Low-Rank Variance Approximation in GMRF Models: Single and Multiscale Approaches

Dmitry M. Malioutov, *Student Member, IEEE*, Jason K. Johnson, Myung Jin Choi, *Student Member, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—We present a versatile framework for tractable computation of approximate variances in large-scale Gaussian Markov random field estimation problems. In addition to its efficiency and simplicity, it also provides accuracy guarantees. Our approach relies on the construction of a certain low-rank aliasing matrix with respect to the Markov graph of the model. We first construct this matrix for single-scale models with short-range correlations and then introduce spliced wavelets and propose a construction for the long-range correlation case, and also for multiscale models. We describe the accuracy guarantees that the approach provides and apply the method to a large interpolation problem from oceanography with sparse, irregular, and noisy measurements, and to a gravity inversion problem.**

***Index Terms*—Approximate variances, Gaussian Markov random fields, multiscale models, wavelets.**

I. Introduction

Markov random fields (MRFs) [1]–[3] are statistical models defined on undirected graphs, where the nodes in the graph correspond to random variables and the edges encode conditional independence relations between pairs of variables. The Markov property of an MRF generalizes the well-known property for Markov processes on a chain, stating that the past and the future are conditionally independent given the present. In MRFs on general graphs, if a set of nodes separates the graph into two disconnected components, then these two components are conditionally independent given the separator. Gauss–Markov random fields (GMRF) are MRFs where the variables are jointly Gaussian. The Markov graph of a GMRF is dictated by the sparsity structure of the inverse of the covariance matrix.

We address estimation in large-scale GMRFs, which arise in a wide variety of applications including computer vision, sensor networks, geostatistics, and oceanography [1], [3], [4]. Assuming that the prior and the observation model are fully specified, this involves computing the estimates and the variances of the hidden variables (note that we are computing these quantities from the model—we *are not* estimating them from samples). A prototypical application is interpolation from sparse, irregular, noisy measurements [4]. As GMRFs are a subclass of jointly Gaussian models, both the estimates (means) and the variances can be obtained via matrix inversion. However, for large-scale problems—arising, for example, in oceanography and seismic imaging with two-dimensional (2-D) or three-dimensional (3-D) fields with millions of variables—exact matrix inversion becomes intractable. Owing to the sparsity of the graph, approximate means can be computed with linear complexity in the number of nodes using iterative solvers such as preconditioned conjugate gradients or multigrid [5], [6]. However, such methods do not provide the variances. The aim of this paper is to develop approaches to find accurate approximate variances—in essence, this is a problem of approximating the diagonal of the inverse of a sparse positive definite matrix.

Variances are a crucial component of estimation, giving the reliability information for the means. They are also useful in other respects: regions of the field where residuals exceed error variances may be used to detect and correct model-mismatch (for example, when smoothness models are applied to fields that contain abrupt edges). Also, as inference is an essential component of learning a model (for both parameter and structure estimation), accurate variance computation is needed when designing and fitting models to data. Another use of variances is to assist in selecting the location of new measurements to maximally reduce uncertainty.

Exact variances can be computed in tree-structured models using belief propagation [7] (which, in trees, corresponds to sparse Gaussian elimination) with linear complexity in the number of nodes. For general models, junction-tree extensions of belief propagation reduce the complexity of exact inference from cubic in the number of variables to cubic in the "tree-width" of the graph [2]. For square and cubic lattice models with N nodes, this leads to complexity $O(N^{3/2})$ and $O(N^2)$, respectively, which, despite being a great improvement from brute-force matrix inversion, is still not scalable for large models.[1] In addition, junction-tree algorithms are quite involved to implement. Approximate methods such as loopy belief propagation (LBP) [8] have linear complexity per iteration but are not guaranteed to converge, and convergence

Manuscript received July 24, 2007; revised May 21, 2008. First published June 20, 2008; current verion published September 17, 2008. The associate editor coordinating the review of this paper and approving it for publication was Dr. Ilya Pollak. This work was supported by the Army Research Office under Grant W911NF-05-1-0207 and the Air Force Office of Scientific Research under Grant FA9550-04-1-0351.

The authors are with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: dmm@mit.edu; jasonj@mit.edu; myungjin@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.2008.927482

[1] A recent method, recursive cavity modeling (RCM) [4], provides tractable computation of approximate variances using a combination of junction-tree ideas with recursive model-thinning. The method in this paper, however, provides analytical guarantees of accuracy, which have not been established for RCM, and is also much simpler in terms of implementation.

may be slow for large problems. Even in case of convergence, LBP may produce rather poor variance approximations.

We propose a simple framework for variance approximations that provides theoretical guarantees of accuracy. In our approach, we use a low-rank aliasing matrix to compute an approximation to the inverse $J^{-1} = P$. By designing this matrix, such that only the weakly correlated terms are aliased, we are able to give provably accurate variance approximations. We propose a few different constructions for the low-rank matrix. We start with a design for single-scale models with short correlation length and then extend it to single-scale models with long correlation length using a wavelet-based aliasing matrix construction. GMRFs with slow correlation decay, e.g., fractional Gaussian noise, are often better modeled using multiple scales. Thus we also extend our wavelet-based construction to multiscale models, in essence making both the modeling and the processing multiscale. This paper builds upon our earlier short conference publications [9], [10].

In Section II, we discuss estimation with GMRF models and also mention multiscale modeling. We introduce our low-rank variance approximation approach, apply it to short-correlation models, and establish accuracy guarantees in Section III. We then describe the spliced-wavelet extension for models with long correlations length in Section IV, also extending the accuracy guarantees. In Section IV-C, we apply the construction to multiscale models. We describe efficient ways of solving the linear system arising in our approach in Appendix B. In Section V, we test our approach with experiments, including estimation problems from oceanography and gravity inversion.

II. GMRF Models

A GMRF model is based on a jointly Gaussian density with certain conditional independence relations, which are summarized by an undirected graph $\mathcal{G} = (V, \mathcal{E})$. Here V is a set of vertices (also called nodes), and $\mathcal{E} \subset \binom{V}{2}$ is a set of edges (unordered pairs of vertices). It is convenient to specify a GMRF model in *information form*

$$p(x) \propto \exp\left\{-\frac{1}{2}x^T Jx + h^T x\right\}. \quad (1)$$

The matrix J is called the information matrix and is symmetric positive definite $(J \succ 0)$ and sparse so as to respect the graph $\mathcal{G}$: if $\{i,j\} \notin \mathcal{E}$, then $J_{ij} = 0$. We call h the potential vector. These quantities are directly related to the usual parameterization of Gaussian densities in terms of the mean $\mu \equiv \mathbb{E}[x]$ and the covariance matrix $P \equiv \mathbb{E}[(x-\mu)(x-\mu)^T]$

$$\mu = J^{-1}h \quad \text{and} \quad P = J^{-1}. \quad (2)$$

Sparsity of J is the link between the graph structure and the conditional independence (Markov) properties of the GMRF: $J_{ij} = 0$ implies that x_i is independent of x_j given the other variables. For $A \subset V$, we define x_A to be the vector $(x_i | i \in A)$ corresponding to the variables in A, and we use $V \backslash A$ to denote the complement of A. Let $N(i) = \{j | \{i,j\} \in \mathcal{E}\}$ denote the *neighbors* of i in the graph. Then the Markov property can be stated as

$$p(x_i | x_{V\backslash i}) = p(x_i | x_{N(i)}), \quad \forall\, i. \quad (3)$$

1) Examples:

Consider an estimation problem with a *thin-membrane prior,* commonly used for data interpolation

$$p(x) \propto \exp\left(-\frac{\alpha}{2}\sum_{\{i,j\}\in\mathcal{E}} (x_i - x_j)^2\right). \quad (4)$$

This prior enforces leveled fields, i.e., it favors that neighbors should have similar values. Note that without any observations, this prior is degenerate (nonintegrable), as any constant field x has the same probability. This degeneracy disappears once observations are added (or with small regularization $\gamma \sum_i x_i^2$), which we assume throughout this paper. The J matrix can be readily deduced from (4): $J_{ij} = 0$ for $i \neq j$ with $\{i,j\} \notin \mathcal{E}$, $J_{ij} = -\alpha$ for $\{i,j\} \in \mathcal{E}$, and $J_{ii} = \alpha d_i$. Here d_i is the degree of node i, $d_i = |N(i)|$. Another common prior in image processing is the *thin-plate* prior

$$p(x) \propto \exp\left(-\frac{\alpha}{2}\sum_{i\in V}\left(x_i - \frac{1}{d_i}\sum_{j\in N(i)} x_j\right)^2\right). \quad (5)$$

The thin-plate prior enforces that each node is close to the average of its neighbors[2] and penalizes curvature.

We can easily incorporate local observations y_i, with Gaussian $p(y_i|x_i)$. Assume that y_i is independent of x_j and other y_j for $j \neq i$: $p(y|x) = \prod_{i\in V} p(y_i|x_i)$. The posterior is now $p(x|y) \propto p(y|x)p(x)$, which is a GMRF that is Markov on the same graph (ys are observed and do not change, so we do not add new nodes). Adding local observations modifies the diagonal of J and the potential vector h.

For a concrete example, consider the linear Gaussian problem, with observations $y = Hx + n$, where x is zero mean with covariance P_{prior} and independent noise n is zero mean and with diagonal covariance Q. Then the Bayes least squares estimate $\mu_{x|y}$ and the error variance $P_{x|y}$ are given by

$$\left(P_{\text{prior}}^{-1} + H^T Q^{-1} H\right)\mu_{x|y} = H^T Q^{-1} y$$
$$P_{x|y} = \left(P_{\text{prior}}^{-1} + H^T Q^{-1} H\right)^{-1}. \quad (6)$$

If $J_{\text{prior}} = P_{\text{prior}}^{-1}$ is a sparse GMRF prior on x and y are local observations, then $J_{x|y} = \left(P_{\text{prior}}^{-1} + H^T Q^{-1} H\right)$ has the same sparsity as J_{prior}, with only the diagonal terms being modified. Now $J_{x|y}$ and $h_{x|y} = H^T Q^{-1} y$ are the information parameters specifying the conditional model given the observations.

Given a model in information form specified by (J, h), it is of interest to compute the (conditional) means μ and the variances P_{ii} for all i. As we have discussed in Section I for large-scale GMRFs, exact computation of P using matrix inversion is intractable. In Section III, we describe our approach.

A. Multiscale GMRF Models

Single-scale models with only local interactions, such as thin membrane and thin plate models, have limitations on the kind of fields they represent. In particular, the tails of the

[2] A thin plate model on a square grid has a more dense Markov graph: neighbors up to two steps away are connected by an edge.

