GMRF VARIANCE APPROXIMATION USING SPLICED WAVELET BASES

Dmitry M. Malioutov, Jason K. Johnson, and Alan S. Willsky

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Ave., Cambridge, MA 02139, USA

ABSTRACT

We consider the problem of computing variances in large-scale Gauss-Markov random field (GMRF) models. In our prior work we considered the short-range correlation case, and we proposed a simple low-rank method which computes approximate variances with linear complexity in the number of nodes. In addition to its low complexity, the method has good guarantees on the quality of the approximation. In this paper we extend our method and analysis using a wavelet-based multi-scale approach which is applicable to models with much longer correlation lengths.

Index Terms— GMRF, approximate variances, wavelets.

1. INTRODUCTION

Gauss-Markov random field models (GMRF) [1] are graphical models on undirected graphs where the variables are jointly Gaussian. The nodes of the graph denote random variables, and the edges represent statistical relations between the variables. We address the problem of estimation in large-scale GMRFs, which arise in a wide variety of applications including computer vision, geostatistics and oceanography. A prototypical application is interpolation from sparse irregular noisy measurements [2].

Since GMRF estimation is a special case of a linear Gaussian problem, in principle the conditional means and variances can be obtained by matrix inversion. However, for large-scale problems with millions of variables, exact algorithms such as Gaussian elimination (with $O(N^3)$ complexity in the number of variables N) are intractable. Approximate mean estimates can be computed with $O(N)$ complexity for sparse graphs by iterative solvers such as conjugate gradients or multigrid. Mean estimates are much less insightful when the confidence in them (the variance) is unknown. However, no general techniques exist to efficiently compute the variances: exact methods are too computationally demanding, while approximate methods such as loopy belief propagation give no guarantees on accuracy. In [3] we have presented a simple approach to compute efficient variance approximations in models with short-range correlations using a low-rank matrix instead of identity when computing the inverse. By designing the low-rank matrix such that only the weakly correlated terms are aliased, we were able to give provably accurate variance approximations.

In this paper we propose a multi-scale construction which extends our approach to handle models with long correlation length. The basis of our construction is the discrete wavelet transform that we use to decompose the correlation across several scales, thus reducing the problem with long-correlation length to a few problems with shorter correlation length. We extend some of the favorable properties of our earlier approach, including unbiasedness and accuracy of variances.

In Section 2 we discuss estimation with GMRF models and review our single-scale low-rank variance approximation approach. We describe the multi-scale extension in Section 3 first for 1D models, and then for 2D. We exhibit the merits of the approach on examples in Section 4.

2. ESTIMATION IN GMRF MODELS

A GMRF model is defined by a graph $\mathcal{G} = (V, \mathcal{E})$ with vertices V and edges $\mathcal{E} \subset \binom{V}{2}$, i.e., some set of two-element subsets of V, and a collection of jointly Gaussian random variables $x = (x_i, i \in V)$ with probability density given in *information form*:

$$p(x) \propto \exp\{-\frac{1}{2}x'Jx + h'x\}. \quad (1)$$

The matrix J is called the information matrix, and it 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 $\mu = \mathbb{E}\{x\}$ and the covariance matrix $P = \mathbb{E}\{(x-\mu)(x-\mu)'\}$:

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

For a concrete example, consider the linear Gaussian problem, with observations $y = Ax + n$, where x is zero-mean with covariance P, and independent noise n is zero-mean and

This research was supported by the Air Force Office of Scientific Research under Grant FA9550-04-1, the Army Research Office under Grant W911NF-05-1-0207, and by a grant from MIT Lincoln Laboratory. We thank V. Chandrasekaran for helpful discussions.

with covariance R. Then the Bayes least-squares estimate $(\hat{x}, \hat{P})$ is given by:

$$(P^{-1} + A'R^{-1}A)\,\hat{x} = A'R^{-1}y, \qquad (3)$$
$$\hat{P} = (P^{-1} + A'R^{-1}A)^{-1}.$$

If $J_x = P^{-1}$ is a sparse GMRF prior on x, and y are local observations, then $J = (P^{-1} + A'R^{-1}A)$ has the same sparsity as J_x, with only the diagonal terms being modified. Now J and $h = A'R^{-1}y$ are the information parameters specifying the conditional model given the observations. As we discussed, for large-scale GMRFs exact inversion is intractable, so approximate approaches have to be considered.

