_1 and m_2 , is sparse since the inter-scale graphical model has a tree structure. The submatrix $J_{[m]}$ corresponds to the information matrix of the *condi-*

tional distribution at scale m conditioned on other scales (see Section 2). As illustrated in Figure 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 matrix J can be decomposed as a sum of J^h , corresponding to the hierarchical inter-scale 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 (with each block corresponding to variables in one scale), its inverse Σ^c is also block-diagonal with each diagonal block equal to $(J_{[m]})^{-1}$. Hence, Σ^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:

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

Each nonzero entry in J^h corresponds to an interscale edge connecting variables at different scales. The block diagonal matrix Σ^c has nonzero entries corresponding to *conjugate* edges within each scale. In the next section, we take advantage of sparsity in *both* J^h and Σ^c for efficient inference.

4. Inference Exploiting Sparsity in Markov and Covariance Structure

Let x be a collection of random variables with a prior distribution $\mathcal{N}(0, J^{-1})$, and y be a set of noisy measurements: $y = Cx + v$ where C is a selection matrix, and v is a zero-mean Gaussian noise vector with a diagonal covariance matrix R . Thus, we have in our setup noisy measurements y available at a subset of the nodes corresponding to the variables x . Then, the MAP estimate \hat{x} is given as follows:

$$\hat{x} = \underset{x}{\operatorname{argmax}} p(x|y) = \mathbb{E}[x|y] = (J + J^p)^{-1}h \quad (2)$$

where $J^p \equiv C^T R^{-1} C$ is a diagonal matrix, and $h \equiv C^T R^{-1} y$. If J corresponds to a tree-structured model, (2) 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 in Section 4.1, and propose a new and efficient inference algorithm that solves (2) for our SIM model in Section 4.2.

4.1. Iterative Algorithms Based on a Matrix Splitting

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

$$\hat{x}^{new} = M^{-1}(h + K\hat{x}^{old}) \quad (3)$$

where \hat{x}^{old} is the value of \hat{x} at the previous iteration, and \hat{x}^{new} is the updated value at the current iteration. The matrix M is called a *preconditioner*, and (3) corresponds to the preconditioned Richardson iterations (Golub & Van Loan, 1990). If solving the equation $M\hat{x} = z$ for a fixed vector z is easy, each iteration can be performed efficiently. There are a variety of ways in which splittings can be defined. For example, Gauss-Jacobi iterations set the preconditioner M as a diagonal matrix with diagonal elements of A , and embedded tree (ET) algorithms (Sudderth et al., 2004) split the matrix so that M has a tree structure.

4.2. 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 (1). Our goal is to solve the equation $(J^h + (\Sigma^c)^{-1} + J^p)\hat{x} = h$ where J^h , Σ^c , and J^p are all sparse matrices. We alternate between two inference steps corresponding to *inter-scale* computation and *in-scale* computation in the MR model. Our inter-scale 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.

Tree Inference In this step, we select the inter-scale tree structure as the preconditioner in (3) by setting $M = J^h + J^p + D$, where D is a diagonal matrix added to ensure that M is positive-definite.²

$$(J^h + J^p + D)\hat{x}^{new} = h - (\Sigma^c)^{-1}\hat{x}^{old} + D\hat{x}^{old} \quad (4)$$

With the right-hand side vector fixed, solving the above equation is efficient since M has a tree structure. 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 Σ^c (in-scale conditional covariance) is sparse and well-conditioned in general; hence the equation can be solved

²In (3), M needs to be invertible, but $(J^h + J^p)$ is singular since the diagonal elements at coarser scales (without measurements) are zero. We use $D = (\text{diag}(\Sigma^c))^{-1}$ where $\text{diag}(\Sigma^c)$ is a diagonal matrix with diagonal elements of Σ^c .

efficiently. In our experiments, we use just a few Gauss-Jacobi iterations (see Section 4.1) to compute z .

In-scale Inference This step selects the in-scale structure to perform computations by setting $M = (\Sigma^c)^{-1}$.

