# *Quantifying Nonlocal Informativeness in High-Dimensional, Loopy Gaussian Graphical Models*

# Quantifying Nonlocal Informativeness in High-Dimensional, Loopy Gaussian Graphical Models

**Daniel Levine**
Lab. for Information and Decision Syst.
Massachusetts Institute of Technology
Cambridge, MA 02139
dlevine@mit.edu

**Jonathan P. How**
Lab. for Information and Decision Syst.
Massachusetts Institute of Technology
Cambridge, MA 02139
jhow@mit.edu

## Abstract

We consider the problem of selecting informative observations in Gaussian graphical models containing both cycles and nuisances. More specifically, we consider the subproblem of quantifying conditional mutual information measures that are *nonlocal* on such graphs. The ability to efficiently quantify the information content of observations is crucial for resource-constrained data acquisition (adaptive sampling) and data processing (active learning) systems. While closed-form expressions for Gaussian mutual information exist, standard linear algebraic techniques, with complexity cubic in the network size, are intractable for high-dimensional distributions. We investigate the use of embedded trees for computing nonlocal pairwise mutual information and demonstrate through numerical simulations that the presented approach achieves a significant reduction in computational cost over inversion-based methods.

## 1 INTRODUCTION

In resource-constrained inferential settings, uncertainty can be efficiently minimized with respect to a resource budget by acquiring or processing the most informative subset of observations – a problem known as *active inference* (Krause and Guestrin, 2005; Williams et al., 2007). Yet despite the myriad recent advances in both understanding and streamlining inference through probabilistic graphical models (Koller and Friedman, 2009), there does not exist a comparable wealth of knowledge regarding how information measures propagate on these graphs. This paper considers the problem of efficiently quantifying a measure of informativeness across nonlocal pairings in a loopy Gaussian graphical model.

This paper assumes a model has been provided, and the ensuing goal is to interpret relationships within this model in the context of informativeness. This assumption is motivated by the hypothesis that, regardless of the specific sensing modalities or communication platforms used in an information collection system, the underlying phenomena can be described by *some* stochastic process structured according to a probabilistic graphical model. The sparsity of that model determines the efficiency of inference procedures. In contrast to methods for estimating information measures directly from raw data (e.g., Kraskov et al., 2004), the approach of this paper does not require the prior enumeration of interaction sets that one wishes to quantify, and the presented algorithm computes *conditional* information measures that account for statistical redundancy between observations.

This paper specifically addresses the common issue of nuisances in the model – variables that are not of any extrinsic importance, but act as intermediaries between random variables that are either observable or of inferential interest. Marginalization of nuisances can be both computationally expensive and detrimental to the sparsity of the graph, which, in the interest of efficient model utilization, one wishes to retain. Ignoring nuisances by treating them as relevant can result in observation selectors fixated on reducing uncertainty in irrelevant portions of the underlying distribution (Levine and How, 2013). In terms of information quantification, nuisances can induce nonlocality in the sense that observations and relevant latent variables are not adjacent in the graph, motivating the study of how information measure propagate through graphical models.

In this paper, we investigate the use of embedded trees (Sudderth et al., 2004) for efficiently quantifying nonlocal mutual information in loopy Gaussian graphs. The formal problem statement and a characterization thereof is described in Section 2. Some preliminary material and prior algorithmic technologies are reviewed in Section 3. Our method for quantify mutual information using embedded trees, ET-MIQ, is described in Section 4 and demonstrated through experimental results in Section 5. A discussion of ET-MIQ in comparison to alternative methods and in anticipation of future extensions is provided in Section 6.

## 2 PROBLEM STATEMENT

Let $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ be a collection of $N$ random variables (or disjoint subvectors) with joint distribution $p_{\mathbf{x}}(\cdot)$. Let index set $\mathcal{V} = \{1, \ldots, N\}$ be partitioned such that $\mathcal{V} = \mathcal{U} \cup \mathcal{S}$, where $\mathcal{U} \subset \mathcal{V}$ indexes latent (unobservable) variables, and where $\mathcal{S} \subset \mathcal{V}$ indexes observable variables, whose realizations $\mathsf{x}_s = x_s, s \in \mathcal{S}$ may be obtained by expending some resource. Let $c : 2^{\mathcal{S}} \to \mathbb{R}_{\geq 0}$ be the cost function that maps subsets of observable variables to a resource cost, and let $\beta \in \mathbb{R}_{\geq 0}$ be a resource budget. Given a subset $\mathcal{R} \subseteq \mathcal{U}$ of *relevant latent variables*, which are of inferential interest and about which one wishes to reduce uncertainty, the general *focused active inference* problem (Levine and How, 2013) is

