MULTISCALE GAUSSIAN GRAPHICAL MODELS AND ALGORITHMS FOR LARGE-SCALE INFERENCE

Myung Jin Choi and Alan S. Willsky

Massachusetts Institute of Technology
Electrical Engineering and Computer Science
77 Massachusetts Ave., Cambridge, MA 02139, USA

ABSTRACT

We propose a class of multiscale graphical models and algorithms to estimate means and approximate error variances of large-scale Gaussian processes efficiently. Based on emerging techniques for inference on Gaussian graphical models with cycles, we extend traditional multiscale tree models to pyramidal graphs, which incorporate both inter- and intra- scale interactions. In the spirit of multipole algorithms, we develop efficient inference methods in which variables far-apart communicate through coarser resolutions and nearby variables interact at finer resolutions. In addition, we propose methods to update the estimates rapidly when measurements are added or new knowledge of a local region is provided.

Index Terms— graphical models, Gauss-Markov random fields, multiresolution, multiscale, large-scale estimation problems

1. INTRODUCTION

The multiscale, or multiresolution modeling framework [1] has attracted much attention in the signal and image processing community for its rich modeling power as well as computational efficiency. Traditional multiscale models use tree-structured graphs (Figure 1 (bottom left)), which provide extremely powerful and efficient algorithms, but have limited modeling power that may lead to blocky artifacts. Other approaches, motivated by *multigrid methods*, use multiple-scale algorithms for computational efficiency but do not have consistent stochastic structures between different scales. These limitations have been recognized by a number of researchers, who consider models that incorporate both intra- and inter-scale interactions [1], [2]. However, due to the resulting model complexity, they either allow only a limited extension of multiscale trees or use computationally expensive methods such as simulated annealing to get solutions.

In recent years, there have been significant advances in understanding and developing efficient inference algorithms for a larger class of Gaussian graphical models [3], [4]. Thanks to these emerging techniques, it is no longer required to limit the graph structure to trees in order to obtain tractable inference algorithms. This paper presents a pyramidal graph in which consistent statistical links exist between neighbors at each scale as well as between adjacent scales. We develop highly efficient algorithms motivated by *multipole methods* [5] to compute the optimal estimates as well as uncertainties of the estimates given noisy measurements at some of the nodes. In addition, using the consistent graphical structure of our model, the estimates can be updated rapidly when measurements are added or new knowledge of a local region (for example, existence of discontinuities in the field) is provided. The problem of fitting the model to best explain the given data is also addressed and simulation results are presented.

2. GAUSS-MARKOV RANDOM FIELDS

A Gaussian random process x can be represented by a graph $\mathcal{G}$ consisting of nodes V and edges $\mathcal{E}$. Each node s is associated with a random variable[1] x_s, and edges connecting the nodes capture the statistical dependencies among the random variables. The pdf of a Gaussian process x, parameterized by its mean μ and covariance matrix P, can be equivalently represented in *information form* $J = P^{-1}$, and $h = P^{-1}\mu$. The inverse covariance matrix J is *sparse with respect to* $\mathcal{G}$: a nonzero off-diagonal element in matrix J indicates the presence of an edge linking the corresponding nodes.

Consider a sparse noisy observation vector $y = Cx + v$, where $v \sim \mathcal{N}(0, R)$ is a Gaussian white noise process. The conditional distribution of x is $p(x|y) \propto \exp(-\frac{1}{2}x^T Jx + x^T h)$, where $J = J_{prior} + C^T R^{-1} C$ and $h = h_{prior} + C^T y$. The optimal estimates and error covariance matrix can, in principle, be computed as

$$\begin{aligned} \hat{x} &= \arg\max p(x|y) = E[x|y] = J^{-1}h & (1) \\ \hat{P} &= E[(x-\hat{x})(x-\hat{x})^T|y] = J^{-1}. & (2) \end{aligned}$$

Research supported in part by Samsung Scholarship, Shell International Exploration and Production, Inc., and by AFOSR under Grant FA9550-04-1-0351.

[1] All analysis in the paper can be easily extended to the case when x_s is a random vector.