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

Evaluating the right-hand side only involves multiplications of a sparse matrix Σ^c and a vector, so \hat{x}^{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 (3), our algorithm is efficient for a fundamentally different reason. In the Richardson iteration (specifically, the ET algorithm) and in our 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.

The concept of performing local in-scale computations can be found in multipole methods (Greengard & Rokhlin, 1987) that use multiple scales to solve partial differential equations. Multipole methods assume that after a solution is computed at coarser resolutions, only *local* terms need to be modified at finer resolutions. The SIM model is aimed at providing a precise statistical framework leading to inference algorithms with solid advantages analogous to those of multipole methods.

5. Learning MR Models with Sparse In-scale Conditional Covariance

5.1. Log-determinant Maximization

Suppose that we are given a target covariance Σ^* 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 is to 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. Thus, standard approaches in Gaussian graphical model selection solve the following log-determinant optimization problem to find an approximate covariance matrix:

$$\begin{aligned} \hat{\Sigma} = \operatorname{argmax}_{\Sigma \succ 0} \quad & \log \det \Sigma \\ \text{s.t.} \quad & |\Sigma_{i,j} - \Sigma_{i,j}^*| \leq \gamma_{i,j}, \quad \forall i, j \end{aligned} \quad (6)$$

where $\gamma_{i,j}$ is a nonnegative regularization parameter. It can be shown that the solution of the above problem has

a sparse inverse, which is a sparse graphical model approximation (Banerjee et al., 2006).

We now turn the tables and 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 above). We again use the log-determinant problem, but now in the information matrix domain:

$$\hat{J} = \underset{J \succ 0}{\operatorname{argmax}} \quad \log \det J$$

$$\text{s.t.} \quad |J_{i,j} - J_{i,j}^*| \leq \gamma_{i,j}, \quad \forall i, j \quad (7)$$

where J^* is a target information matrix. The solution \hat{J} has a sparse inverse, leading to a sparse covariance approximation. In our MR modeling approach, we apply this sparse covariance approximation method to model the conditional distribution at each scale conditioned on other scales.

5.2. Learning a SIM Model

Suppose that we are given a target covariance Σ_F^* of the variables at the finest scale. Our objective is to introduce hidden variables at coarser scales and learn a SIM model, so that when we marginalize out all coarser scale variables, the marginal covariance at the finest scale is approximately equal to Σ_F^* . Our learning procedure consists of three steps. First, we learn the inter-scale part of the SIM model (i.e., J^h in Figure 5) by learning an MR tree approximation. Next, a sparse in-scale conditional covariance Σ^c is learned by solving a convex optimization problem similar to (7), but before this step, we compute the target information matrix (for the full process across *all* scales) which plays the same role as J^* in (7).

Step 1. Learning the inter-scale model J^h To begin with, we select an MR tree structure (without any in-scale connections) with *additional hidden* variables at coarser scales and the original variables at the finest scale. For some processes, there exists a natural hierarchical structure: for example, the MR tree models in Figure 1 for regular one-dimensional or two-dimensional processes, and the hierarchy defined by the Standard Industrial Classification (SIC) system for the stock returns. For problems in which the hierarchical structure is not clearly defined, any clustering algorithm can be applied to group variables together and insert one coarser scale variable per group. Once the structure is fixed, we apply the EM algorithm to choose the parameters that best match the given target covariance Σ_F^* for the finest scale variables. This procedure is efficient for a tree-structured model and converges to a local maximum.

Step 2: Finding the target information matrix J^* From Step 1, we have an information matrix J_{tree} corresponding to an MR tree model. Note that J_{tree} has a structure as in Figure 5 and thus can be written as $(J^h + J^c)$

except that J^c is a diagonal matrix. This diagonal in-scale conditional structure results in artifacts that correspond to inaccurate matching of finest-scale covariances, so we fix J^h and modify J^c in the remaining steps. The goal of this step is to compute the target information matrix $J^* = J^h + J^{c*}$ so that the finest scale submatrix of $(J^*)^{-1}$ is exactly equal to the given target covariance Σ_F^* . In other words, we design a matrix J^{c*} such that $(J^h + J^{c*})$ becomes an “exact” target MR model in which the marginal covariance at the finest scale equals the given target covariance Σ_F^* . We describe the detailed computation in the Appendix (see also (Choi et al., 2009)).

