MAXIMUM ENTROPY RELAXATION FOR MULTISCALE GRAPHICAL MODEL SELECTION

Myung Jin Choi, Venkat Chandrasekaran, and Alan S. Willsky

Laboratory for Information and Decision Systems
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139

ABSTRACT

We consider the problem of learning multiscale graphical models. Given a collection of variables along with covariance specifications for these variables, we introduce hidden variables and learn a sparse graphical model approximation on the entire set of variables (original and hidden). Our method for learning such models is based on maximizing entropy over an exponential family of graphical models, subject to divergence constraints on small subsets of variables. We demonstrate the advantages of our approach compared to methods that do not use hidden variables (which do not capture long-range behavior) and methods that use tree-structure approximations (which result in blocky artifacts).

Index Terms— Graphical models, multiscale models, maximum entropy principle, model selection, hidden variables

1. INTRODUCTION

Multiscale priors have been widely used in large-scale signal processing applications to model statistical dependencies among variables [1]. Multiscale representations can lead to efficient algorithms, and are useful when the phenomenon of interest, the available data, or the estimation objectives involve behavior at multiple scales. In this paper, we view multiscale models as graphical models [2] in which the nodes of the graph represent random variables and the edge structure specifies the conditional independence (Markov) properties among the variables.

Multiscale models defined on trees lead to efficient linear-complexity estimation and inference algorithms [1]. However, tree-structured models possess limited modeling capabilities; for example, tree-structured Gaussian models often lead to blocky artifacts in the resulting covariance approximations [1]. In order to model a richer class of statistical dependencies among variables, one often requires graphical models containing cycles. As a result, significant effort has been, and still is being, devoted to tractable approximate estimation algorithms for loopy graphical models. Motivated by the development of such algorithms, we consider the problem of *learning* multiscale graphical models containing cycles. Specifically, given a collection of variables x and a desired covariance structure among these variables η_x, we construct additional coarse-scale *hidden* variables z and build a multiscale graphical model on the entire collection $\{x, z\}$ with two objectives. First, we attempt to find a global multiscale model on $\{x, z\}$ so that the marginal statistics on variables x are close to the specified η_x. Second, we seek a global structure that is a *sparse* graphical model on $\{x, z\}$, which would permit tractable approximate estimation and inference algorithms. Our approach for finding such models is based on maximizing entropy subject to divergence constraints on small subsets of variables [3]. When appropriately viewed in the context of exponential families, this formulation reduces to a convex optimization program that can be efficiently solved using a primal-dual interior-point algorithm.

In Section 2, we provide a brief background on graphical models and exponential families. We discuss our approach for learning multiscale graphical models in Section 3. We also provide simulation results that demonstrate the advantages of the multiscale framework in terms of modeling and estimation performance. Our discussion is focused on Gaussian models, with results for discrete models deferred to a longer paper. Finally, we conclude with a brief discussion in Section 4.

2. BACKGROUND

2.1. Graphical models and Exponential families

A *graphical model* [2] is a collection of random variables indexed by the vertices of a graph $\mathcal{G} = (V, \mathcal{E})$; each vertex $v \in V$ corresponds to a random variable x_v, and where for any $A \subset V$, $x_A = \{x_v | v \in A\}$. The set $\mathcal{E}$ is some subset

The research described in this paper was supported in part by Shell International Exploration and Production, Inc. and in part by the Army Research Office under Grant W911NF-05-0-0207. Myung Jin Choi is supported in part by Samsung Scholarship.

of $\binom{V}{2}$, the set of all pairs of edges.[1] A distribution $p(x)$ is *Markov* with respect to a graph $\mathcal{G} = (V, \mathcal{E})$ if for any subsets $A, B \subset V$ that are separated by some $S \subset V$ (each path from a node in A to a node in B passes through a node in S), the subset of variables x_A is conditionally independent of x_B given x_S, i.e. $p(x_A, x_B|x_S) = p(x_A|x_S) \cdot p(x_B|x_S)$.