2.1. Low-rank variance calculation

Our low-rank variance approximation approach [3] is based on the fact that for sparse J, while computing $P = J^{-1}$ may be hard, calculating the i-th column of P can be done efficiently to a given tolerance by a sparse iterative solver. This is done by solving a linear system $JP_i = e_i$, where e_i is the i-th standard basis vector. To get all N columns of P, this would have to be done N times, at each node in the graph: $JP = [e_1, ..., e_N] = I$ with complexity $O(N^2)$. This is still intractable for large-scale models (we do not need the full P with N^2 elements; computing the N variances suffices). In [3] we proposed to use a low-rank matrix BB', with $B \in \mathbb{R}^{N\times M}$ and $M \ll N$, instead of the identity. The system $J\hat{P} = BB'$ can be solved with $O(MN)$ complexity in two steps: first we solve $JR = B$ using iterative solvers. Then, we post-multiply R by B', i.e. $\hat{P}_{ii} = [RB']_{ii}$ (which requires MN operations, as we only need the diagonal).

To get accurate variance approximations, B must be designed appropriately, taking the graph into consideration. Let all rows b_i of B have unit norm: $b_i'b_i = 1$. Consider the quantity $\mathrm{diag}(J^{-1}(BB'))$:

$$\hat{P}_{ii} \triangleq [J^{-1}(BB')]_{ii} = P_{ii} + \sum_{i\neq j} P_{ij}\, b_i'b_j. \qquad (4)$$

We need the error terms $P_{ij}\, b_i'b_j$ to be nearly zero for all pairs of nodes. In [3] we assume that the correlation P_{ij} rapidly decays with distance from i to j. Hence, error terms are automatically small for pairs i and j that are far away compared to the correlation length. It remains to design B so that b_i and b_j are orthogonal for nearby nodes i and j.

Constructing B for 1D models: single-scale version. To be concise, we outline a construction of B on 1D models. Refer to [3] for a construction on 2D lattices and arbitrary graphs.

Consider a 1D model of length N (allowing nearest neighbor, as well as long-range interactions in J). We group nodes into classes, which we call colors, such that nodes of the same color are a distance M apart. We will have a column B_c of B for each color c. We assign $B_c(i) = \pm 1$ randomly for each node i of color c, and $B_c(j) = 0$ for other nodes. An illustration appears in Figure 1. We plot a column of B on the left, and the result of $J^{-1}B_i$ on the right. To find $\hat{P}$, we have to repeat this for all the colors, add them, and apply B'.

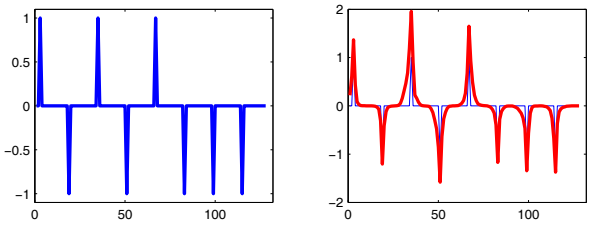

Fig. 1. A column B_3 of B and the corresponding PB_3.

Properties of $\hat{P}$. In [3], we have analyzed the approximation $\hat{P}$, and showed that it has very favorable properties. Due to the random nature of B, $\hat{P}$ is an unbiased approximation of P, and for models with short correlation length, by choosing the local region to be large enough, we establish bounds on its variance. Next, we extend the single-scale construction to multi-scale via a discrete wavelet transform. Some of the favorable properties are inherited by the extension.

3. MULTI-SCALE LOW-RANK VARIANCE APPROXIMATION VIA WAVELETS

When the correlation length is comparable to the size of the signal, the single-scale approach gives no computational savings. To address this problem, we propose to decompose the signal into several frequency bands. A very convenient tool for such a decomposition is the discrete wavelet transform.

A discrete wavelet decomposition is specified by a scaling function $\phi(t)$ and a wavelet function $\psi(t)$ which are the solutions to certain related dilation equations [4]. A scaling function generates a family of dilations and translations, $\phi_{s,k}(t) = \frac{1}{2^{s/2}}\phi(2^{-s}t - k)$. For a fixed scale s, the set $\{\phi_{s,k}(t)\}_k$ generates the approximation space $\mathcal{V}_s$. The spaces $\mathcal{V}_s$ are nested: $\mathcal{V}_1 \supset \mathcal{V}_2 \supset \mathcal{V}_3$ ..., with higher s corresponding to coarser scales.