Step 3: Obtaining sparse in-scale conditional covariance Consider the target information matrix computed from Step 2: $J^* = J^h + J^{c*}$. The inter-scale part J^h is a tree model but J^{c*} is not sparse and does not have a sparse inverse (i.e., $\Sigma^{c*} \equiv (J^{c*})^{-1}$ is *not* sparse). We find a SIM model that approximates J^* by solving the following problem:

$$\hat{J} = \underset{J \succ 0}{\operatorname{argmax}} \quad \sum_m \log \det J_{[m]}$$

$$\text{s.t.} \quad |J_{i,j} - J_{i,j}^*| \leq \gamma_{i,j}, \quad \forall \{i, j\} \in \mathcal{E}_{inscale}$$

$$J_{i,j} - J_{i,j}^* = 0 \quad \forall \{i, j\} \in \mathcal{E}_{inter} \quad (8)$$

where $J_{[m]}$ is the in-scale information matrix at scale m and $\mathcal{E}_{inscale}$ and \mathcal{E}_{inter} are the set of all possible in-scale and inter-scale edges, respectively. If we look at the terms involving scale m (i.e., elements of the matrix $J_{[m]}$), the above problem maximizes the log-determinant of $J_{[m]}$ subject to element-wise constraints. Therefore, as in Section 5.1, 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 problem in (8) is convex and can be efficiently solved using general techniques for convex optimization (Löfberg, 2004). The regularization parameter $\gamma_{i,j}$ is chosen by a heuristic method (see (Choi et al., 2009)).

6. Experimental Results

In this section, we present the modeling and inference performance of our SIM model. The results are compared with a single-scale approximate model where we learn a sparse graphical model using (6) without introducing hidden variables, a tree-structured MR model, and a sparse MR model of the form introduced in (Choi & Willisky, 2007) that has sparse graphical model structure at each scale. We measure the modeling accuracy of approximate models by computing the divergence between the specified target distribution and the approximate distribution learned.³

³For MR models we use the marginal distribution at the finest scale to compute this divergence.

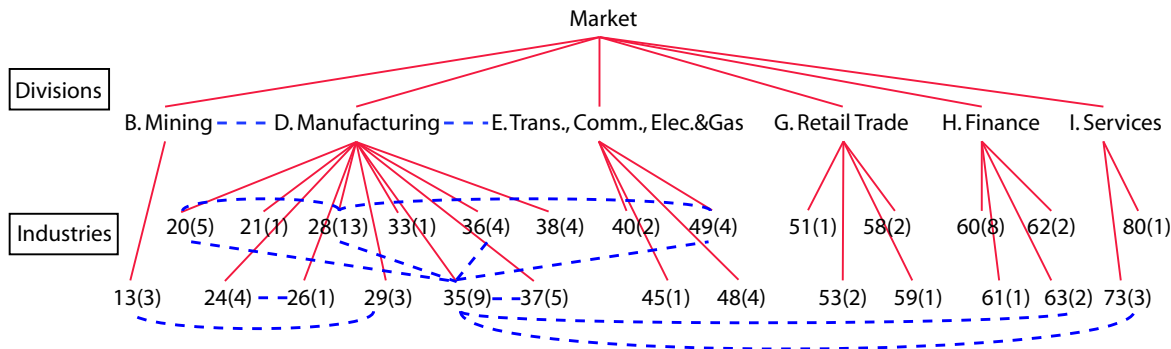


Figure 6. The structure of the SIM model approximation for Stock data.