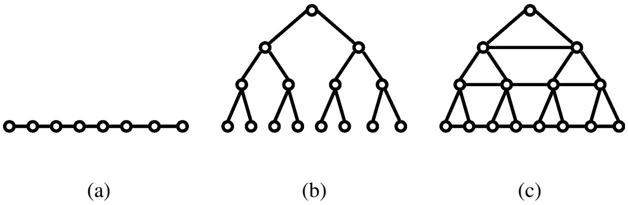


Fig. 1. (a) Single-scale model. (b) Tree-structured multiscale model. (c) Loopy multiscale model on a pyramidal graph.

correlation for such models fall off exponentially fast, so to represent long-range correlations with slower decay, other models are needed. One can certainly accomplish this by using far denser single-scale graphs, with long-range interactions, but this defeats the sparsity needed for efficient algorithms. An alternative is to make use of multiscale models, which represent the phenomenon of interest at multiple scales or resolutions. Coarser scales correspond to local aggregates of finer scales: coarse-scale variables capture summaries of local regions at the finer scale. The multiple scales may represent physically meaningful quantities with measurements acquired at different scales. Alternatively, coarser scales may be artificially introduced hidden variables without measurements, which facilitate more efficient estimation. The scales may be disjoint, with estimates in coarser scales used to simplify estimation in the finer scales [5], [11], or they may be linked together into a coherent statistical model, with either deterministic or stochastic interactions between scales [12]–[14]. A significant effort has been devoted to the development of extremely efficient tree-structured [see Fig. 1(b)] multiscale models [12]. The main drawback of tree-structured models is that certain neighbors in the fine-scale model may become quite distant in the tree-structured model, which leads to blocky artifacts in the estimates. To avoid these artifacts, multiscale models that allow loops also received attention, e.g., [13] and [14]. We consider a class of multiscale models on pyramidal graphs with loops described in [13] and [15]. The different scales in this model constitute a coherent statistical model with nondeterministic interscale interactions.

The Markov graph for the model is illustrated in Fig. 1(c). We show each scale to be one-dimensional, but they can also be two- and three-dimensional. The model has a pyramidal structure including interactions within the scale and between neighboring scales. The model has many small loops, so exact methods for tree-structured graphs do not apply, but the model is much richer representationally than tree-structured ones. The motivation for this multiscale model is to represent or approximate a single-scale model with slow correlation decay. The correlations in the single-scale model get distributed among scales in the multiscale model, and the long correlations are mostly accounted for through coarse-scale interactions. Conditioned on the coarse-scale variables, the conditional correlations among the fine-scale variables are more local. We leave the question of learning such pyramidal models to [13] and describe an extension of our low-rank variance approximation to find variances when such a model is specified in Section IV-C.

III. Low-Rank Variance Approximation

Finding the means of a GMRF corresponds to solving the linear equations $J\mu = h$. For sparse graphs, a variety of efficient, iterative algorithms exist for solving such equations with total complexity that grows roughly linearly with the number N of nodes in the graph (see Appendix B for details) [6]. However, except for models on trees, such linear complexity is not readily available for the computation of the covariance matrix. One way in which one might imagine performing this computation is to embed it in a set of N linear equation solvers. Let $v_i \in \mathbb{R}^N$ be the ith standard basis vector; then the ith column of P can be obtained by solving $JP_i = v_i$. To get all N columns of P, this would have to be done N times, once at each node in the graph: $JP = [v_1, \ldots, v_N] = I$ with complexity $O(N^2)$. This is still intractable for large-scale models. Note that the full P matrix has N^2 elements, so quadratic complexity is a lower bound to compute all of P.

However, in many cases, we are most interested only in the diagonal elements P_{ii} of P (i.e., the individual variances),[3] and this raises the question as to whether we can compute or approximate these elements with procedures with only linear complexity. Of course the direct computation $\text{diag}(P) = \text{diag}(J^{-1}I)$ is costly. Instead we propose to design a low-rank matrix BB^T, with $B \in \mathbb{R}^{N\times M}$ and $M \ll N$, and use it instead of I. The system $J\hat{P} = BB^T$ can be solved with $O(MN)$ complexity in two steps: first we solve $JR = B$ using iterative solvers. Then, we postmultiply R by B^T, i.e., $\hat{P}_{ii} = [RB^T]_{ii}$ (which requires MN operations, as we only need the diagonal).

To get accurate variance approximations, B must be designed appropriately, taking the graph and the correlation structure of the model into consideration. Let all rows b_i of B have unit norm $b_i^T b_i = 1$. Consider the diagonal of $\hat{P} = J^{-1}(BB^T)$

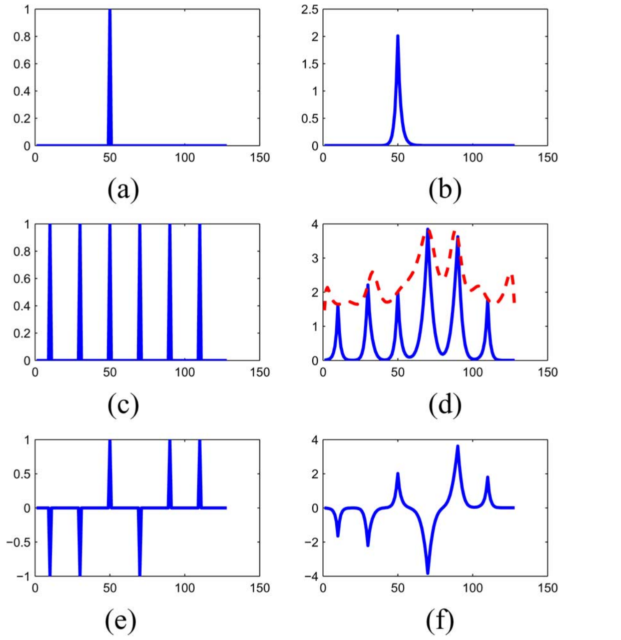
$$\hat{P}_{ii} \triangleq [J^{-1}(BB^T)]_{ii} = P_{ii} + \sum_{i\neq j} P_{ij} b_i^T b_j. \tag{7}$$

To force $\hat{P}_{ii}$ to be accurate approximations of the variances we need the aliased terms $P_{ij} b_i^T b_j$ to be nearly zero for all pairs of nodes. We analyze two different cases. For models with short-range correlations P_{ij} decays fast and is nearly zero for most pairs, so we only have to take care of the nearby nodes. In the long-range correlation case, we use a wavelet decomposition to decompose the correlation across several scales, thus producing several problems with short correlation length. Moreover, by adding randomness to the choice of B (and perhaps computing approximations with several such random choices), we can obtain unbiased approximations of the true covariances. We describe the short-range correlation construction next, and a wavelet-based extension for models with long correlations in Section IV.

A. *Constructing B for Models With Short Correlation*

The key idea here is that to make $P_{ij} b_i^T b_j$ small, we need either P_{ij} or $b_i^T b_j$ to be small. Suppose that P_{ij} decays fast with

[3] It is also possible to use our approach to find accurate approximations of the elements of the covariance which correspond to nearby nodes. For sparse models, there are $O(N)$ such elements.



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

Fig. 2. An illustration of the low-rank matrix construction for a 1-D chain. (a) Single spike v_i results in (b) fast-decaying response $J^{-1}v_i$. Next, in (c) we add together several well-separated spikes, $z = \sum_{i \in c} v_i$, and in (d) show the resulting response $J^{-1}z$, which at the peaks is close to the correct variances P_{ii} (dashed). Next, we introduce random sign-flips σ_i. In (e), we plot $B_c = \sum_{i \in c} \sigma_i v_i$. In (f), we show the response $R_c = J^{-1}B_c$.

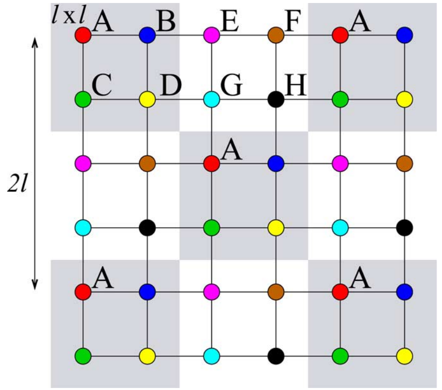
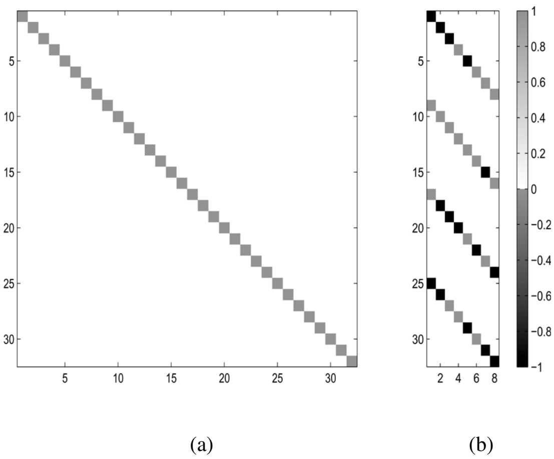


Fig. 3. Local 2×2 regions for square lattice. Colors: $\{A, \ldots, H\}$ with first four colors in shaded blocks and last four colors in transparent blocks. The blocks appear in a checkerboard pattern.

distance from node i to j. Then, for nodes that are far apart in the graph (farther than the correlation length[4]) the correlation P_{ij} and the corresponding error-terms in (7) are negligible. For pairs of nodes i and j that are nearby, we have to design B such that b_i and b_j are orthogonal: this is a problem of designing an overcomplete basis $\{b_i \in \mathbb{R}^M\}$ that is nearly orthogonal with respect to a graph $\mathcal{G}$. We describe such a construction for chains and rectangular lattices, and suggest an approach for arbitrary sparse graphs.



(a) (b)

Fig. 4. (a) Identity and (b) locally orthogonal B matrix formed by adding certain columns of I together and changing signs randomly.

For the sake of clarity, we start with a simple 1-D chain example.[5] We assume that the correlation between nodes decays rapidly with distance (e.g., in many models correlation decays exponentially with distance $d(i,j)$ between i and j: $|P_{ij}| \leq A\beta^{d(i,j)}$ with $0 \leq \beta < 1$). Consider Fig. 2(a) and (b). We plot the ith standard basis vector v_i with $i = 50$ in (a) and the ith column P_i of P, the solution to the system $JP_i = v_i$ in (b). There is a spike of P_{ij} at $j = i$, a fast decaying response for j near i, and most of other entries are nearly zero. Now let $z = v_{i_1} + v_{i_2} + \cdots v_{i_K}$, where all indexes' i_ks are mutually well separated. In Fig. 2(c) and (d), we show z and the solution w to $Jw = z$. We also show P_{ii} (dashed). At each i_k, we have a spike and a fast-decaying response. This operation can be seen as a convolution of a spike-train with a time-varying kernel. If the spikes are well- separated, then the interference from other spikes is small and $w_{i_k} \approx P_{i_k,i_k}$ for each k. This is the basic idea behind the construction of our B matrix for the short-range correlation case.