1-4244-1198-X/07/$25.00 ©2007 IEEE

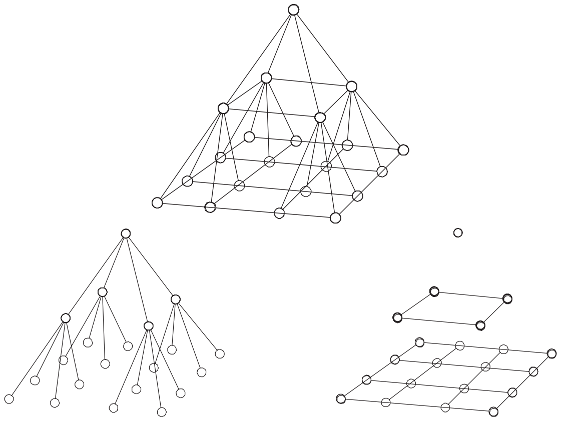

Fig. 1. (top) A pyramidal graphical model for two-dimensional processes, and its decomposition into (bottom left) a quadtree and (bottom right) nearest-neighbor grids.

2.1. Inference using tractable subgraphs

For problems with a large number of variables, the matrix inversion in (1) and (2) becomes intractable. Tree-structured graphs provide efficient linear complexity algorithms to compute both conditional means and error variances [1], but have limited modeling capabilities. For inference on graphs with cycles, *embedded subgraph algorithms* [3] utilize tractable subgraphs such as trees or subgraphs consisting of a small number of variables at each iteration to solve (1). For *walk-summable models* [3], it can be proven that the iterations converge for any sequence of subgraphs as long as each edge and node is updated infinitely often. This allows us to choose subgraphs adaptively for each iteration to reduce estimation error quickly as possible.

The *Lagrangian Relaxation (LR) method* [4] decomposes an intractable graph explicitly into tractable subgraphs and uses the estimates in each subgraph to perform approximate inference for the entire graph. At each iteration, nodes and edges shared by a set of subgraphs exchange potentials to match marginal statistics. For the Gaussian case, this algorithm converges to the true conditional means and gives upper bounds on the variances.

3. PYRAMIDAL GRAPHS

The convergence rate of iterative inference algorithms can be significantly improved by introducing auxiliary variables that represent the field of interest at coarser resolutions. Although the pyramidal graph we are proposing here can easily incorporate data or user objectives at multiple resolutions, we focus on the case that the coarser scales are merely acting to help inference at the finest scale. Let's assume that the field of interest is two-dimensional and originally can be described at a single resolution. We construct a pyramidal graphical model shown in Figure 1 (top) by placing the original field at the bottom of the hierarchy and introducing hidden variables at coarser scales. Unlike multigrid methods and the models considered in [2], the measurements are not replicated at coarser scales. We denote the coarsest scale in our pyramidal graph as *Scale 1* and the finest scale as *Scale M*.

Suppose that the field we are estimating is smooth overall, with the possible exception of a few discontinuities. The *thin-membrane model* penalizes the differences between the neighboring nodes: $p(x) \propto \exp(-\alpha \sum_{i \in V} \sum_{j \in \mathcal{N}(i)} (x_i - x_j)^2)$, where $\mathcal{N}(x_i)$ is the set of neighboring nodes of i, and α is a parameter that controls the strength of constraints. We extend this thin-membrane model to define prior in the pyramidal graph, which consists of two components: $J_{prior} = J_t + J_s$. The quadtree structure in Figure 1 (bottom left) is represented by J_t, which imposes the constraint that each parent node has a value close to its children. J_s corresponds to the nearest neighbor grid model for each scale as shown in the bottom right plot and imposes smoothness within each scale. Without loss of generally, we assume that $h_{prior} = 0$. As long as all parameters are nonnegative, it can be easily shown that the pyramidal graph is walk-summable.

The resulting marginal covariance at the finest scale has long-range correlations compared to its monoscale counterpart thanks to coarser scale variables. However, the conditional correlation of one scale, conditioned on adjacent scales, decays fast since long-range correlations are captured by coarser scale nodes [6]. This indicates that far-field effects can be well approximated at coarser scales, and each fine scale can only compute interactions among nearby nodes. A similar approximation technique is used in multipole methods [5].