Table 1. Top 4 strongest conjugate edges at Scale 3 of Figure 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 |

6.1. Stock Returns

Our first experiment is modeling the dependency structure of monthly stock returns. We compute the empirical covariance using the monthly returns from 1990 to 2007, and learn a SIM model approximation for the 84 companies in the S&P 100 stock index⁴ using the hierarchy defined by the Standard Industrial Classification (SIC) system.⁵ Our MR models have 4 scales, representing the market, 6 divisions, 26 industries, and 84 individual companies, respectively, at scales from the coarsest to the finest.

Figure 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 Figure 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 1 shows 4 conjugate edges at Scale 3 in the order of their absolute magnitude (i.e., the top 4 strongest in-scale conditional covariance).

Figure 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 that belong to the same industry, because there is no hidden variable to capture the correlations at a coarser resolution. Figure 7(b) shows the information matrix at the finest scale of a sparse MR model approximation (Choi & Willsky, 2007). Although the graphical model is sparser than the single-scale approximation, some of the companies still have densely connected edges. This suggests that the SIM model structure is a more natural representation for capturing in-scale statistics. As shown in the caption of Figure 7, the SIM model approximation provides the smallest divergence of all approximations.

6.2. Fractional Brownian Motion

We consider fractional Brownian motion (fBm) with Hurst parameter $H = 0.3$ defined on the time interval $(0, 1]$ with the covariance function: $\Sigma(t_1, t_2) = \frac{1}{2}(|t_1|^{2H} + |t_2|^{2H} - |t_1 - t_2|^{2H})$. Figure 8 shows the covariance realized by each model using 64 time samples. Our SIM approximation in Figure 8(d) is close to the original covariance in Figure 8(a), while the single-scale approximation in Figure 8(b) fails to capture long-range correlations and the tree model covariance in Figure 8(c) appears blocky.

Fig. 9(a) displays a 256-point sample path using the exact statistics and (b) displays noisy observations of (a), which are only available on $(0, 1/3]$ and $(2/3, 1]$. Fig. 9 (c-e) show the estimates based on the approximate single-scale model, the MR tree model (with 5 scales), and the SIM model, respectively, together with the optimal estimate based on the exact statistics. The estimate based on our SIM model approximation is close to the optimal estimate and does not

⁴We disregard 16 companies listed on S&P 100 after 1990.

⁵http://www.osha.gov/pls/imis/sic_manual.html

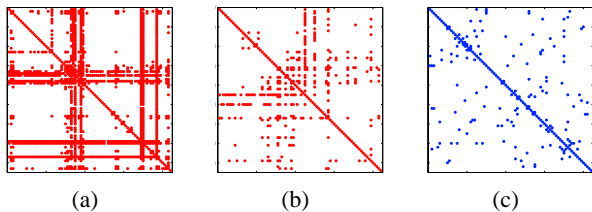


Figure 7. Stock returns modeling example. Sparsity pattern of the information matrix of (a) the single-scale (122.48), and (b) the sparse MR approximation (28.34). (c) Sparsity pattern of the in-scale conditional covariance of the SIM approximation (16.36). All at the finest scale. We provide the divergence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

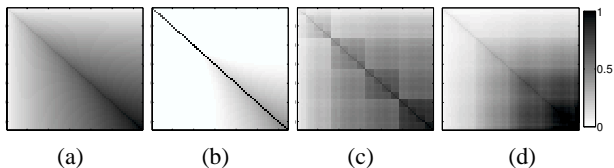


Figure 8. Covariance for fBm-64. (a) Original model, (b) Single-scale approximation, (c) Tree approximation, (d) SIM model.

have blocky artifacts unlike the estimate based on the MR tree model. The sparse MR model of (Choi & Willsky, 2007) does not lead to blocky artifacts either, but we observe that the SIM model can achieve a smaller divergence with a smaller number of parameters than the sparse MR model (see Table 2). Note that the number of parameters (number of nodes plus the number of (conjugate) edges) in the SIM model is much smaller than in the original model and in the approximate single-scale model.