The wavelet function $\psi(t)$ also generates a family of dilations and translations: $\psi_{s,k}(t) = \frac{1}{2^{s/2}}\psi(2^{-s}t - k)$. The span of $\{\psi_{s,k}(t)\}_k$ at a given scale s gives the detail space $\mathcal{W}_s = \mathcal{V}_{s-1} \ominus \mathcal{V}_s$. We can decompose the fine scale $\mathcal{V}_1$ into $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus ... \oplus \mathcal{W}_{N_s} \oplus \mathcal{V}_{N_s}$, where N_s is the number of scales. We focus on orthogonal wavelet families 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.

A discrete wavelet basis for the space $\mathcal{V}_1$ is constructed by collecting the scaling functions at the desired coarsest scale, and the wavelet functions at all finer scales as columns of a matrix W. Let S_s and W_s consist of the translations of the scaling and wavelet functions, respectively, at scale s. Then

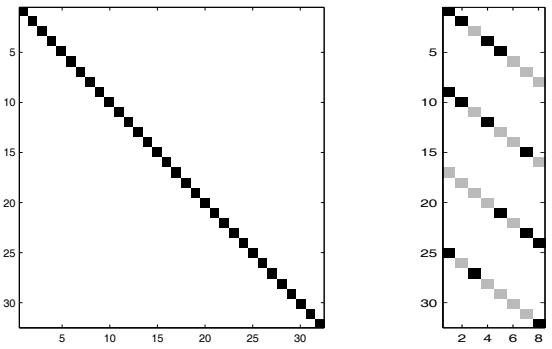

Fig. 2. (left) identity and (right) locally orthogonal B matrix.

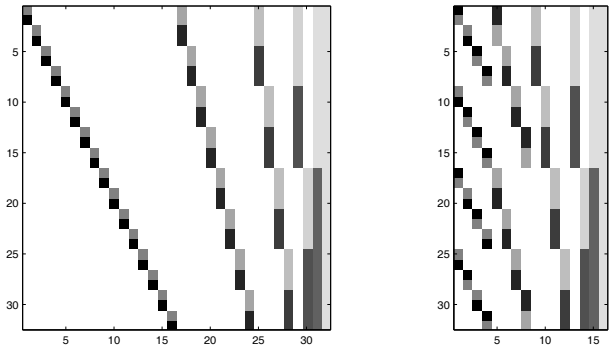

Fig. 3. (left) A discrete wavelet basis (right) B matrix obtained by splicing certain columns of W at each scale.

our orthogonal basis consists of

$$W = [W_1\ W_2\ ....\ W_{N_s}\ S_{N_s}]. \quad (5)$$

We consider wavelets on intervals, and use periodic boundary conditions. At level s we have 2^{N_s-s} possible translations, hence that many columns in W_s. An illustration of a Haar wavelet basis for $N = 32$ is given in Figure 3 (left).

Constructing multi-scale spliced B for 1D. To motivate the multi-scale construction in 1D, we refer back to the single-scale one. Using I instead of B in (4) results in a perfect recovery of P, as there is no interference between the different rows of I: $e_i' e_j = 0$ for $i \neq j$. When the correlation length is much smaller than N, there is negligible interference between nodes far enough away. By splicing the *columns* of B as illustrated in Figure 2, we get a matrix B which is locally orthogonal. By splicing we mean adding groups of columns (corresponding to nodes of the same color, see Section 2.1) together, after each column is negated with probability $1/2$. This can be represented as $B = IC$, where C is $N \times M$.

For the multi-scale construction we apply this splicing operation at each scale, $B_s = W_s C_s$. This construction attempts to distribute the errors evenly across scales. The correlation length between nodes i and j at scale s (where P is filtered with the wavelet at scale s) changes with scale. For common GMRF models, such as thin-plate and thin-membrane, we expect it to double as we move to a coarser scale[1]. We combine only those columns of W_s that lead to small interference of any pair of vectors b_i and b_j at that scale. Thus, due to longer correlations, for coarser scales we can only combine half as many vectors. However, the number of columns of W_s also decreases by 2 as we move to a coarser scale, so the resulting number of columns of B_s stays the same for all s. With such a construction, the number of columns of B will be $O(\log_2 N)$ instead of N for the full wavelet basis W.