4. MULTIPOLE-MOTIVATED INFERENCE ALGORITHMS

4.1. Computation of estimates and re-estimates

The optimal estimates on the pyramidal graph can be computed iteratively using a tractable subgraph at each iteration. We first develop a simple algorithm in which the order of inference steps follows the spirit of multipole algorithms, and then extend the idea to a more sophisticated algorithm that selects subgraphs adaptively.

The multipole-motivated inference algorithm starts by getting rough estimates at all nodes in the pyramidal graph using only the J_t component in our prior model:

$$\hat{x}^{(0)} = (J_t + C^T R^{-1} C)^{-1} h.$$

When the coarsest scale of the pyramidal graph has multiple nodes, the J_s component at the coarsest scale is also included in this initial step to get globally consistent estimates[2]. Then,

[2] The number of variables at the coarsest scale is significantly smaller than that of the finest scale, so we assume that exact inference within the coarsest scale is tractable.

we alternate between the *in-scale inference step* (equivalent to a coarse-to-fine sweep) and the *tree inference step* (a fine-to-coarse sweep) until convergence. Let $\bar{x}^{(n)}$ and $\hat{x}^{(n)}$ denote the estimates computed at the n^{th} in-scale and tree iteration, respectively. We use the notation $J_{[i,j]}$ to represent the submatrix of J corresponding to scale i and scale j, and x_m to represent the subvector of x corresponding to scale m.

In the in-scale inference step, we decompose (1) by scale: $J_{[m,m]}\hat{x}_m = h_m - J_{[m,m-1]}\hat{x}_{m-1} - J_{[m,m+1]}\hat{x}_{m+1}$. Then, starting from the coarsest scale ($m = 1$) and proceeding downward, the nodes at scale m are updated using the just-computed estimates at its coarser neighbor, $m-1$, and the previous tree-inference estimates at the next finer scale, $m+1$. Exact inference is not tractable for scales with a large number of nodes, so $J_{[m,m]}$ is again decomposed into J_a, which corresponds to a tractable subgraph embedded in the grid model at scale m, and $K_a = J_{[m,m]} - J_a$. Then, this inference step is equivalent to computing the following equation:

$$\begin{aligned}\bar{x}_m^{(n)} &= J_a^{-1}(h_m - K_a\hat{x}_m^{(n-1)} \\ &\quad -J_{[m,m-1]}\bar{x}_{m-1}^{(n)} - J_{[m,m+1]}\hat{x}_{m+1}^{(n-1)})\end{aligned}$$

Utilizing the multipole idea, we choose J_a to be $diag(J_{[m,m]})$, a diagonal matrix with entries taken from $J_{[m,m]}$, which corresponds to a fully disconnected graph at each scale. Then, this step is essentially applying a single Gauss-Jacobi iteration within each scale.

In the tree inference step, the quadtree(s) connecting different scales[3] is used as a tractable subgraph. Although it is sufficient for this step to pass messages upward, to facilitate convergence analysis in terms of embedded subgraph algorithms, we pass messages both upward *and* downward to perform exact inference on the quadtree(s). Let J_n be defined as the associated J matrix corresponding to the quadtree(s): $J_n = J_t + diag(J_s) + C^T R^{-1} C$. Then, the tree inference step using the quadtree structure can be represented as

$$\hat{x}^{(n+1)} = J_n^{-1}(h - K_n\bar{x}^{(n)})$$

where $K_n = J - J_n$.

Instead of using the fixed subgraphs as described above, the *adaptive Embedded Tree (ET) algorithm* [3] can also be applied to each iteration step. For the tree-inference step, a spanning tree of the pyramidal graph is selected to minimize the residual $h - J\hat{x}^{(n)}$, and similarly for the in-scale inference step, a spanning tree in each grid model can be adaptively chosen. It has been observed that alternating the adaptive in-scale and tree inference steps results in much faster convergence than applying the adaptive ET algorithm to the pyramidal graph without any guidance of its hierarchical structure. Note that from the walk-summability of the pyramidal graph, both non-adaptive and adaptive iterations are guaranteed to converge [3].

Assume that we already have solved an estimation problem based on a large number of measurements, and then wish to modify the estimates to account for new local information. We refer this problem as *re-estimation*, which can arise in two possible scenarios. The first case is when a new set of measurements are introduced in a local region. The second case is modifying the prior model locally to weaken the smoothness constraints across surface discontinuities so that high-frequency components can be recovered. For either case, the re-estimation problem can be posed as the following: given the estimates $\hat{x} = J^{-1}h$, compute the updated estimates $\tilde{x} = (J + \Delta J)^{-1}(h + \Delta h)$, where ΔJ and Δh have nonzero elements only in a localized area.