6.3. Polynomially Decaying Covariance for a 2-D Field

We consider a collection of 256 Gaussian random variables arranged spatially on a 16×16 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 $\Sigma_{x_s, x_t} = d(s, t)^{-\frac{1}{2}}$, where $d(s, t)$ is the spatial distance between nodes s and t . Such processes with polynomially-decaying covariance have long-range correlations (unlike processes with exponentially-decaying covariance), and are usually not well-modeled by a single-scale sparse graphical model. 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. 10 shows the *conjugate* graph of the SIM model approximation within each 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.

We generate random noisy measurements using the specified statistics and compare the computation time to solve

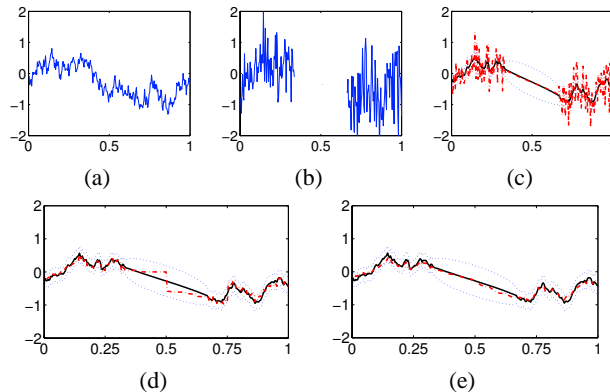


Figure 9. Estimation for fBm-256. (a) Sample-path using exact statistics. (b) Noisy and sparse observations of (a), Estimation using (c) single-scale approximation, (d) tree model, and (e) SIM model are shown in the dash-dot red lines, with the optimal estimate based on exact statistics in the solid black line. The dashed blue line shows plus/minus one standard deviation error bars.

Table 2. FBm-256 approximation

| | #var | #param* | Div. | RMS** |
|------------|------------|-------------|-------------|---------------|
| 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 |

* # nodes + # graphical or conjugate edges

** root-mean-square error w.r.t. the optimal estimate

the inference problem for the SIM model (using the inference algorithm in Section 4.2), the original and the single-scale approximate model (using the ET algorithm described in Section 4.1), and the sparse MR model (using the algorithm in (Choi & Willsky, 2007)). The SIM modeling approach provides a significant gain in convergence rate over other models as displayed in Figure 11.

7. Conclusion and Future Work

We propose a method to learn a Gaussian MR model with sparse in-scale conditional covariance at each scale and sparse inter-scale graphical structure connecting variables across scales. By decomposing the information matrix of the resulting MR model into a sparse matrix (information matrix corresponding to inter-scale graphical structure) and matrix that has a sparse inverse (in-scale conditional covariance), we develop an efficient inference algorithm that exploits sparsity in both the Markov and covariance structure. Our learning algorithm first learns a good MR tree model that approximates the given target covariance at the finest scale and then augments each scale with a sparse conjugate graph using a convex optimization procedure based on log-determinant maximization. While our focus in this paper is on the Gaussian model, applying similar principles to discrete models is also of interest, and under investigation.

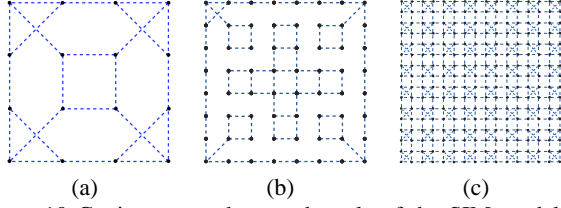


Figure 10. Conjugate graph at each scale of the SIM model for polynomially decaying covariance approximation. (a) Scale 3 (4×4), (b) Scale 4 (8×8), (c) Scale 5 (16×16).

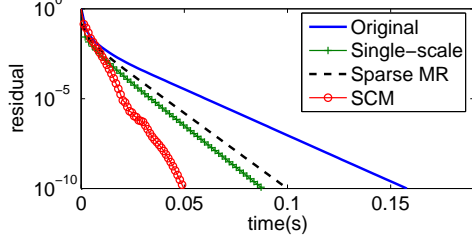


Figure 11. Residual error vs. computation time to solve the inference problem in the polynomially decaying covariance example.