Now to find such groups of well-separated nodes, we partition the nodes into classes, which we call *colors,* such that nodes of the same color are a distance M apart. For chains, this can be done simply by periodically cycling through the M colors. We will have a column B_c of B for each color c. We assign $B_c(i) = \sigma_i = \pm 1$ independent identically distributed (i.i.d.) random signs for each node i of color c, and $B_c(j) = 0$ for other nodes. An illustration appears in Fig. 2(e) (and Fig. 4). We assign random signs to entries of B_c in order to have destructive interference between the error terms, and we later show that it leads to unbiased variance approximations. In Fig. 2(e), we plot a column B_c of B, and in (f) $R_c = J^{-1}B_c$. Next we apply B_c^T, thus selecting the entries for nodes of color c. After repeating these steps for all the colors and adding them together, we get our approximation $\hat{P}$.

For rectangular-grid models, the idea is very similar. We partition the nodes into several color classes such that nodes of the same color have a certain minimum distance between them. One

[4] We define the correlation length to be a distance in the graph beyond which the correlation coefficient between any two nodes becomes negligible (smaller than some specified threshold). For models with exponential decay, this is consistent with the conventional definition but also applies to models with other modes of correlation decay.

[5] Here we consider a chain in a generalized sense, meaning that the nodes have an inherent 1-D ordering but the Markov graph does not have to be a chain and may have links a few steps away in the ordering.

such construction with eight colors appears in Fig. 3. By off-setting the blocks in a checkerboard pattern, the minimum distance can be increased to twice the dimension of each square. The rest of the procedure is the same as in the 1-D case: we assign $B_c(i)$ to be ± 1 randomly (i.i.d. flips of a fair coin) for each node i of color c and solve $JR_c = B_c$ for all c.

For chains and lattices, the nodes are easy to color by inspection. For arbitrary sparse graphs, we suggest to use approximate graph-coloring to define B. To get a minimum distance l, one could augment the graph by connecting nodes up to l steps away and solve the graph-coloring problem on it (assigning colors such that nodes of the same color do not share an edge). Finding an optimal coloring is very hard, but approximate solutions (allowing for some violations, and using more than the minimum number of colors) can be approached using spectral methods [16] or the max-product form of belief propagation. Upon defining the colors, we can follow the same steps as we have described for chains and grids.

Next we analyze the diagonal elements of $\hat{P}$ and show that they are unbiased and that the errors can be made arbitrarily small by increasing the minimum separation.

B. Properties of the Approximation $\hat{P}$

Our construction of B can be viewed as aliasing of the columns of the standard basis I: groups of columns that correspond to nodes of the same color are added together. We refer to this process as *splicing*, see Fig. 4 for illustration. It can be represented as $B = IC$. Here the cth column C_c contains nonzero entries only for nodes of color c. The exact covariance P is the solution to linear system $JP = I$. We approximate it by solving $J\hat{P} = BB^T = ICC^TI$, i.e., $\hat{P} = J^{-1}CC^T$, and the error is

$$E = \hat{P} - P = J^{-1}(CC^T - I). \tag{8}$$

The matrix $(CC^T - I)$ serves the role of a signed adjacency matrix, showing which pairs of columns of I are aliased together. Let $\mathcal{C}(i)$ be the set of nodes of the same color as i; then

$$(CC^T - I)_{i,j} = \begin{cases} \sigma_i \sigma_j, & \text{if } i \in C(j), j \neq i \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

We are interested in the diagonal entries of E:

$$\begin{aligned} E_{ii} &= \left(P(CC^T - I)\right)_{ii} = \sum_j P_{ij}(CC^T - I)_{ji} \\ &= \sum_{j \in C(i)\backslash i} \sigma_i \sigma_j P_{ij} = P_i^T \delta_{C(i)\backslash i}. \end{aligned} \tag{10}$$

The term $\delta_{C(i)\backslash i}$ is a signed indicator of the components aliased to i, i.e., $\delta_{C(i)\backslash i}(j) = \sigma_i \sigma_j = \pm 1$ if $j \in C(i)\backslash i$, and zero otherwise.

1) Unbiased: The approximations $\hat{P}_{ii}$ are unbiased. The expectation of $\hat{P}$ over $\{\sigma_i\}$ is $\mathbb{E}_\sigma[\hat{P}_{ii}] = P_{ii} + \mathbb{E}_\sigma[E_{ii}]$. We have $\mathbb{E}_\sigma[E_{ii}] = \sum_{j \in C(i)\backslash i} P_{ij} \mathbb{E}_\sigma[\sigma_i \sigma_j] = 0$,as σ_i and σ_j are independent and zero mean. Hence $\mathbb{E}_\sigma[\hat{P}_{ii}] = P_{ii}$. We stress that unbiasedness involves averaging over choices of σ. However, if the variance of $\hat{P}$ is small, then even one sample σ provides accurate approximations $\hat{P}$.[6]

2) Variance of the Approximations: Suppose that the correlations P_{ij} fall off exponentially with the distance $d(i,j)$ between i and j, i.e., $|P_{ij}| \leq A\ \beta^{d(i,j)}$, with $0 \leq \beta < 1$. This is true for a wide class of models including Markov models on bipartite graphs. Now, $\mathrm{Var}_\sigma(\hat{P}_{ii}) = \mathbb{E}_\sigma\left[(\hat{P}_{ii} - P_{ii})^2\right] = \mathbb{E}_\sigma\left[E_{ii}^2\right] = \mathbb{E}_\sigma\left[\left(\sum_{j \in C(i)\backslash i} \sigma_i \sigma_j P_{ij}\right)^2\right]$. We have

$$\begin{aligned} \mathrm{Var}_\sigma(\hat{P}_{ii}) &= \mathbb{E}_\sigma\left[\left(\sum_{j \in C(i)\backslash i} \sigma_i \sigma_j P_{ij}\right)^2\right] \\ &= \sum_{j,j' \in C(i)\backslash i} \mathbb{E}_\sigma\left[\sigma_i^2 \sigma_j \sigma_{j'}\right] P_{ij} P_{ij'} \\ &= \sum_{j \in \mathcal{C}(i)\backslash i} P_{ij}^2. \end{aligned} \tag{11}$$

In the second line, we use the fact that $\sigma_i^2 = 1$ and that $\mathbb{E}_\sigma\left[\sigma_j \sigma_{j'}\right] = 1$ if $j = j'$, and zero otherwise.

In a 2-D lattice model with our construction, the number of nodes of a given color that are $(2l)n$ steps away is $8n$ (all the distances between nodes of the same color are integer multiples of $2l$). Using the exponential decay bound, for nodes j with $d(i,j) = 2nl$, $P_{ij} \leq A\ \beta^{2nl}$. Hence

$$\sum_{j \in \mathcal{C}(i)\backslash i} P_{ij}^2 \leq \sum_{n=1}^{\infty} 8nA^2 \beta^{4nl} = 8A^2 \frac{\beta^{4l}}{(1 - \beta^{4l})^2}. \tag{12}$$

We have used the following series: $\sum_{n=1}^{\infty} n\beta^n = (\beta/(1-\beta)^2)$. Thus, $\mathrm{Var}_\sigma(\hat{P}_{ii}) \leq 8A^2\left(\beta^{4l}/(1 - \beta^{4l})^2\right)$. Since $|\beta| < 1$, we can choose l large enough such that the variance of the approximation is below any desired threshold. In practice, l should be chosen to be comparable to the correlation length of the model.

Now let us repeat the analysis for 2-D lattices with a slower, power-law rate of decay, i.e., $P_{ij} \leq A\ d(i,j)^{-p}$, where $p > 0$. Then the sum in (12) changes to

$$\sum_{j \in \mathcal{C}(i)\backslash i} P_{ij}^2 \leq A^2 \sum_{n=1}^{\infty} \frac{8n}{(4nl)^{2p}} = \frac{8A^2}{(4l)^{2p}} \sum_n n^{1-2p}. \tag{13}$$

If $p > 1$, then the sum $\sum_n n^{1-2p}$ converges (and is equal to $\zeta(2p-1)$, the Riemann zeta function), and the errors can be made arbitrarily small by increasing l. However, if $p \leq 1$, then for any l, the sum diverges.[7] In Section IV, we show that the wavelet-based construction can dramatically reduce the errors for such power-law decay and can go beyond these limitations.

We can also bound the absolute error itself (rather than its variance): $|E_{ii}| \leq \sum_{j \in \mathcal{C}(i)\backslash i} |P_{ij}|$. For example, with exponential decay of P_{ij}, we have $|E_{ii}| \leq 8A\left(\beta^{2l}/(1 - \beta^{2l})^2\right)$. The stochastic bound in (12) is tighter, but it requires taking expectation over the random signs σ.

[6] If the correlation decay is not known, then repeated trials over σ can also provide empirical variances of the approximation.

[7] Here we are focusing on 2-D models. More generally, the required p depends on the dimension of the lattice. In d dimensions, there are $O(n^{d-1})$ aliased terms at distance n, and the sum in (13) becomes $\propto \sum n^{(d-1)-2p}$. Thus, we need $p > d/2$ for convergence.

IV. Constructing Wavelet-Based B for Models With Long Correlation

In our construction of matrix B in the last section, we set the separation length between nodes of the same color to be comparable to the correlation length in the model. When the correlation length is short, the approach is very efficient. However, when the correlation length is long, the approach is no longer attractive: making the separation length long will make the computational complexity high. Alternatively, if we violate the correlation length and use a short separation, then the method still gives unbiased variance approximations, but the variance of these variance approximations becomes very high (see examples in Section V).

To address long-range correlations, we propose using wavelets to decompose the aliasing matrix B across several scales, so that the correlation length in each scale is short. Note that in this section, the GMRF model has just *one scale.* Multiple scales come from the wavelet decomposition. In Section IV-C, we apply the method to a *multiscale model,* where the GMRF has hidden variables representing coarser scales and allows sparse representation of processes with slow correlation falloff.