$$\begin{aligned} \text{maximize}_{\mathcal{A} \subseteq \mathcal{S}} \quad & I(\mathbf{x}_{\mathcal{R}}; \mathbf{x}_{\mathcal{A}}) \\ \text{s.t.} \quad & c(\mathcal{A}) \leq \beta, \end{aligned} \tag{1}$$

where $I(\cdot, \cdot)$ is the mutual information measure (cf. Section 3.3).

It is well known that (1) is **NP**-hard (Ko et al., 1995; Krause and Guestrin, 2009). Despite this, suboptimal heuristics such as greedy selection have been analyzed in the context of submodularity (Nemhauser et al., 1978), leading to various performance bounds (Golovin and Krause, 2010; Krause and Guestrin, 2005; Williams et al., 2007). In the focused case, where there are nuisances $\mathcal{U} \setminus \mathcal{R}$ in the problem, the objective in (1) is in general not submodular (Krause et al., 2008), although online-computable performance bounds can be established through submodular relaxations (Levine and How, 2013).

However, efficient computation of the mutual information objective in (1) has remained elusive for all but simple models – symmetric discrete distributions (Choi et al., 2011) and Gaussian trees (Levine and How, 2013). Just as covariance analysis in the Kalman filtering framework can be used to anticipate the uncertainty evolution in a linear-Gaussian state space model, which is Markov to a minimal tree-shaped Gaussian graph, this paper aims to provide a general *preposterior analysis* of uncertainty reduction in nontree Gaussian systems.

This paper specifically considers the class of Gaussian distributed vectors $\mathbf{x} \sim \mathcal{N}^{-1}(\mathbf{0}, J)$ with inverse covariance matrices $J$, each of which is Markov to an undirected graph with cycles. For large $N$, evaluating the MI objective in (1) via matrix inversion is cubic in $N$, which may be prohibitively expensive. The aim of this paper is explicating an iterative algorithm for computing MI whose complexity per iteration is linear in $N$, and whose convergence is often subquadratic in $N$, leading to a relative asymptotic efficiency over naïve linear algebraic techniques.

## 3 BACKGROUND

### 3.1 MARKOV RANDOM FIELDS

A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$ linking pairs of vertices, can be used to represent the conditional independence structure of a joint distribution $p_{\mathbf{x}}(\cdot)$ over a collection $\mathbf{x} = (\mathsf{x}_1, \ldots, \mathsf{x}_N)$ of $N$ random variables (or disjoint random subvectors). This paper considers the class of distributions represented by undirected graphs, also known as Markov random fields (MRFs).

The topology of an MRF can be characterized, in part, by its set of paths. A path is a sequence of distinct adjacent vertices $(v_1, \ldots, v_m)$ where $\{v_k, v_{k+1}\} \in \mathcal{E}$, $k = 1, \ldots, m-1$. If for any two distinct vertices $s, t \in \mathcal{V}$ there is more than one path joining $s$ to $t$, then $\mathcal{G}$ contains a cycle. A graph without cycles is called a *tree* (or, if it is disconnected, a *forest*).

An MRF can represent conditional independences of the form given by the global Markov condition: For disjoint subsets $A, B, C \subset \mathcal{V}$, $\mathbf{x}_A \perp\!\!\!\perp \mathbf{x}_B \mid \mathbf{x}_C$ iff $A$ and $B$ are graph-separated by $C$ (all paths between a vertex in $A$ and a vertex in $B$ must pass through $C$). The edge set $\mathcal{E}$ satisfies the pairwise Markov property: For all $i, j \in \mathcal{V}$, $\{i, j\} \notin \mathcal{E}$ iff $\mathsf{x}_i \perp\!\!\!\perp \mathsf{x}_j \mid \mathbf{x}_{\mathcal{V} \setminus \{i,j\}}$. A distribution is said to be Markov with respect to a graph $\mathcal{G}$ if it satisfies the conditional independences implied by $\mathcal{G}$.