Properties of the multi-scale approximation $\hat{P}$. In the single-scale case we showed that $\hat{P}$ is unbiased, and analyzed the variance. We extend these results to 2D. The error is equal to

$$E = P - \hat{P} = P - PBB' = P(WW' - BB'). \quad (6)$$

We are only interested in the variances, hence in $\text{diag}(E)$. We can decompose E across scale: $E = P(WW' - BB') = P\sum_{s=1}^{N_s}(W_s W_s' - B_s B_s')$. Let $c_s(i)$ be the sign assigned to column i of W_s, let $\mathcal{C}(i;s)$ be the set of columns that get merged with i, and denote column i of W_s by $W_s(i)$. Then $E[B_s B_s'] = W_s W_s' + \sum_{i\neq j}^{j\in\mathcal{C}(i;s)} E[W_s(i)W_s(j)' c_s(i)c_s(j)] = W_s W_s'$. The error terms cancel out because $E[c_s(i)c_s(j)] = 0$ for $i \neq j$. Thus, the approximation $\hat{P}$ is unbiased.

To analyze the variance we consider $\text{tr}(E)$. We have:

$$\text{tr}(E_s) = \text{tr}(P(W_s W_s' - B_s B_s')) = \text{tr}(W_s' P W_s - B_s' P B_s) = \text{tr}(W_s' P W_s - C_s' W_s' P W_s C_s) = \text{tr}(P_s(I - C_s C_s')). \quad (7)$$

where $P_s \triangleq W_s' P W_s$ is the covariance of the wavelet coefficients at scale s. Then, via the same analysis as in [3]:

$$\text{tr}(P_s(I - C_s C_s')) = \sum_{i\neq j, j\in\mathcal{C}(i;s)} P_s(i,j)c_s(i)c_s(j). \quad (8)$$

Putting all the scales together, we have $\text{tr}(E) = \sum_s \text{tr}(E_s) = \sum_s \sum_{i\neq j, j\in\mathcal{C}(i;s)} P_s(i,j)c_s(i)c_s(j)$. Now, $\text{Var}(\text{tr}(E)) \leq \sum_s \sum_{i\neq j, j\in\mathcal{C}(i;s)} P_s(i,j)^2$, as $c_s(i)$ are i.i.d. with variance 1. Now, assuming that the errors at different nodes are only weakly correlated, which we justify with experiments in Section 4, we have $\sum_i \text{Var}(E_{ii}) \approx \text{Var}(\sum E_{ii}) = \text{Var}(\text{tr}(E)) \leq \sum_s \sum_{i\neq j, j\in\mathcal{C}(i;s)} P_s(i,j)^2$. By combining columns of W_s such that $P_s(i,j)$ is small for interfering terms, the total variance of the errors is kept small.

[1]The correlation length of wavelet coefficients $d_{s;k}$ of self-similar and certain stationary processes is roughly the same at all scales: $E[d_{s;k_1} d_{s;k_2}] \propto |k_1 - k_2|^\alpha$, at any scale, where α depends on the regularity of the wavelet, and the properties of the random process [5]. However, lag of $k_1 - k_2$ at scale s translates into $2^{(s-1)}(k_1 - k_2)$ at the finest scale.

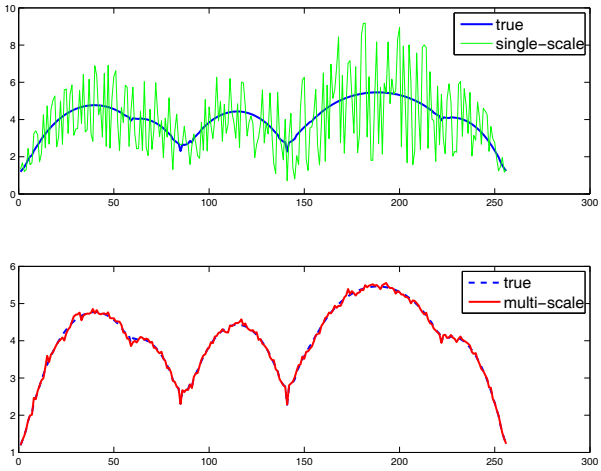

Fig. 4. Low-rank variance approximation for a 1D signal, $N = 256$: (top) single-scale B, with $M = 32$, and (bottom) multi-scale B, with $M = 28$, Coifman wavelet basis.