A distribution being Markov with respect to a graph implies that it can be decomposed into local functions in a very particular way [2]. We elaborate on this connection for exponential family distributions [4]. Let $\mathbb{X}$ be either a continuous or discrete sample space. We consider parametric families of probability distributions with support $\mathbb{X}^{|V|}$ defined by

$$p_\theta(x) = \exp\{\theta^T \phi(x) - \Phi(\theta)\}, \tag{1}$$

where $\phi : \mathbb{X}^{|V|} \to \mathbb{R}^d$ are the *sufficient statistics*, θ are the *exponential parameters*, and $\Phi(\theta) = \log \int \exp(\theta^T \phi(x)) dx$ is the *log-partition function*.[2] The family is defined by the set $\Theta \triangleq \left\{\theta \in \mathbb{R}^d : \Phi(\theta) < \infty\right\} \subset \mathbb{R}^d$ of all normalizable θ. A class of graphical models is obtained by defining the collection of statistics ϕ to be *local* functions over small subsets of variables. Let $\phi = \{\phi_v(x_v), v \in V\} \cup \{\phi_E(x_E), E \in \binom{V}{2}\}$ define a collection of node and pairwise statistics, where each $\phi_E(x_E)$ (or $\phi_v(x_v)$) is only a function of the variables x_E (or variable x_v). Specializing the Hammersley-Clifford theorem [2] to such exponential family distributions, we have that if p_θ is Markov with respect to $\mathcal{G} = (V, \mathcal{E})$, then θ is sparse according to $\mathcal{G}$, i.e. $\theta_E = 0$ for $E \notin \mathcal{E}$.

By taking expectations of the statistics ϕ with respect to $p_\theta(x)$, we obtain the *moment parameters* $\eta = \mathbb{E}_{p_\theta}\{\phi(x)\}$. Let $\mathcal{M}$ denote the set of *realizable* moment parameters that can be obtained under expectations with respect to some $\theta \in \Theta$. The above expectation defines a bijective map $\Lambda : \Theta \to \mathcal{M}$; thus, an exponential family distribution has an alternate moment parameterization given by $p_\eta(x) = p_{\Lambda^{-1}(\eta)}(x)$.

Entropy and Divergence: The *entropy* $H(\eta) \triangleq H(p_\eta(x))$ of an exponential family distribution parameterized by moments η is the negative of the convex conjugate of the log-partition function; thus, $H(\eta)$ is concave as a function of η. The *Kullback-Leibler divergence* $D(\eta\|\nu) \triangleq D(p_\eta(x)\|p_\nu(x))$ is the *Bregman* distance induced by the entropy function. As a result, $D(\eta\|\nu)$ is convex with respect to the moment parameters η, keeping the moments ν fixed. We refer the reader to [4] for more background.

Gaussian models as an exponential family: Consider a zero-mean[3] Gaussian graphical model with a symmetric positive-definite covariance matrix P [2]. A natural parameterization for such a model that provides a connection to exponential families is in terms of the *information matrix* $J = P^{-1}$, so that $p(x) \propto \exp\left\{-\frac{1}{2}x^T J x\right\}$. Thus, if $p(x)$ is Markov with respect to $\mathcal{G} = (V, \mathcal{E})$, then $J_{vu} = J_{uv} = 0$ if and only if the edge $\{v, u\} \notin \mathcal{E}$ for every pair of vertices $v, u \in V$. Defining statistics $\phi_v(x_v) = x_v^2, \forall v \in V$, and $\phi_{v,u}(x_v, x_u) = x_v x_u, \forall \{v, u\} \in \binom{V}{2}$, we obtain θ parameters as $\theta = \left\{-\frac{1}{2}J_{vv}, \forall v\right\} \cup \{-J_{vu}, \forall \{v, u\}\}$ and η parameters as $\eta = \{P_{vv}, \forall v\} \cup \{P_{vu}, \forall \{v, u\}\}$. A key point here is that the marginal density for a subset of variables is determined by the corresponding subset of the moment parameters (a principle submatrix of P).

[1] This notion can be generalized to include high-order edges involving more than two variables.

[2] The integral must be replaced by a sum for discrete models.

[3] The mean vector does not play a significant role in model selection.