We start with one-dimensional wavelets in continuous time to simplify discussion and analysis. A wavelet decomposition is specified by a scaling function $\phi(t)$ and a wavelet function $\psi(t)$, which generate a family of dilations and translations [17]

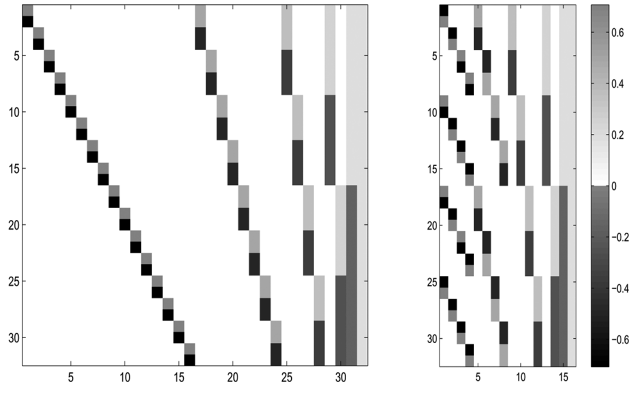
$$\phi_{s,k}(t) = \frac{1}{2^{s/2}}\phi(2^{-s}t-k)$$
$$\psi_{s,k}(t) = \frac{1}{2^{s/2}}\psi(2^{-s}t-k). \tag{14}$$

For a fixed scale s, the set $\{\phi_{s,k}(t)\}_k$ generates the approximation space $\mathcal{V}_s$. These spaces $\mathcal{V}_s$ are nested $\mathcal{V}_1 \supset \mathcal{V}_2 \supset \mathcal{V}_3 \ldots$, with higher s corresponding to coarser scales. The span of the wavelets $\{\psi_{s,k}(t)\}_k$ at a given scale s gives the detail space $\mathcal{W}_s = \mathcal{V}_{s-1} \ominus \mathcal{V}_s$ (we use $\ominus$ to denote the orthogonal complement of $\mathcal{V}_s$ in $\mathcal{V}_{s-1}$). We can decompose the fine scale $\mathcal{V}_1$ over N_{sc} scales

$$\mathcal{V}_1 = \mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \cdots \oplus \mathcal{W}_{N_{sc}} \oplus \mathcal{V}_{N_{sc}}. \tag{15}$$

We focus on orthogonal[8] wavelet families with compact support, where $\psi_{s,k}(t)$ is orthogonal to all other translations and dilations of $\psi(t)$ and to scaling functions at scale s and coarser.

To deal with discrete-time signals, we make the standard assumption that discrete samples f_k are the scaling coefficients $\langle \phi_{s_1,k}, f(t)\rangle$ of a continuous wavelet transform of some smooth function $f(t)$ at scale s_1[17]. Let $s_1 = 1$ without loss of generality. Now, a discrete wavelet basis for the space $\mathcal{V}_1$ is constructed by collecting the scaling functions at the coarsest scale and the wavelet functions at all finer scales as columns of a matrix W. Let S^s and W^s contain the scaling and wavelet functions, respectively, at scale s. In general, we do not need to go all the way to the coarsest scale $N_{sc} = \log_2(N)$. Stopping the decomposition earlier with $N_{sc} < \log_2(N)$ also provides an orthogonal basis for the space $\mathcal{V}_1$. Our orthogonal basis is[9]



(a) (b)

Fig. 5. (a) A discrete wavelet basis, with columns corresponding to wavelets at different scales and translations, and (b) B matrix obtained by aliasing certain columns of W within each scale. In the wavelet basis W, the number of columns doubles with each finer scale, but in B it stays constant.

$$W = [\,W^1 \quad W^2 \quad \ldots \quad W^{N_{sc}-1} \quad S^{N_{sc}}\,]. \tag{16}$$

An illustration of a Haar wavelet basis for $N = 32$ is given in Fig. 5(a). Columns (wavelets) are grouped by scale, and horizontal axis corresponds to translation. At scale s, we have $2^{N_{sc}-s}$ possible translations, and hence that many columns in W^s.

A. Wavelet-Based Construction of B

There is now a well-established literature [18]–[21] describing that, for many classes of random processes, their wavelet coefficients have faster decaying correlation than the original process itself. In our approach, we do not transform the random process—instead, we consider solutions R_k to $JR_k = W_k$ (W_k is a column of W) and show that R_k exhibits fast decay (we also say correlation decay), which will allow compression of B and computational efficiency. Roughly speaking, we create a scale-dependent B, with a construction similar to Section III at each scale. We now present our wavelet-based construction and then analyze it.

In the original single-scale construction, we find an approximation $\hat{P}$ to P by solving $J\hat{P} = BB^T$ instead of $JP = I$. The matrix B is an aliased version of I, with $B = IC$. For the multiscale construction, we start by expressing the exact covariance as the solution to the system $JP = WW^T = I$. We approximate it by applying the aliasing operation at each scale $B^s = W^sC^s$ (note, we do not alias wavelets across scales). We call this aliasing operation *wavelet splicing*. The kth column of W^s contains $\psi_{s,k}(t)$ and corresponds to the kth wavelet at scale s. We group these coefficients, and hence, the columns, into M^s groups (colors) such that any two coefficients of the same color are well separated with respect to the correlation length at scale s (i.e., correlation length for R_k at scale s). Each column of C^s

[8] One could also use biorthogonal wavelets [17] in our approach: instead of having an orthogonal wavelet basis W, we would have an analysis basis W_a and a synthesis basis W_s, such that $W_aW_s^T = I$.

[9] Ideally one would use boundary wavelets at the edges of the signal [17]. We do not pursue this: we use $N_{sc} < \log_2(N)$ and assume that the support of the wavelets at the coarsest scale N_{sc} is small compared to the size of the field, and hence edge-effects have negligible impact in our approach.

contains nonzero entries only for nodes of a particular color. Similar to Section III, we set $C_c^s(k) = \sigma_k^s = \pm 1$, for $k \in c$, and zero otherwise. The signs σ_k^s are equiprobable and i.i.d. Combining all the scales together, this gives

$$B = WC \tag{17}$$

where $B = [B^1, \dots B^{N_S}]$, $W = [W^1, \dots, W^{N_{sc}-1}, S^{N_{sc}}]$, and $C = \text{blockdiag}([C^1, \dots, C^{N_{sc}}])$. We illustrate matrices W and B in Fig. 5(a) and (b) respectively. The rest of the procedure follows that for the short correlation length: we solve for the diagonal of $\hat{P}$ using $J\hat{P} = BB^T$, as described in Section III.

In the wavelet decomposition, the majority of the coefficients are at fine scales. In the next section, we describe that for well-behaved GMRFs, R_k decays faster at finer scales.[10] While at the finer scales in W there are more coefficients (and columns), they can be aliased together more aggressively; see Fig. 5(b). We show that under certain assumptions, the correlation length can be assumed to decrease twofold with each finer scale, so the resulting number of columns of B^s stays the same for all scales. In this manner, the number of columns of B is $O(\log_2(N))$ instead of N for the wavelet basis W, giving significant computational savings in our approach.

1) Construction of B for 2-D: We use the separable wavelet construction, which takes products of 1-D functions to create a family of two-dimensional triplets [17][11]

$$\begin{aligned}
\psi^{(1)}_{s;k_1,k_2}(x,y) &= \phi_{s,k_1}(x)\psi_{s,k_2}(y)\\
\psi^{(2)}_{s;k_1,k_2}(x,y) &= \psi_{s,k_1}(x)\phi_{s,k_2}(y)\\
\psi^{(3)}_{s;k_1,k_2}(x,y) &= \psi_{s,k_1}(x)\psi_{s,k_2}(y).
\end{aligned} \tag{18}$$

Stacking $\psi^{(i)}_{s;k_1,k_2}$ as columns of a matrix creates an orthogonal basis $\bar{W}$ for two-dimensional fields. To produce the corresponding aliasing matrix $\bar{B}$ as in (17), we first create one-dimensional spliced matrices $B^s = W^s C^s$ and $\tilde{B}^s = S^s C^s$ containing linear combinations of wavelet and scaling functions at each scale. Then we create triplets using columns of B^s and $\tilde{B}^s$ in the same manner as in (18).

B. Error Analysis

In Section III-A, we analyzed the errors in the single scale construction $E_{ii} = P_i^T \delta_{C(i)\backslash i}$. When the separation between nodes of the same color is smaller than the correlation length, the errors are significant (see Fig. 6). We will now justify why the wavelet construction can dramatically reduce the errors for models with long-range correlations.

The variance approximation in the wavelet-based construction of B is $\hat{P} = J^{-1}BB^T = J^{-1}WCC^TW^T$. The aliasing matrix C is block diagonal with a block for each scale. Let $R = J^{-1}W$. Its kth column $R_k = J^{-1}W_k$ is the response of the linear system $JR_k = W_k$ to the wavelet W_k. An illustration appears in Fig. 7. We show the response $R_k = PW_k$ for a wavelet W_k at two different scales $s = 6$ and $s = 5$. We also show the wavelets W_l that are aliased to k with dashed lines. It is clear that R_k decays much faster than P_i in Fig. 6. We discuss this decay in more detail later in this section. The regions where R_k and W_l overlap contribute to the errors in $\hat{P}_i$ for i falling in the support of W_l. The error is

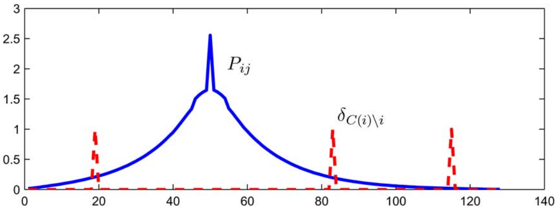


Fig. 6. Errors with the aliased standard basis: the error is obtained by an inner product between P_i and $\delta_{C(i)\backslash i}$ (both signals are a function of j). Here $i = 50$. We set all the signs $\sigma_j = 1$ for simplicity.

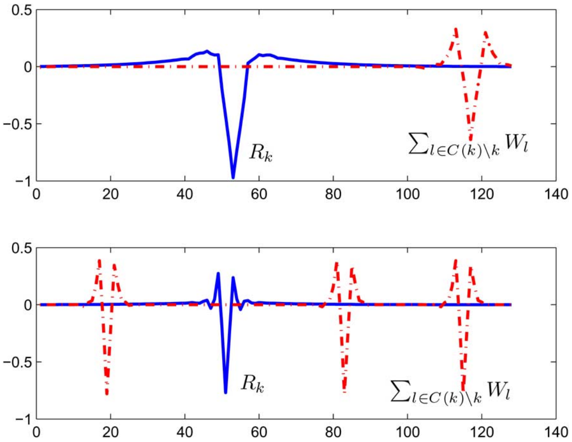


Fig. 7. R_k and W_l for $l \in C(k)\backslash k$. (Top) Scale 6. (Bottom) Scale 5. Regions where R_k and W_l overlap contribute to the errors in $\hat{P}_i$ for i in the support of W_l. The original P_i is shown in Fig. 6.

$$E = \hat{P} - P = J^{-1}W(CC^T - I)W^T = R(CC^T - I)W^T. \tag{19}$$

We have $(CC^T - I)_{k,l} = \sigma_k\sigma_l = \pm 1$ only if $k \neq l$ and the wavelets W_k and W_l are aliased together. In particular, if k and l belong to different scales, then $(CC^T - I)_{k,l} = 0$. Now the errors in variances are

$$\begin{aligned}
E_{ii} &= \sum_k \sum_l R_{ik}(CC^T - I)_{kl} W_{il}\\
&= \sum_k \sum_{l \in C^s(k)\backslash k} \sigma_k \sigma_l R_{ik} W_{il}.
\end{aligned} \tag{20}$$

We will analyze $\mathbb{E}_\sigma\left[E_{ii}^2\right]$ in Proposition 1 and show that the interference of R_k and W_l decays fast with separation. We will show that for large fields, as $N \to \infty$, the error is stable (i.e., bounded), the approximation can be made accurate to any desired level by controlling aliasing and that the multiscale construction is much more accurate than the single-scale one for GMRFs that have substantial energy over multiple scales.