The re-estimation problem can be solved iteratively by updating a subset of variables at each iteration. Let S denote the region at the finest scale in which changes have been made, i.e. in which either ΔJ or Δh is nonzero. Also, let $\mathcal{T}_S$ denote the set of (disjoint) quadtrees, each of which is rooted at a single node at the coarsest scale and has non-empty intersection with the nodes in S. Our algorithm alternates between tree inference iterations on $\mathcal{T}_S$ and the *adaptive block Gauss-Seidel iterations* [3] in order to choose a subset of variables to be updated. The latter steps provide rapid estimate adjustments, primarily at finer scales and in the vicinity of S, while the tree inference steps propagate these estimates more broadly across the field.

4.2. Computation of variances

The diagonal elements of the error covariance matrix P correspond to the uncertainties in the estimates at each node. Note that by decomposing the pyramidal graph into the quadtree(s) and separated vertical and horizontal chains within each scale, the LR method [4] can be applied to compute not only the optimal estimates but also upper bounds on error variances.

Alternatively, we may use the fact that error variances conditioned on adjacent scales decay fast and thus can be efficiently computed using the *low-rank approximation algorithm* [7]. In [6], we derive a set of equations for approximate variances which can be iterated using coarse-to-fine sweeps. Let V_m be the set of nodes at scale m, and let $\bar{p}_{ij}$ be the variance between $i \in V_m$ and $j \in V_m$ conditioned on the adjacent scales. Then, the approximate variance of i computed at the n^{th} iteration is given by

$$\sigma_i^{(n)} = \bar{p}_{ii} + \sum_{j,k \in (\mathcal{N}(i) \cup \{i\}) \cap V_m} \bar{p}_{ij} \cdot \bar{p}_{ik} \cdot (\tilde{Q}_m^{(n)})_{jk} \qquad (3)$$

where $\tilde{Q}_m^{(n)}$ is defined as

$$\tilde{Q}_m^{(n)} = J_{[m,m-1]}\Sigma_{[m-1]}^{(n)} J_{[m-1,m]} + J_{[m,m+1]}\Sigma_{[m+1]}^{(n-1)} J_{[m+1,m]}$$

$\Sigma_{[m]}^{(n)}$ is a diagonal matrix with each element corresponding

[3] Note that when the coarsest scale have multiple nodes, the quadtree structure is a set of disjoint quadtrees, each of which has a root node at the coarsest scale.

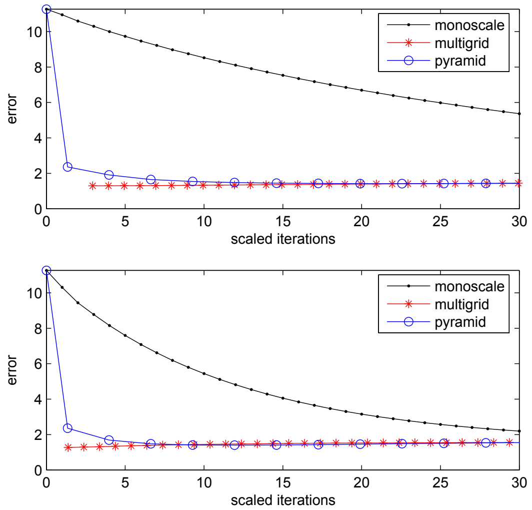

Fig. 2. The convergence of RMS errors in surface estimation. The horizontal axes are in units of equivalent monoscale iterations. (top) Non-adaptive iterations. (bottom) Adaptive iterations.

to approximate variances of variables at scale m computed at the n^{th} coarse-to-fine sweep.

It can be proven that the approximate variances in (3) provide lower bounds on the true error variances [6]. The lower bounds closely approximate the true values as long as the conditional correlations decay fast. However, for some models with sparse measurements, even conditional correlations may have relatively slow decay. An alternative that provides accurate variances even in such cases is the *wavelet-based low-rank approach* [7]. The structure of the pyramidal graph allows efficient and simple implementation of the wavelet-based algorithm.