2.2. Maximum entropy modeling

The maximum-entropy principle states that subject to linear constraints on a set of statistics, the entropy-maximizing distribution among *all* distributions lies in the exponential family based on those statistics used to define the constraints. Consider the following restricted maximum-entropy problem within the framework of exponential family distributions [4]. Let η be the moment parameters of an exponential family, and let η_V and $\eta_\mathcal{E}$ represent the subset of moments corresponding to the set of all vertices V and a set of edges $\mathcal{E}$ respectively. We constrain these moments to be equal to some η_V^* and $\eta_\mathcal{E}^*$:

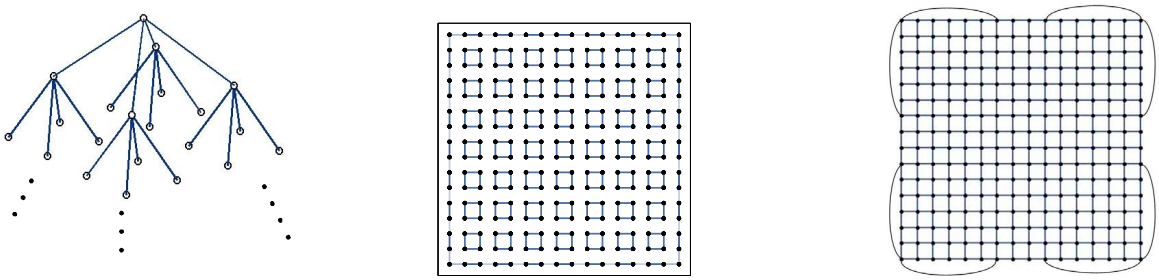
$$\text{(ME)} \quad \begin{array}{ll} \arg\max_{\eta \in \mathcal{M}} & H(\eta) \\ \text{s.t.} & \eta_\mathcal{E} = \eta_\mathcal{E}^*, \eta_V = \eta_V^*. \end{array}$$

Based on the maximum-entropy principle, we can conclude that the optimal distribution (if it exists) of this ME problem over the *entire* exponential family $\{p_{\Lambda^{-1}(\eta)} : \eta \in \mathcal{M}\} = \{p_\theta : \theta \in \Theta\}$ is Markov with respect to the graph $(V, \mathcal{E})$. This suggests that entropy, when used as a maximizing objective function, favors *sparse* graphical models.

Motivated by the maximum-entropy principle, consider the following *relaxed* maximum-entropy formulation [3]:

$$\text{(MER)} \quad \begin{array}{ll} \arg\max_{\eta \in \mathcal{M}} & H(\eta) \\ \text{s.t.} & D_E(\eta_{\bar{E}}\|\eta_{\bar{E}}^*) \le \delta_E, \ \forall E \in \mathcal{E} \\ & D_v(\eta_v\|\eta_v^*) \le \delta_v, \ \forall v \in V. \end{array}$$

Here, D_E and D_v are the marginal divergences on $E \in \mathcal{E}$ and $v \in V$ respectively, the edge set $\mathcal{E}$ serves to specify the constraint set, and $\delta = \{\delta_E, E \in \mathcal{E}\} \cup \{\delta_v, v \in V\}$ are a specified set of tolerances on marginal divergences. The moments $\eta_{\bar{E}}$ specify the moments of the marginal distribution of variables x_E; for example, $\eta_{\{\bar{v,u}\}} = \{\eta_v, \eta_u, \eta_{vu}\}$. The moments $\eta_{\bar{E}}^*$ and η_v^* denote the specified statistics on edge E and vertex v respectively. We have that MER is a convex program due to the convexity properties of H and $D(\cdot\|\cdot)$, and the convexity of the set $\mathcal{M}$ [3]. The MER solution is Markov with respect to the graph $\mathcal{G} = (V, \mathcal{E})$. Further, it is also Markov with respect to the graph specified by just the set of *active* edge constraints, i.e. the *subset* of edge constraints in $\mathcal{E}$ that are satisfied with equality by the MER solution. One expects that MER identifies a sparse Markov model within the constraint set, with the degree of sparsity obtained in the solution being controlled by the tolerances δ. When the specified statistics $\eta_{\bar{E}}^*$ and η_v^* are empirical moments computed from samples, using the relaxed constraints with δ avoids parameter overfitting.



(a) (b) (c)

Fig. 1. (a) Multiscale quad-tree model. (b) Finest scale of the multiscale MER solution. (c) Single-scale MER solution.