[10] We measure distance and separation relative to scale: separation of K at scale s corresponds to separation of $K2^{s-1}$ at scale 1.

[11] This is different from taking outer products between each pair of columns of W in (16). That would also give an orthogonal basis but has the undesirable effect of mixing wavelets from different scales.

We can also bound $|E_{ii}|$ (and hence the ℓ_∞-norm of $e \triangleq \mathrm{diag}(E)$, $\|e\|_\infty = \max_i |E_{ii}|$)

$$|E_{ii}| = \left| \sum_k \sum_{l \in C^s(k) \setminus k} \sigma_k \sigma_l R_{ik} W_{il} \right| \leq \sum_k \sum_{l \in C^s(k) \setminus k} |R_{ik}||W_{il}|. \tag{21}$$

The tighter stochastic bound on $\mathbb{E}_\sigma\left[E_{ii}^2\right]$ in (24) involves expectation over σ, so these bounds are not redundant.

1) Correlation Decay: We now analyze the decay of $R_k(i)$. Note that, while our analysis is similar in spirit to other work involving wavelets and covariance matrices, our objectives and indeed our analysis differ in significant ways. In particular, conventional analysis focuses on the covariance matrix of the wavelet coefficients, i.e., $P_W = W^T P W$. In contrast, our analysis is based on viewing the rows of P as deterministic signals and considering their transforms—i.e., on the matrix $R = PW$. That said, we will comment on possible ties to more conventional wavelet analysis at the end of this section.

We first recall some relevant facts from wavelet analysis [17]. Suppose a continuous-time function $f(t)$ is α-Lipschitz[12] (this is related to how many times $f(t)$ is continuously differentiable). Also suppose that the wavelet family $\psi_{s,k}(t)$ has m vanishing moments,[13] with $m > \alpha$. Then the wavelet coefficients $Wf(s,k) = \langle \psi_{s,k}(t), f(t) \rangle$ satisfy $|Wf(s,k)| = O(2^{s(m+1/2)})$. If $m \geq 1$, then the magnitude of the wavelet coefficients in smooth regions drops fast for each finer scale.

However, this fast decay does not happen near a point of singularity of $f(t)$, say, t_0. Suppose that the wavelet at scale 1 has support K. At a coarser scale s, the support is $K2^{s-1}$. To avoid the point of singularity, the wavelet at scale s has to be outside the interval $t_0 \pm K2^{s-1}$, which gets twice as wide with each coarser scale. This set over all scales is called the "cone of influence," and it contains unusually high values of wavelet coefficients, a region of disturbance caused by the singular point [17].

For our analysis, we view P as samples of a continuous-time function and assume that the correlation function P_{ij} may have a singularity at $i = j$, and that it is smooth otherwise. Consider scale s, $R^s = PW^s$. The ith row of R^s contains the scale-s wavelet coefficients of the ith row of P. The singularity of P_{ij} at $i = j$ will produce a disturbance region with high wavelet coefficients near that value of k for which W_k peaks at row i. Recall that the rows of R^s are indexed by nodes, and the columns correspond to wavelet coefficients. The disturbance region at node i in R^s will be roughly $K2^s$ rows wide, and K columns wide (since wavelet coefficients involve downsampling by 2^s). When columns of W_s are aliased together, we have to make sure that the cones of influence do not overlap. The region of disturbance is twice as narrow (in terms of the number of rows) at each finer scale, so roughly twice as many wavelets can be aliased with each finer scale.

As an illustration, consider Fig. 7. The region of disturbance of $R_k(i)$ near $i = 50$ can be seen in Fig. 7 for scales 6 and 5. The original P_i is shown in Fig. 6 and has a singularity at $i = 50$. It is evident that by going to a finer scale, from $s = 6$ to $s = 5$, $R_k(i)$ decays faster, and more columns of W can be aliased without sacrificing the accuracy.

2) Properties of the Wavelet-Based Approximation $\hat{P}$: In the single-scale case, we showed that $\hat{P}$ is unbiased and bounded the variance of the errors. We extend these results to our wavelet-based approximation. The total error is equal to

$$E = P - \hat{P} = P\left(WW^T - BB^T\right). \tag{22}$$

3) Unbiased: Let $\mathcal{C}(k)$ be the set of columns that get merged with column k. Then taking an expectation over $\{\sigma_k\}$, $\mathbb{E}_\sigma[BB^T] = WW^T + \sum_k \sum_{l \in \mathcal{C}(k) \setminus k} W_k W_l^T \mathbb{E}_\sigma[\sigma_k \sigma_l] = WW^T$. The error terms cancel out because $\mathbb{E}_\sigma[\sigma_k \sigma_l] = 0$ for $k \neq l$. Thus, the approximation $\hat{P}$ is *unbiased.*

4) Variance of the Approximations: We now obtain a bound based on the expression in (20). Since $\hat{P}$ is unbiased, we have $\mathrm{Var}_\sigma(\hat{P}_i) = \mathbb{E}_\sigma\left[E_{ii}^2\right]$. Using (20), it follows:

$$\mathbb{E}_\sigma\left[E_{ii}^2\right] = \mathbb{E}_\sigma\left[\left(\sum_l \sum_{k \in C(l) \setminus l} \sigma_k \sigma_l R_{ik} W_{il}\right)^2\right]. \tag{23}$$

The terms $\sigma_k \sigma_l$ and $\sigma_{k'} \sigma_{l'}$ are uncorrelated unless $(k,l) = (k',l')$ or $(k,l) = (l',k')$, so this expectation reduces to $\mathbb{E}_\sigma\left[E_{ii}^2\right] = \sum_l \sum_{k \in C(l) \setminus l} R_{ik}^2 W_{il}^2 + \sum_l \sum_{k \in C(l) \setminus l} R_{ik} W_{il} R_{il} W_{ik}$. Also, the second term is zero, as we require that the supports of the aliased terms W_l and W_k do not overlap, i.e., $W_{il} W_{ik} = 0$ for $k \in C(l) \setminus l$. Hence

$$\mathbb{E}_\sigma\left[E_{ii}^2\right] = \sum_l \sum_{k \in C(l) \setminus l} R_{ik}^2 W_{il}^2. \tag{24}$$

To bound this sum, we consider a model with exponential and power-law decay of correlations and assume that the wavelet has m vanishing moments. Also, we *do not* use $N_{sc} = \log_2(N)$ scales in the decomposition but rather set $N_{sc} \propto \log_2(L)$, where L is the correlation length of the model. Once the size of the field exceeds L, there is no advantage in including coarser scales that contain negligible energy.

Proposition 1 (Bounded Errors): Suppose for a 1-D GMRF, $P_{ij} \sim \beta^{d(i,j)}$ or $P_{ij} \sim d(i,j)^{-p}$. Then, as the size of the field tends to infinity, the errors in (24) stay bounded, provided that the number of vanishing moments of the wavelet function satisfies $m \geq 1$. Also, by increasing the separation length, i.e., the distance between nearmost aliased terms, the errors can be made arbitrarily small.

We establish this stability property in Appendix A. We avoid the issue of boundary effects as we fix the number of scales of the wavelet decomposition when the field size tends to infinity. In the Appendix, we show that the errors in the wavelet-based construction can be much smaller than in the single-scale one if the GMRF has power distributed over multiple scales. For

[12] A function is pointwise α-Lipschitz [17] at t_0 if there exists $\gamma > 0$ and a polynomial p_v of degree $m = \lfloor \alpha \rfloor$ such that $\forall\, t \in \mathbb{R}$, $|f(t) - p_v(t)| \leq \gamma |t - t_0|^\alpha$, $(\alpha > 0)$. It is uniformly Lipschitz over an interval if it is pointwise Lipschitz with γ not dependent on t.

[13] A wavelet with n vanishing moments is orthogonal to polynomials of degree $n - 1$, i.e., $\int_{-\infty}^{\infty} t^k \psi(t) dt = 0$ for $0 \leq k < n$.

higher dimensional lattices with power-law rate of decay, the required number of vanishing moments also has to satisfy $m + p > (d/2)$, where d is the dimension.

5) Alternative Variance Analysis: We also consider another line of analysis that makes ties to covariances of wavelet coefficients $P_W^s \triangleq (W^s)^T P W^s$ (rather than $R^s = PW^s$). It is important to emphasize that this analysis is approximate and does not lead to bounds. Consider $\mathrm{tr}(E)$, and decompose it by scale. We have

$$\begin{aligned}\mathrm{tr}(E_s) &= \mathrm{tr}\left[P(W^s(W^s)^T - B^s(B^s)^T)\right]\\ &= \mathrm{tr}\left[(W^s)^T P W^s - (B^s)^T P B^s\right]\\ &= \mathrm{tr}\left[(W^s)^T P W^s - (C^s)^T (W^s)^T P W^s C^s\right]\\ &= \mathrm{tr}\left[P_W^s (I - C^s(C^s)^T)\right]. \end{aligned} \tag{25}$$

Then, via the same analysis as in Section III-B, we have

$$\mathrm{Var}_\sigma\left(\mathrm{tr}\left[P_W^s(I - C^s(C^s)^T)\right]\right) = \sum_k \sum_{l \in \mathcal{C}(k)\backslash k} ((P_W^s)_{k,l})^2. \tag{26}$$

Putting all the scales together, $\mathrm{Var}_\sigma(\mathrm{tr}(E)) = \sum_s \sum_k \sum_{l \in \mathcal{C}(k)\backslash k} \left((P_W^s)_{k,l}\right)^2$. This equality holds since the signs σ_k at different scales are independent. Now, assuming that the errors at different nodes are only weakly correlated, which we justify with experiments in Section V, we have $\sum_i \mathrm{Var}_\sigma(E_{ii}) \approx \mathrm{Var}_\sigma(\sum E_{ii}) = \mathrm{Var}_\sigma(\mathrm{tr}(E)) = \sum_s \sum_k \sum_{l \in \mathcal{C}(k)\backslash k} \left((P_W^s)_{k,l}\right)^2$. We obtain an estimate of the variance of our approximation that explains how the errors are decomposed across scale. The accuracy of this approach relies on more detailed knowledge of the structure of the covariance than the bound we have presented earlier. That said, since the statistics of wavelet coefficients of various random processes have been analyzed in prior work [18]–[21], there are certainly classes of processes in which this alternate variance approximation can be quite accurate. Moreover, taking advantage of such additional knowledge of covariance structure may suggest alternative bases to W, and in turn to B, that are adapted to the process structure and yield tighter bounds.[14]

C. Multiscale Models for Processes With Long-Range Correlations

In our analysis, the errors in variance approximations mainly depend on the covariance structure of P, and the information matrix J does not play a direct role. However, J plays a crucial role during estimation—the model has to be Markov with respect to a sparse graph to be able to store it efficiently, and to solve the linear system $J\mu = h$ efficiently. Some processes with slow correlation falloff do not have a sparse information matrix in a one-scale representation, so they do not fit well into our approach. However, slow correlation falloff can be modeled using sparse multiscale representations with hidden variables, as we discussed in Section II. A pyramidal model with a stochastic relationship between scales was proposed in [13] and [15]. We consider the problem of finding approximate variances in such a model.