Constructing multi-scale B for 2D - separable wavelets. We start with the separable construction of a 2D wavelet basis, which takes outer products between 1D functions. We combine translations of wavelet and scaling functions at each scale to create a family of triplets [4]:

$$\psi^{(1)}_{s;k_1,k_2}(x,y) = \frac{1}{2^s}\phi(2^{-s}x-k_1)\psi(2^{-s}y-k_2), \quad (9)$$
$$\psi^{(2)}_{s;k_1,k_2}(x,y) = \frac{1}{2^s}\psi(2^{-s}x-k_1)\phi(2^{-s}y-k_2),$$
$$\psi^{(3)}_{s;k_1,k_2}(x,y) = \frac{1}{2^s}\psi(2^{-s}x-k_1)\psi(2^{-s}y-k_2).$$

We stack $\psi^{(i)}_{s;k_1,k_2}$ as columns of W to create an orthogonal basis. To produce a corresponding matrix B, we apply the same operations to spliced wavelets and scaling functions, i.e., to columns of W_sC_s, and S_sC_s.

4. RESULTS

Our first experiment involves a 1D model with length $N = 256$, with connections from each node to nodes up to 4 steps away. Noisy observations are added at a few randomly selected nodes. The J matrix is close to singular, and the correlation length in the model is long. In Figure 4 we illustrate the results using both the single-scale (top plot) and the multi-scale (bottom plot) 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 useless. For the multi-scale case, using a smaller matrix B with $M = 28$, constructed by splicing a Coifman wavelet basis, we are able to find very accurate variance approximations.

Next we apply the approach to a 2D thin-membrane model of size 256×256, with sparse noisy measurements. We use

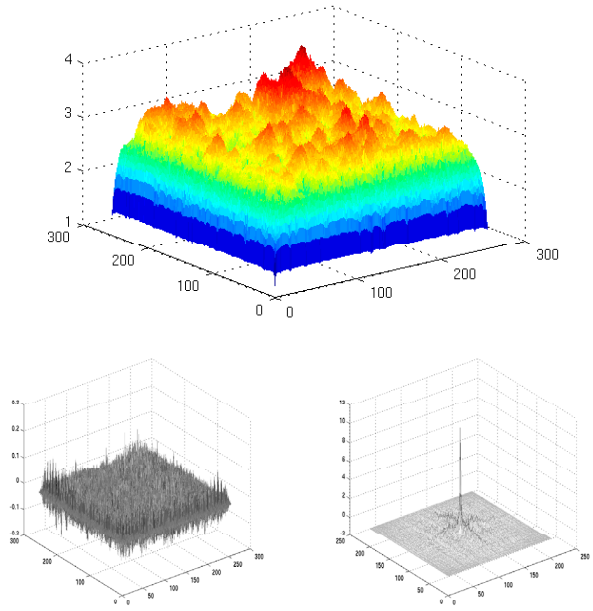

Fig. 5. 2D thin-membrane example: (top) approximate variances, (bottom left) errors, and (bottom right) 2D auto-correlation of errors.

separable Coifman wavelets, and the B matrix is 65536×304. This is a very significant reduction in the number of columns, compared to W. The results appear in Figure 5 (top plot). Our approximate solution is a close match to the exact solution, which can still be computed for models of this size (our multi-scale approach can also be applied to larger problems, where exact computation is intractable). The errors and their 2D auto-correlation appear in Figure 5 (bottom left and right, respectively). The errors are only weakly correlated, supporting our error analysis in Section 3.

5. REFERENCES

[1] H. Rue and L. Held, *Gaussian Markov Random Fields Theory and Applications*, Chapman and Hall, CRC, 2005.

[2] J.K. Johnson and A.S. Willsky, "A recursive model-reduction method for approximate inference in Gaussian Markov random fields," *IEEE Trans. Imag. Proc.*, 2006 (in review).

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

[4] S. Mallat, *A wavelet tour of signal processing*, Academic Press, 1998.

[5] B. Vidakovic, *Statistical Modeling by wavelets*, John Wiley and Sons, 1999.