3. LEARNING MULTISCALE GRAPHICAL MODELS

3.1. Multiscale Modeling

An important goal in constructing a graphical model to approximate a specified set of statistics on a collection of variables is to later use the model for estimation. Algorithms for (approximate) estimation are typically more tractable for sparser graphical models than for densely connected ones. However, when the specified statistics correspond to processes that have long-range correlations (variables at distant spatial locations having high correlation), the resulting graphical model approximations tend to be very densely connected. In order to model such processes, one approach is to introduce additional *hidden* variables at *coarser* scales that capture the long-range statistics, leaving the original variables at the finest scale to capture short-range statistics.

Tree-structured multiscale models (for example, see Figure 1(a)) have been widely studied to provide tractable approximations to processes with long-range correlations [1]. However, the principal limitation with such models is that they lead to blocky covariance approximations [1]. Other multiscale modeling methods include designing *fixed* structures that have *connected* graphs at each scale [5].

We propose a method to *learn* a multiscale graphical model using the MER approach for jointly Gaussian random variables. We define coarse-scale variables that are aggregates of the variables at child nodes as follows:

$$x_p = \frac{1}{|\mathcal{C}(x_p)|} \sum_{c \in \mathcal{C}(x_p)} x_c + v, \qquad (2)$$

where $\mathcal{C}(x_p)$ is the set of child nodes of x_p, and v is zero-mean white Gaussian noise with variance σ^2. Our focus here is on collections of variables that are spatially located on a grid. For example, coarse-scale variables correspond to variables at the non-leaf nodes of the quad-tree in Figure 1(a). The set $\mathcal{C}(x_p)$ for each non-leaf node p are the four immediate children of p.

Given a desired set of covariances on the original collection of variables (at the finest scale), one can compute the covariances between all pairs of nodes (original and newly-defined hidden variables) based on (2). We supply these covariances as input to MER in order to *learn* a graphical model approximation on the *entire* collection of original and hidden variables. We note here that while the coarse-scale variables are defined according to parent-child relationships on a quad-tree, the MER solution can in general contain edges that are not present in the quad-tree.

In order to solve MER in a tractable manner [3], we restrict the Markovianity of the MER solution based on the parent-child relationships that define the aggregate computation in (2). We constrain the edge set $\mathcal{E}$ in the MER problem to allow arbitrary edges within each scale but only edges that belong to the quad-tree to connect variables across scales. Thus, the MER solution is Markov on a *subgraph* of a graph that consists of a quad-tree and fully connected components at each scale (as a result, the MER solution can contain edges within each scale that are not present in the quad-tree).

3.2. Simulation results

We provide experimental results comparing the performance of MER with and without the addition of coarse-scale hidden variables. For more details on primal-dual interior-point methods to solve MER, see [3]. We consider a collection of 256 Gaussian random variables arranged spatially on a 16×16 grid. The variance of each variable is given by $\eta_{x_s} = 1.5$ and the covariance between each pair of variables is given by $\eta_{x_s,x_t} = d(s,t)^{-\frac{1}{2}}$, where $d(s,t)$ is the distance between nodes s and t. Polynomial decay in covariance is typically found in models with long-range correlations (as opposed to exponential decay in processes with short-range correlation).

We solve the MER problem directly with these covariance specifications (i.e. with no hidden variables) using $\delta_v = \gamma$ and $\delta_E = 3\gamma$, with $\gamma = 0.015$ (see [3] for more details on choosing these tolerance parameters). The resulting single-scale MER solution is shown in Figure 1(c). Note that this model is densely-connected, and also has several edges connecting variables that are far away spatially. Next, we introduce coarse-scale hidden variables by aggregating variables

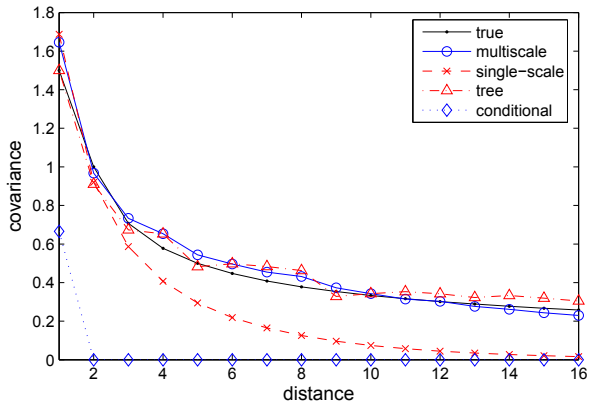


Fig. 2. Covariance behavior of various modeling approaches. The conditional covariance is the covariance of the finest scale in the multiscale MER solution conditioned on coarser scales.