A representative structure for the model is illustrated in Fig. 1(c). The variables in the bottom (fine) scale correspond to some physical phenomenon that is being modeled. The variables at coarser scales represent aggregates over local regions. They may or may not be of interest in the estimation, but they serve to induce a sparse graph structure (once they are integrated out, the fine-scale model in general has a complete, nonsparse, information matrix). Aggregation can mean that the coarser scale variable represents an average, or some weighted combination of the variables in the finer scale over a small region. However, the relationship across scale is nondeterministic, allowing for uncertainty. The graph is sparse but has many loops.

The structure of the information matrix is such that variables in one scale are only connected to nearby scales. Hence the J matrix for a multiscale model with four scales has the following chain structure (with scale 1 being the finest and 4 the coarsest):

$$J = \begin{pmatrix} J_1 & J_{12} & & \\ J_{21} & J_2 & J_{23} & \\ & J_{32} & J_3 & J_{34} \\ & & J_{43} & J_4 \end{pmatrix}. \tag{27}$$

Suppose that we are mainly interested in computing the variances of the variables at the finest scale (the other ones are auxiliary), i.e., in the block of J^{-1} corresponding to scale 1. Hence in our approach, we only need to approximate $\mathrm{blockdiag}(I, 0, 0, 0)$ and not the full I matrix. We use the matrix $B = \begin{pmatrix} B_1 \\ 0 \end{pmatrix}$, with zero for all coarser scales.[15] Here B_1 is a spliced wavelet basis corresponding to variables at scale 1 (the same construction as in Section IV).

Our error analysis takes into account only the covariance structure of the fine scale variables $P_1 = [J^{-1}]_1$. Hence, it is oblivious to the hidden variables representation and only depends on the properties of the marginal covariance block P_1. Experimental results with this multiscale model for processes with long-range correlations are presented in Section V.

V. Computational Experiments

Our first experiment involves a 1-D thin-membrane model with length $N = 256$, with nearest neighbor connections. Noisy observations are added at a few randomly selected nodes. This model has a short correlation length; see Fig. 8(top). We apply the single-scale low-rank method from Section III-A and plot the errors in variances (absolute error in percent, averaged over all nodes) versus the separation length in Fig. 8(bottom). The errors decay fast with separation length, in line with our analysis in Section III-B.

Next we consider a 1-D thin-membrane model with connections from each node to nodes up to four steps away. The J matrix is close to singular, and the correlation length in the model is long, see Fig. 9(top). We illustrate the results using both the

[14] For example, one could use partial wavelet decompositions that stop at intermediate scales, and more generally wavelet packets [17] adapted to the statistics of wavelet coefficients at different scales.

[15] Alternatively, if the variances at coarser scales are of interest, we use the matrix $\mathrm{blockdiag}(B_1, B_2, B_3, B_4)$, where B_i is a spliced wavelet basis corresponding to scale i. The errors are decoupled: errors from B_i at scale i are not propagated to other scales.

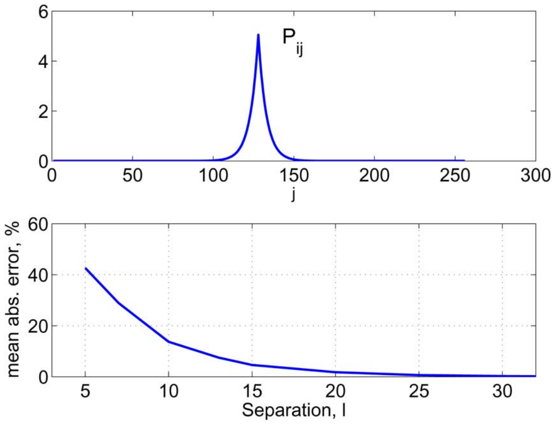


Fig. 8. (Top) Correlation P_{ij} from the center node. (Bottom) Errors in variances (mean absolute error, in percent) versus separation length l.

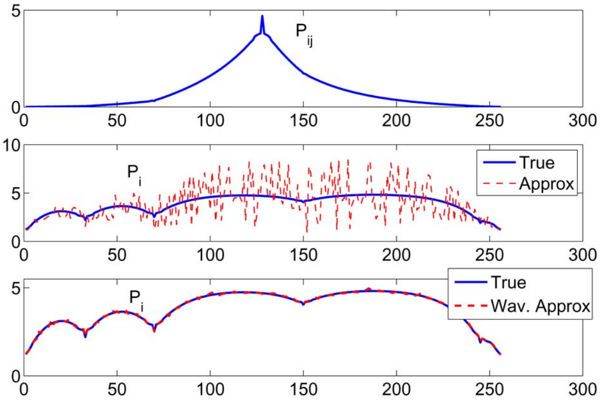


Fig. 9. One-dimensional example with long-correlation. (Top) Correlation P_{ij} from the center node. (Center) True variance, and low-rank approximate variance using one scale. (Bottom) True variance, and low-rank wavelet-based approximate variance.

single-scale (middle) and the wavelet-based (bottom) low-rank methods. We use $M = 32$ for the single-scale approach, which is too small compared to the correlation length. While the approximation is unbiased, its high variance makes it practically useless. For the wavelet-based case, using a smaller matrix B with $M = 28$, constructed by splicing a Coifman wavelet basis (Coiflet basis) [22], we are able to find very accurate variance approximations as seen in Fig. 9(bottom).

Next we apply the approach to a 2-D thin-membrane model of size 256 × 256, with correlation length about 100 pixels, and with sparse noisy measurements taken at randomly selected locations. The underlying true field is flat. We use separable Coifman wavelets, and the resulting sparse B matrix has size 65 536 × 304. This is a very significant reduction in the number of columns, compared to W. The results appear in Fig. 10: the errors (bottom left) are small compared to the variances (top). Our approximate solution is a close match to the exact solution, which can still be computed for models of this size. The 2-D autocorrelation of the errors appears in Fig. 10 (bottom right): the errors are weakly correlated, supporting our alternative error analysis based on P_W in Section IV-B. Next, we apply our low-rank variance approximation method to ocean surface height data collected along the tracks of Jason-1 satellite[16] over the Pacific Ocean region. The data are sparse and highly irregular. We use the thin-plate model for the data. The measurements in general fall between the grid points, and they are modeled as bilinear interpolation $y_k = h_k x + n_k$ of the nearest four nodes in the grid (h_k has four nonzero entries) with added white Gaussian noise $n_k \sim \mathcal{N}(0, \gamma)$. The posterior information matrix combines the thin-plate prior J_{tp} with the measurements $J = J_{tp} + (1/\gamma) H^T H$. It is sparse because the measurements only induce local connections within each cell in the grid.

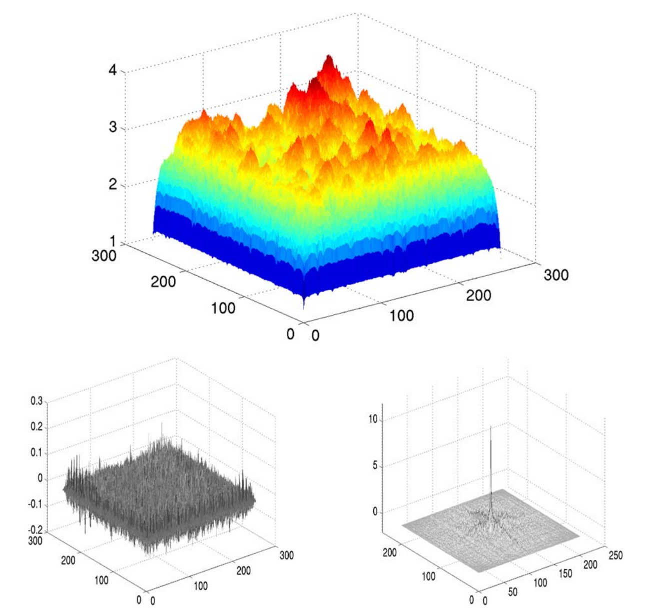


Fig. 10. Two-dimensional thin-membrane example. (Top) approximate variances, (bottom left) errors, and (bottom right) 2-D autocorrelation of errors. The approximations are accurate (errors are much smaller than the variances), and the errors are weakly correlated.

The size of the field is 1024 × 1024, i.e., over a million variables. Computing the variance in a model of this size is beyond what is practical with exact methods on a single workstation. We use our approximate variance calculation method. The correlation length is moderate, so using just two wavelet scales suffices, and the B matrix has only 448 columns. The resulting approximate variances using a version of the embedded trees (ET) iterative solver (described in Appendix B) appear in Fig. 11. The regions over land are ignored (in black). The variances are lowest near the measurements (along the tracks), as expected.

Next, we consider a gravity inversion problem, where one is interested in estimating the underground geological structure of a 3-D volume based on gravity measurements on its surface. For simplicity, we consider a 2-D version of this problem. We divide the 2-D subsurface region into small blocks and model the mass x in the blocks as a thin-plate GMRF. The gravity measurements y on the surface come from a discretization of Newton's law

[16] This altimetry dataset is available from the Jet Propulsion Laboratory: http://www.jpl.nasa.gov. It is over a ten-day period beginning December 1, 2004. The data are normalized to remove seasonal spatially varying average sea levels.

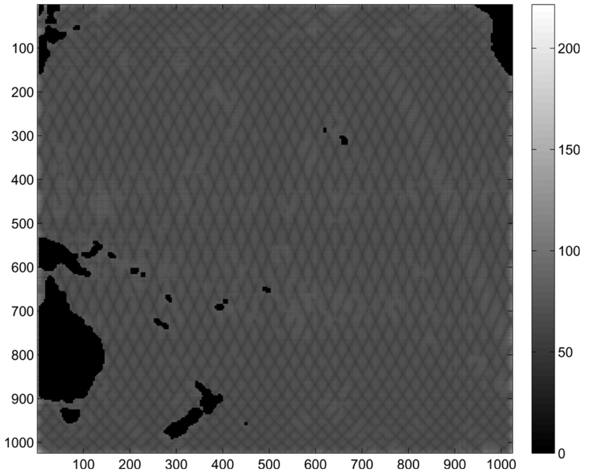
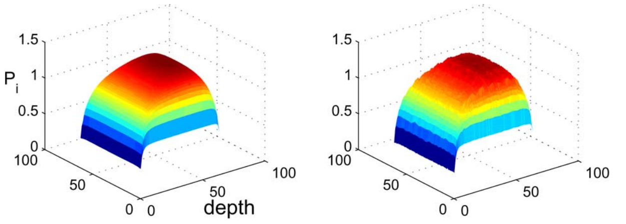


Fig. 11. Approximate uncertainty (millimeters) of Pacific Ocean surface height based on measurements along satellite tracks, 1024 × 1024 grid.