Appendix. Computing J^* in Section 5.2

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

$$\Sigma_{[m]} = A_m \Sigma_{[m+1]} A_m^T + Q_m \quad (9)$$

where A_m and Q_m are determined by J_{tree} .⁶ Since we wish to find an MR model such that the covariance matrix at the finest scale becomes Σ_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 (9).

Let $J^* = J_{tree}$. We modify J^* in a *coarse-to-fine* way to match the target marginal covariance at each scale as obtained above (9). Suppose that we have replaced $J_{[1]}^*$ through $J_{[m-1]}^*$, and let us consider computing $J_{[m]}^*$. We partition J^* into 9 submatrices with the information matrix at scale m at the center:

$$J^* = \begin{pmatrix} J_c^* & J_{c,[m]}^* & 0 \\ J_{[m],c}^* & J_{[m]}^* & J_{[m],f}^* \\ 0 & J_{f,[m]}^* & J_f^* \end{pmatrix} \quad (10)$$

In order to set the marginal covariance at scale m equal to the target covariance matrix $\Sigma_{[m]}$ in (9), we replace $J_{[m]}^*$ in (10) with the following matrix

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

and proceed 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 (Banerjee et al., 2006), so computing $J_{[m]}^*$ is not a bottleneck of the learning process.

⁶Let $B_m = (J_{tree})_{[m-1],[m]}^{-1}$ and $D_m = (J_{tree})_{[m]}^{-1}$. Then, $A_m = B_m D_m^{-1}$ and $Q_m = D_{m-1} - B_m D_m^{-1} B_m^T$.

Acknowledgments

We thank Prof. Hui Chen for discussions about the stock returns example. This research was supported in part by AFOSR through Grant FA9550-08-1-1080, in part under a MURI through AFOSR Grant FA9550-06-1-0324, and in part by Shell International Exploration and Production, Inc. M. J. Choi was partially funded by a Samsung Scholarship.

References

- Banerjee, O., El Ghaoui, L., d'Aspremont, A., & Natsoulis, G. (2006). Convex optimization techniques for fitting sparse Gaussian graphical models. *International Conference on Machine Learning (ICML)* (pp. 12–18).
- Bouman, C. A., & Shapiro, M. (1994). A multiscale random field model for Bayesian image segmentation. *IEEE Transactions on Image Processing*, 3, 162–177.
- Choi, M. J., Chandrasekaran, V., & Willsky, A. S. (2009). Gaussian multiresolution models: Exploiting sparse Markov and covariance structure. *MIT LIDS Technical Report #2806*.
- Choi, M. J., & Willsky, A. S. (2007). Multiscale Gaussian graphical models and algorithms for large-scale inference. *IEEE Stat. Signal Proc. Workshop* (pp. 229–233).
- Cox, D. R., & Wermuth, N. (1996). *Multivariate dependencies: Models, analysis and interpretation*. Chapman & Hall/CRC.
- Golub, G. H., & Van Loan, C. H. (1990). *Matrix computations*. Baltimore, MD: The Johns Hopkins Univ. Press.
- Greengard, L., & Rokhlin, V. (1987). A fast algorithm for particle simulations. *Journal of Computational Physics*, 73, 325–348.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford, U.K.: Oxford University Press.
- Löfberg, J. (2004). Yalmip : A toolbox for modeling and optimization in MATLAB. *IEEE Computer-Aided Control System Design (CACSD) Conference* (pp. 284–289).
- Osindero, S., & Hinton, G. (2007). Modeling image patches with a directed hierarchy of Markov random fields. *Neural Information Processing Systems (NIPS)* (pp. 1121–1128).
- Sudderth, E. B., Wainwright, M. J., & Willsky, A. S. (2004). Embedded Trees: Estimation of Gaussian processes on graphs with cycles. *IEEE Transactions on Signal Processing*, 52, 3136–3150.
- Willsky, A. S. (2002). Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90, 1396–1458.