according to the quad-tree structure of Figure 1(a), and by using noise v with a variance of $\sigma^2 = 0.1$ in (2). The resulting MER solution (using $\delta_v = \gamma$ and $\delta_E = 3\gamma$, with $\gamma = 0.01$) consists of a quad-tree that connects variables across scales and additional edges at the finest scale as shown in Figure 1(b) (the only extra edges in addition to the quad-tree are at the finest scale). The interesting aspect about most of these finest scale edges is that they appear to connect variables that do not have the same immediate common parent in the quad-tree. Further, this multiscale model appears to be sparser than the single-scale MER solution of Figure 1(c). We note here that the divergence between the single-scale MER solution and the original distribution is approximately 14.1, and the divergence between the multiscale MER solution marginalized to the finest scale and the original distribution is approximtely 14.6.

Figure 2 describes the covariance behavior of the multiscale and single-scale models learned by MER, and of a simple multiscale quad-tree-structured model that doesn't have any loops as in Figure 1(a).[4] The plot shows the covariance of a "row" of variables. Both the multiscale model computed by MER, and the multiscale tree-structured model capture the long-range correlation behavior in the original model, while the single-scale model learned by MER is poor in terms of capturing such long-range behavior. As discussed previously, the multiscale tree-structured model provides a covariance approximation that is blocky. Further, the *conditional covariance* of the variables at the finest scale of the multiscale MER solution, conditioned on coarser scales, decays very rapidly. Thus, the coarse-scale variables capture long-range behavior, while the variables at the finest scale capture short-range statistics. We conclude that using MER to learn multiscale graphical models provides good modeling capabilities for processes with long-range correlation.

[4]The divergence between this model and the true distribution is 34.2.

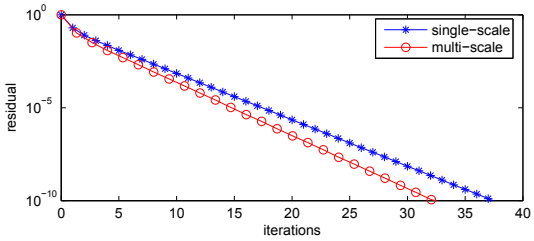


Fig. 3. Comparison of estimation performance.

Finally, in Figure 3 we present a comparison of the estimation performance of the multiscale and single-scale models learned by MER. We generate a sample $y = Cx + n$, where x is generated according to the true underlying distribution, and n is zero-mean white Gaussian noise with variance 1. Here, C is a "selection" matrix that randomly selects only 50% of the entries of x. Such a scenario with sparse, noisy measurements is commonly encountered in many applications [5]. We compute estimates of x using a tractable estimation algorithm called the embedded trees iteration [5]. The plot shows the residual error vs. the number of iterations (each iteration has comparable complexity) for the two models. The multiscale approach provides a 10% gain in convergence rate.

4. CONCLUSION

We describe a multiscale modeling approach based on maximum entropy relaxation that learns multiscale graphical model structure. Our method provides good modeling performance especially for processes that exhibit long-range covariance behavior. We focus on Gaussian models, and defer a similar study of discrete models to a longer paper.

5. REFERENCES

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

[2] S. L. Lauritzen, *Graphical Models*, Oxford University Press, Oxford, U.K., 1996.

[3] J. K. Johnson, V. Chandrasekaran, and A. S. Willsky, "Learning Markov Structure by Maximum Entropy Relaxation," in *Artificial Intelligence and Statistics*, Puerto Rico, March 2007.

[4] S. Amari, "Information geometry on a hierarchy of probability distributions," *IEEE Trans. Info. Theory*, vol. 47, no. 5, pp. 1701–1711, July 2001.

[5] M. J. Choi and A. S. Willsky, "Multiscale Gaussian graphical models and algorithms for large-scale inference," in *IEEE Statistical Signal Processing Workshop*, Madison, Wisconsin, August 2007.