(a)

(b)

Fig. 12. Gravity inversion example: (a) exact variances and (b) accurate approximate variances using the wavelet-based low-rank approach. The variances increase with depth.

of universal gravitation. They are linear in the unknowns x but nonlocal—they couple all the nodes in the GMRF

$$y_i = G\sum_j \frac{u_{ij}x_j}{d_{ij}^2} + n_i. \tag{28}$$

Here y_i is the two-component (horizontal and vertical) gravity measurement at point i on the surface and x_j is the unknown mass at the jth subsurface node; see Fig. 12(a). Also, G is the gravitational constant, d_{ij} and u_{ij} are, respectively, the distance and the unit vector from the location of the jth node to ith measurement point, and n_i is Gaussian noise with diagonal covariance Q. Combining the linear measurement model $y = Hx + n$ with the thin-plate prior J_{tp} for the unknown field x, the posterior variance that we would like to approximate is

$$P = \left(J_{tp} + H^T Q^{-1} H\right)^{-1}. \tag{29}$$

Note that this problem does not simply correspond to a sparse matrix J: in addition to the sparse component J_{tp}, there is also a low-rank nonsparse component $H^T Q^{-1} H$ due to the nonlocal measurements. However, using a version of block Gauss–Seidel (see Appendix B), we still obtain fast solution of the resulting linear system. We consider a square region with 64 × 64 nodes, with gravity measurements at 64 locations at the top.[17] We plot the true variances and those obtained using a wavelet-based low-rank approach with four scales and 206 columns of B (instead of 4096). Despite the long-range correlation induced by the observation model, and the addition of the nonsparse $H^T Q^{-1} H$ term, the method still gives accurate variances, as we show in Fig. 12.

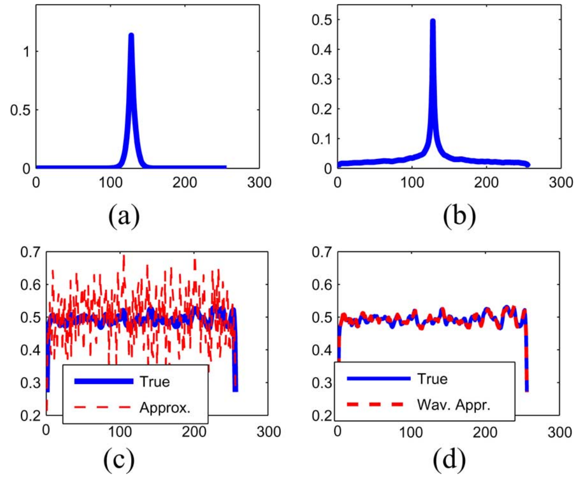


Fig. 13. Multiscale example: (a) conditional and (b) marginal correlation at the fine scale. (c) Approximate variances using the low-rank approach: spliced standard basis. (d) Accurate approximate variances using the wavelet-based low-rank approach.

Finally, we apply our reduced-rank approach to a multiscale model on a pyramidal graph, as described in Section II-A. The model has 256 variables in the finest scale and five coarser levels, with the number of variables decreasing twofold for each coarser level. The total number of variables is 496. In Fig. 13(a), we show the fast-decaying conditional correlation at the fine scale (conditioned on the coarser scales) and in (b) the slow-decaying marginal correlation at the fine scale. The fast decay of conditional correlations allows efficient solutions of the linear systems in our approach. However, the errors in our low-rank variance approximations depend on the long-range marginal correlations, requiring the use of the wavelet-based approach. In Fig. 13, we show the results of computing approximate variance using the single-scale approach in (c) and the wavelet-based approach in (d). The sizes of the resulting aliasing matrices B are 496 × 32 and 496 × 28, respectively. It can be seen that the single-scale approach is inadequate, while the wavelet-based B yields very accurate variances, even though it uses an aliasing matrix B with fewer columns. This is as expected—the model has a long marginal correlation length at the fine scale, which only the wavelet-based approach is able to handle.

VI. Conclusion

We have presented a simple computationally efficient scheme to compute accurate variance approximations in large-scale GMRF models. The scheme involves designing a low-rank aliasing matrix that is used during matrix inversion. By a

[17] We assume that the density is approximately known outside the square region (this is not required, but it simplifies the problem).

judicious choice of the aliasing matrix, the errors in the approximation can be made unbiased and with small variances. We have designed aliasing matrices for both the short-range and smooth long-range correlation cases and applied them to single and multiscale GMRF models.

There are many interesting directions for further research: using wavelet packets to better adapt to the statistics of the GMRF; using diffusion wavelets [23] to extend the wavelet-based construction of B to arbitrary (nonregular) graphs; and interpreting our approach in the walk-sum framework for Gaussian inference [8]. In addition, for multiscale GMRF models, we are interested to find ways to design a low-rank aliasing matrix that exploits the short correlation length of the conditional model within each scale, rather than using wavelet-based constructions.

Appendix A
Stability of Errors in Wavelet Based Approximation

We provide the analysis for Proposition 1. Recall the expression for $\mathbb{E}_\sigma\left[E_{ii}^2\right]$ that we obtained in (24) and decompose it according to scale s

$$\mathbb{E}_\sigma\left[E_{ii}^2\right] = \sum_s \sum_{l \in s} \sum_{k \in C(l) \backslash l} R_{ik}^2 W_{il}^2. \tag{30}$$

Since W_l has compact support, W_{il} is nonzero only for some constant (independent of N and s) number of wavelets at each scale that contain i in the support. Let K be an upper bound on this constant. Also, W_{il}^2 at scale s is bounded by 2^{-s} since $\|W_l\|_2 = 1$, and $\psi_{s,k}(t) = (1/2^{s/2})\psi(2^{-s}t - k)$. Thus we have

$$\mathbb{E}_\sigma\left[E_{ii}^2\right] \le K \sum_s \sum_{k \in C(l_s^*) \backslash l_s^*} R_{ik}^2 2^{-s}. \tag{31}$$

Here l_s^* is the index that achieves the maximum sum over k at scale s. We bound the other terms in the sum over l by this maximum, giving a factor of K in front.

First, suppose that we are dealing with a one-dimensional GMRF and that outside the region of disturbance, P_{ij} decays exponentially with $d(i,j)$, i.e., $P_{ij} = A\beta^{d(i,j)}$, $|\beta| < 1$. Then the response $R_k(i)$ also decays exponentially with the same decay rate outside the region of disturbance $R_k(i) = A_s \beta^{d(i,j(k))}$, where $j(k)$ corresponds to the peak of W_k. This happens because exponentials are eigenfunctions of linear time-invariant filters. However, the constant A_s decreases rapidly with each finer scale. If our wavelet has m vanishing moments, then $A_s = O(2^{((s-N_{sc})(m+1/2)})$ for k that belongs to scale s, $s \in \{1, \ldots, N_{sc}\}$. Recall that N_{sc} is the number of scales we use in the wavelet basis, which depends on the correlation length L of the process: we set $N_{sc} \propto \log_2(L)$.

We can write $\sum_k R_{ik}^2 = A_s^2 Q_\beta(s)$, where we define $Q_\beta(s) = \sum_{k \in C(l_s) \backslash l_s} \beta^{2d(i,j(k))} = \sum_{n \neq 0} \beta^{2d_s|n|}$, with n indexing the aliased terms, and we use d_s to denote the separation length at scale s. Consider how $Q_\beta(s)$ depends on s. The separation length in our construction is $d_s = d_1 2^{s-1}$, where d_1 is the separation at the finest scale. The number of aliased terms doubles with each finer scale, and the distance between them decreases by a factor of two. For one-dimensional signals, this (unscaled) error roughly doubles with each finer scale: $Q_\beta(s) = \sum_{n \neq 0} \beta^{(|n| d_1 2^s)}$ satisfies $Q_\beta(s+1) \le (1/2) Q_\beta(s)$.

Hence $Q_\beta(s) \le 2^{-(s-1)} Q_\beta(1)$. Note that the term $Q_\beta(1)$ is equal to the error in the original (wavelet-less) construction with separation distance d_1. Putting all the pieces together, the total error in (31) is bounded by

$$\begin{aligned} K \sum_{s=1}^{N_{sc}} 2^{-s} \sum_k R_{ik}^2 &\le K \sum_s 2^{-s} A_s^2 Q_\beta(1) 2^{(1-s)} \\ &\le 2K Q_\beta(1) \sum_s 2^{-2s} 2^{(s-N_{sc})(2m+1)} \\ &\le 2K Q_\beta(1) 2^{-2N_{sc}} \sum_{s=1}^{N_{sc}} 2^{(s-N_{sc})(2m-1)} \\ &\le 4K Q_\beta(1) 2^{-2N_{sc}}, \text{ if } m \ge 1. \end{aligned} \tag{32}$$

In the last line, if the number of vanishing moments satisfies $m \ge 1$, then the sum $\sum_{s=1}^{N_{sc}} 2^{(s-N_{sc})(2m-1)} \le 2$ for any N_{sc}. That means that the total error is bounded by a constant multiple of $2^{-2N_{sc}} Q_\beta(1)$. Since $N_{sc} \propto \log_2(L)$, this is roughly $L^{-2} Q_\beta(1)$. As we mentioned, $Q_\beta(1)$ roughly corresponds to the error in the standard basis construction (without wavelets). From Section III, we know that $Q_\beta(1)$ is bounded so the errors in wavelet-based construction are also bounded, and it can be seen that using wavelets, we get a much smaller error. We also know that by controlling d_1, the error $Q_\beta(1)$ can be made arbitrarily small. Hence, the same is true for the error in the wavelet-based construction.