4.3. Parameter estimation

In order to fit our pyramidal graph to best explain the given data, parameters can also be estimated from the measurements. Here, we consider estimating two parameters: φ that controls the strength of the smoothness constraints and γ, the reciprocal of the measurement noise variance. Let J_P the prior matrix J_{prior} with a unit parameter value, then $J = \varphi J_P + \gamma C^T C$. The tractable methods for the computation of estimates and variances allow us to derive an efficient EM algorithm to estimate the parameters.

In the E-step, the expected values of potential functions are evaluated using the conditional means $\hat{x}^{(n-1)}$ and error variances $\hat{P}^{(n-1)}$ computed from the parameters estimated at iteration $(n-1)$:

$$\begin{aligned}\eta_1 &= tr(J_P \hat{P}^{(n-1)}) + (\hat{x}^{(n-1)})^T J_P \hat{x}^{(n-1)} \\ \eta_2 &= \parallel y - C\hat{x}^{(n-1)} \parallel^2 + tr(C\hat{P}^{(n-1)}C^T)\end{aligned}$$

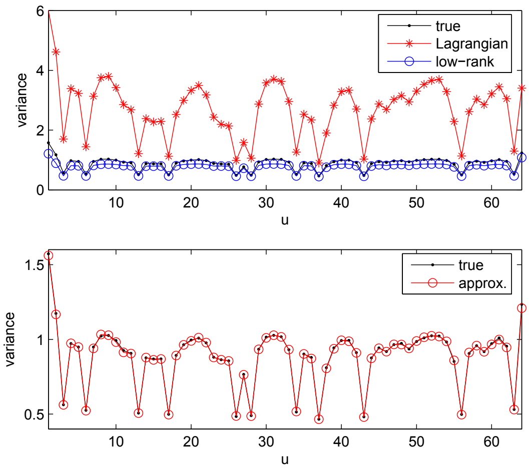

Fig. 3. A cross section of approximate variances computed by (top) the LR method and the coarse-to-fine low-rank algorithm, (bottom) the wavelet-based low-rank algorithm.

Due to the sparsity of J_P and C, we only need the variances of individual nodes and covariances between the pairs of neighboring nodes to compute both values [4].

The M-step, leads to the following simple expressions for the next parameter estimates:

$$\varphi^{(n)} = \frac{N}{\eta_1} \qquad \gamma^{(n)} = \frac{N_{meas}}{\eta_2}$$

where N and N_{meas} are the number of nodes and measurements, respectively.

5. SIMULATION RESULTS

We test our multipole algorithm on a synthetic surface of size 64 × 64 variables in which noisy ($\sigma^2 = 1$) measurements are available only at randomly chosen 10% of the variables. Figure 2 shows the convergence of RMS errors for the non-adaptive (top) and adaptive (bottom) algorithms on the pyramidal graph, together with the corresponding multigrid and monoscale algorithms. The multigrid algorithm uses the estimates at coarser versions of the problem to guide inference at finer scales, and the monoscale algorithm applies Gauss-Jacobi (non-adaptive) or adaptive iterations directly on a single-scale thin-membrane model. The monoscale algorithm converges much slowly than the pyramidal graph which achieves performance comparable to multigrid methods after a few iterations. Note that due to the lack of consistent stochastic structure, it is not straightforward to estimate error variances or to solve the re-estimation problem using multigrid methods.

Figure 3 (top) shows one cross section of the bounds on variances of the synthetic surface. The upper bounds show estimates computed by the LR method, and lower bounds are

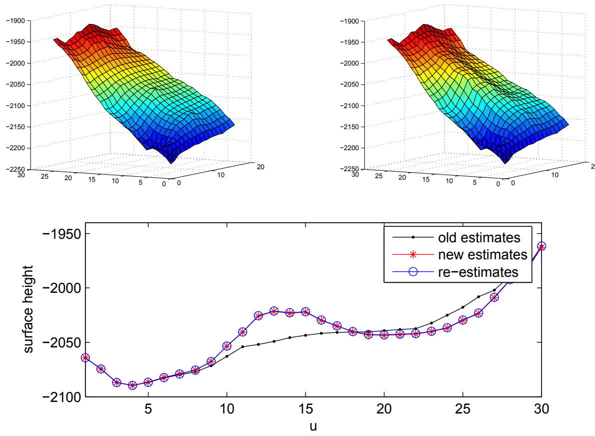

Fig. 4. Re-estimation applied to the problem of updating estimates to incorporate a new set of measurements in a local region. (top left) Estimates before adding measurements. (top right) Re-estimates. (bottom) A cross section of re-estimates.