Now let us consider power-law decay of correlations (again outside of the disturbance region) $P_{ij} = A d(i,j)^{-p}$, with $p > 0$. In contrast to the exponential decay, the power-law decay changes when wavelets are applied. A wavelet with m vanishing moments acts as local smoothing followed by mth order differentiation [17], so if P_{ij} decays as $d(i,j)^{-p}$, then $R_k(i)$ decays as $d(i, j(k))^{-(p+m)}$. This means that the tails of $R_k(i)$ decay faster than the tails of P_{ij}. We define $Q_p(s) = \sum_n (2d_s|n|)^{-(p+m)}$. The bound for $Q_p(s)$ in terms of $Q_p(1)$ changes $\sum_n ((d_s/2)|n|)^{-(p+m)} = 2^{(p+m)} \sum_n (d_s|n|)^{-(p+m)}$; hence $Q_p(s) = 2^{-(p+m)(s-1)} Q_p(1)$. Putting everything together, the error in (31) is bounded by

$$\begin{aligned} &K \sum_{s=1}^{N_{sc}} 2^{-s} R_{ik}^2 \\ &\quad \le K \sum_s 2^{-s} A_s^2 Q_p(1) 2^{(p+m)(1-s)} \\ &\quad \le K Q_p(1) 2^p \sum_s 2^{-(p+m+1)s} 2^{(s-N_{sc})(2m+1)} \\ &\quad \le K Q_p(1) 2^p 2^{-(p+m+1)N_{sc}} \sum_{s=1}^{N_{sc}} 2^{(s-N_{sc})(m-p)} \\ &\quad \le K Q_p(1) 2^{p+1} 2^{-(p+m+1)N_{sc}}, \text{ if } m > p+1. \end{aligned} \tag{33}$$

In the last line, if $m > p+1$, then the sum $\sum_{s=1}^{N_{sc}} 2^{(s-N_{sc})(m-p)} \le 2$ and the total error is bounded by a constant multiple of $2^{-(p+m+1)N_{sc}} Q_p(1)$, or roughly $L^{-(p+m+1)} Q_p(1)$. If $1 \le m < p$, then the sum is dominated by the largest term $\sum_{s=1}^{N_{sc}} 2^{(s-N_{sc})(m-p)} \approx 2^{(p-m)N_{sc}}$, and the total error is a constant multiple of $2^{-(p+m+1)N_{sc}} 2^{(p-m)N_{sc}} Q_p(1) = 2^{(-2m-1)N_{sc}} Q_p(1)$, or roughly $L^{-2m} Q_p(1)$. In either case, the total error is bounded by a small multiple of $Q_p(1)$. For the power-law decay, $Q_p(1)$ is in fact smaller than the error using the standard basis, as using wavelets we change the power of decay from p to $p+m$. Hence $Q_p(1)$ roughly corresponds to the error in the single-scale construction with p replaced by $p+m$. Using our results for the standard basis in Section III, we can conclude that the total errors are bounded and can be made arbitrarily small by controlling d_1.

Also note that wavelet-based construction is especially advantageous in lattices of higher dimension (with d dimensions): there, the convergence of $Q_p(1)$ requires $p > (d/2)$.[7] fHowever, with the wavelet construction, we only need $p + m > (d/2)$. This means that for the case where the errors in the standard-basis construction diverge, we can still make them converge using wavelets with sufficient number of vanishing moments.

Appendix B
Efficient Solution of Linear Systems

In our approach, we compute the variances by solving a small number $M \ll N$ of linear systems $JR_i = B_i$, all sharing the same matrix J. Whenever a fast solver for J is available, the overall variance approximation scheme is also fast.

Iterative approaches such as Richardson iterations and conjugate gradient methods are very appropriate for our approach, as multiplication by a sparse J is very efficient, so the cost per iteration is low. The number of iterations can be controlled by using a good preconditioner for J, one that is easy to evaluate and serves as an approximation of J^{-1}.

An efficient set of preconditioners based on embedded trees has been developed in [24] for the lattice GMRF model. The idea is that for models with a tree-structured graph $\mathcal{G}$, solving the system $J\mu = h$ (i.e., applying J^{-1} to a vector) is highly efficient—it can be done in $O(N)$ operations. Hence, for general graphs $\mathcal{G}$, [24] uses spanning trees $T \subset \mathcal{G}$ with preconditioner J_T^{-1}. We use a similar strategy based on block Gauss–Seidel iterations that uses thin induced subgraphs as blocks. We partition the lattice into narrow overlapping horizontal and vertical strips. Estimation in the strip (conditioned on other variables being fixed) can be done efficiently with the cost linear in the length and cubic in the width of the strip. By iterating over the strips, convergence to the correct means is guaranteed.[18] We use this approach in the experiments in Section V. The same approach applies to gravity inversion with sparse plus low-rank structure when small blocks are used instead of strips.

There are several directions for designing potentially even more efficient preconditioners. Recently, [25] proposed an adaptive scheme based on ET that picks the spanning trees adaptively to have the most impact in reducing the error. This should be beneficial within the context of block Gauss–Seidel as well. Also, for single-scale models with long-range correlations, using multiscale solvers such as [5] can dramatically improve convergence. Alternatively, when the MRF model itself has multiple scales (as in Section IV-C), then estimation approaches in [13] and [15] can be used. There the model is decomposed into a tractable tree-structured component and disjoint horizontal components (one for each scale), which, conditioned on the coarser scale variables, have short conditional correlations and are also tractable. By iterating between these tractable subproblems, estimation in the whole multiscale model can be done efficiently.

[18] We note that, in general, the convergence of ET iterations is not guaranteed. By also requiring the subtrees to be induced, we force ET to be equivalent to Gauss–Seidel, guaranteeing its convergence.

Acknowledgment

The authors would like to thank V. Chandrasekaran for helpful discussions.

References

[1] M. I. Jordan, “Graphical models,” *Statist. Sci. (Special Issue on Bayesian Statistics)*, vol. 19, pp. 140–155, 2004.

[2] S. L. Lauritzen*, Graphical Models*. Oxford, U.K.: Clarendon, 1996.

[3] H. Rue and L. Held*, Gaussian Markov Random Fields Theory and Applications*. London, U.K.: Chapman and Hall/CRC Press, 2005.

[4] J. K. Johnson and A. S. Willsky, “A recursive model-reduction method for approximate inference in Gaussian Markov random fields,” *IEEE Trans. Image Process.*, vol. 17, pp. 70–83, Jan. 2008.

[5] U. Trottenberg, C. W. Oosterlee, and A. Schuller*, Multigrid*. New York: Academic, 2001.

[6] D. Spielman and S. H. Teng, “Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems,” in *Proc. ACM Symp. Theory Comput.*, 2004.

[7] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhater*, Probabilistic Networks and Expert Systems*. Berlin, Germany: Springer, 2003.

[8] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, “Walk-sums and belief propagation in Gaussian graphical models,” *J. Mach. Learn. Res.*, vol. 7, Oct. 2006.

[9] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, “Low-rank variance estimation in large-scale GMRF models,” in *Proc. IEEE ICASSP*, May 2006.

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

[11] B. Gidas, “A renormalization group approach to image processing problems,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 164–180, Feb. 1989.

[12] A. S. Willsky, “Multiresolution Markov models for signal and image processing,” *Proc. IEEE*, vol. 90, pp. 1396–1458, 2002.

[13] M. J. Choi, “Multiscale Gaussian graphical models and algorithms for large-scale inference,” M.S. thesis, Massachusetts Inst. of Technology, Cambridge, 2007.

[14] C. A. Bouman and M. Shapiro, “A multiscale random field model for Bayesian image segmentation,” *IEEE Trans. Image Process.*, vol. 3, pp. 162–177, Mar. 1994.

[15] M. J. Choi, V. Chandrasekaran, D. Malioutov, J. K. Johnson, and A. S. Willsky, “Multiscale stochastic modeling for tractable inference and data assimilation,” *Comput. Meth. Appl. Mechanics Eng.*, vol. 197, pp. 3492–3515, 2008.

[16] B. Aspvall and J. R. Gilbert, “Graph coloring using eigenvalue decomposition,” *SIAM J. Alg. Disc. Meth.*, vol. 5, pp. 526–538, 1984.

[17] S. Mallat*, A Wavelet Tour of Signal Processing*. New York: Academic, 1998.

[18] *Theory and Applications of Long-Range Dependence*, P. Doukhan, G. Oppenheim, and M. S. Taqqu, Eds. Berlin, Germany: Birkhauser, 2003.

[19] P. Flandrin, “Wavelet analysis and synthesis of fractional Brownian motion,” *IEEE Trans. Inf. Theory*, vol. 38, pp. 910–917, Mar. 1992.

[20] E. Masry, “The wavelet transform of stochastic processes with stationary increments and its application to fractional Brownian motion,” *IEEE Trans. Inf. Theory*, vol. 39, pp. 260–264, Jan. 1993.

[21] R. W. Dijkerman and R. R. Mazumdar, “Wavelet representation of stochastic processes and multiresolution stochastic models,” *IEEE Trans. Signal Process.*, vol. 42, pp. 1640–1652, Jul. 1994.

[22] I. Daubechies*, Ten Lectures on Wavelets*. Singapore: SIAM, 1992.

[23] R. R. Coifman and M. Maggioni, "Diffusion wavelets," *Appl. Comp. Harm. Anal.*, vol. 21, no. 1, pp. 53–94, 2006.

[24] E. Sudderth, M. J. Wainwright, and A. S. Willsky, "Embedded trees: Estimation of Gaussian processes on graphs with cycles," *IEEE Trans. Signal Process.*, vol. 52, pp. 3136–3150, Nov. 2004.

[25] 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, pp. 1916–1930, May 2008.

Dmitry M. Malioutov (S'01) received the B.S. degree in electrical and computer engineering from Northeastern University, Boston, MA, in 2001 and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 2003 and 2008, respectively.

Currently, he is a Postdoctoral Researcher at Microsoft Research, Cambridge, U.K. His research interests include statistical signal and image processing, machine learning, graphical models and message passing algorithms, and sparse signal representation.

Jason K. Johnson received the S.B. degree in physics and the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1995, 2003, and 2008, respectively.

He was a Member of Technical Staff (1995–2000) with Alphatech, Inc., Burlington, MA, where he developed algorithms for multiscale signal and image processing, data fusion, and multitarget tracking. Currently, he is a Postdoctoral Researcher at Los Alamos National Laboratory. His current research interests include Markov models, tractable methods for inference and learning, and statistical signal and image processing.

Myung Jin Choi (S'06) received the B.S. degree in electrical engineering and computer science from Seoul National University, 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 pursuing the Ph.D. degree in the Stochastic Systems Group.

Her research interests include statistical signal processing, graphical models, and multiresolution algorithms.

Ms. Choi is a Samsung scholarship recipient.

Alan S. Willsky (S'70–M'73–SM'82–F'86) joined the Massachusetts Institute of Technology, Cambridge, in 1973, where he is the Edwin Sibley Webster Professor of Electrical Engineering and Acting 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 *Signals and Systems* (Englewood Cliffs, NJ: Prentice-Hall, 1983). His research interests are in the development and application of advanced methods of estimation, machine learning, and statistical signal and image processing.

Dr. Willsky has received numerous 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 1988 IEEE Control Systems Society Distinguished Member Award, and the 2004 IEEE Donald G. Fink Prize Paper Award. He received a Doctorat Honoris Causa degree from the Université de Rennes, France, in 2005.