computed by applying 5 coarse-to-fine sweeps of the low-rank approximation method. The upper bounds obtained by the LR method are rather loose, but they follow the shape of the true variances, and note that the bounds are obtained while computing the optimal estimates without any additional cost. Figure 3 (bottom) shows the variances estimated by the wavelet-based low-rank methods. It can be observed that the estimates are close to the true variances.

Next, we apply the re-estimation algorithm to a real problem: estimating the top surface of a large salt deposit located below the sea floor of Gulf of Mexico. The measurements, provided by Shell International Exploration, Inc., consist of $377,384$ picks by analysts interpreted from seismic data. After estimating the surface heights using a pyramidal graph with four scales, we introduce 100 new measurements in a small region. Figure 4 (top right) shows the re-estimates of the local region after 10 iterations of the re-estimation algorithm, which shows more detailed surface delineations compared to the estimates before adding the measurements (top left). The bottom plot shows one cross section of the re-estimates. To compare the performance, the figure also shows the updated estimates using a naive method: after modifying J and h to model the new measurements, we simply perform inference on the entire pyramid. Using this naive implementation, 3 million nodes are updated at each iteration. The re-estimation algorithm updates less than 1000 nodes at each iteration, yet after 10 iterations, they converge to the same result.

Lastly, Figure 5 shows the estimation results of φ using 5 sets of measurements generated from different sizes of pyramidal graphs. As the number of nodes grows larger, the estimate of φ converges to the correct value.

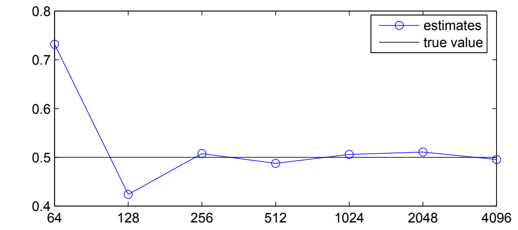

Fig. 5. Parameter φ estimated from 5 sets of measurements generated by the pyramidal graph. The x-axis show the number of nodes at the finest scale of the pyramidal graph.

6. CONCLUSION

In this paper, we have introduced a class of multiscale Gaussian graphical models defined on pyramidal lattices, and developed efficient algorithms for inference problems. Our algorithms take advantage of the fact that long-range correlations are well approximated at coarser scales, and alternates global propagation of information using an embedded spanning tree of the pyramidal graph and local computations within each scale. The hierarchical structure of our model also leads to efficient methods to modify an estimated field when local changes are made to the prior model or to the available data.

7. REFERENCES

[1] A. S. Willsky, “Multiresolution Markov models for signal and image processing,” *Proc. of IEEE*, vol. 90, no. 8, pp. 1396–1458, 2002.

[2] Z. Kato, M. Berthod, and J. Zerubia, “A hierarchical Markov random field model and multitemperature annealing for parallel image classification,” *Graphical Models and Image Proc.*, vol. 58, no. 1, pp. 18–37, 1996.

[3] V. Chandrasekaran, J. K. Johnson, and A. S. Willsky, “Estimation in Gaussian graphical models using tractable subgraphs: A walk-sum analysis,” *IEEE Trans. Signal Proc.* (accepted).

[4] J. K. Johnson, *Recursive Variational Methods for Approximate Inference in Graphical Models*, Ph.D. thesis, MIT, 2007, (in preparation).

[5] L. Greengard and V. Rokhlin, “A fast algorithm for particle simulations,” *J. Comp. Physics*, vol. 73, no. 2, pp. 325–348, 1987.

[6] M. J. Choi, “Multiscale Gaussian graphical models and algorithms for large-scale inference,” S. M. thesis, MIT, May 2007.

[7] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, “GMRF variance approximation using spliced wavelet bases,” in *Proc. ICASSP*, Apr. 2007.