### 3.2 INFERENCE ON GAUSSIAN MRFS

A multivariate Gaussian distribution in the information form $p_{\mathbf{x}}(\mathbf{x}) \propto \exp\{-\frac{1}{2}\mathbf{x}^T J \mathbf{x} + \mathbf{h}^T \mathbf{x}\}$, with (symmetric, positive definite) *precision* or *inverse covariance matrix* $J$ and potential vector $\mathbf{h}$, is Markov with respect to a Gaussian MRF (GMRF; Speed and Kiiveri, 1986) if $J$ satisfies the sparsity pattern of $\mathcal{E}$: $(J)_{i,j} = (J)_{j,i}^T \neq \mathbf{0} \Leftrightarrow \{i, j\} \in \mathcal{E}$. The parameters of the information form are related to the covariance $P = J^{-1}$ and mean $J^{-1}\mathbf{h}$. Thus, estimating the mean of a Gaussian is equivalent to solving the system of equations

$$J\hat{\mathbf{x}} = \mathbf{h}. \tag{2}$$

Assume without loss of generality[1] that each component $\mathsf{x}_i$ of $\mathbf{x}$ is a subvector of dimension $d \in \mathbb{N}^+$, whereby $J \in \mathbb{R}^{Nd \times Nd}$ can be partitioned into an $N \times N$ grid of $d \times d$ block submatrices. Solving (2) by inverting $J$ requires $\mathcal{O}((Nd)^3)$ operations, which can be prohibitively expensive for large $N$. If the graph contains no cycles, then Gaussian belief propagation (GaBP) (Pearl, 1988; Weiss and Freeman, 2001) can be used to compute the conditional mean, as well as marginal variances, in $\mathcal{O}(Nd^3)$, providing a significant computational savings for large $N$. For graphs

[1] Extension to the case of varying subvector dimensions with $d \triangleq \max_{i \in \mathcal{V}} \dim(\mathsf{x}_i)$ is straightforward.

with cycles, various estimation procedures have been recently developed to exploit available sparsity in the graph (cf. Sections 3.4 and 3.5).

Marginalization and conditioning can be conceptualized as selecting submatrices of $P$ and $J$, respectively. Let disjoint sets $A$ and $B$ form a partition of $\mathcal{V} = \{1, \dots, N\}$. The marginal distribution $p_{\mathbf{x}_A}(\cdot)$ over $\mathbf{x}_A$ is parameterized by covariance matrix $(P)_{A,A}$, the block submatrix of $P$ corresponding to the rows and columns indexed by $A$. Similarly, the conditional distribution $p_{\mathbf{x}_A|\mathbf{x}_B}(\cdot|\mathbf{x}_B)$ of $\mathbf{x}_A$ conditioned on $\mathbf{x}_B \coloneqq \mathbf{x}_B$ is parameterized by the $(J)_{A,A}$ block submatrix of $J$. In the inferential setting, one has access to $J$ and not $P$.

### 3.3 MUTUAL INFORMATION

Mutual information (MI) is an information-theoretic measure of dependence between two (sets of) random variables. Its interpretation as a measure of entropy reduction appeals to its use in uncertainty mitigation (Caselton and Zidek, 1984). For disjoint subsets $A, B, C \subset \mathcal{V}$ (with $C$ possibly empty), conditional mutual information is defined as

$$I(\mathbf{x}_A; \mathbf{x}_B|\mathbf{x}_C) \triangleq h(\mathbf{x}_A|\mathbf{x}_C) - h(\mathbf{x}_A|\mathbf{x}_B, \mathbf{x}_C), \qquad (3)$$

where, for continuous random variables $\mathbf{x}_\mathcal{V}$, $h(\cdot)$ is the differential entropy functional

$$h(q_\mathsf{x}(\cdot)) = -\int_\mathcal{X} q_\mathsf{x}(x) \log q_\mathsf{x}(x)\, dx.$$

Note that MI is always nonnegative and is symmetric with respect to its first two arguments. For convenience, we will often use only the index sets (and not the random variables they index) as the arguments of mutual information.

Let $P_{A|C}$ denote the (marginal) covariance of $\mathbf{x}_A$ given $\mathbf{x}_C$. For multivariate Gaussians, the conditional MI (Cover and Thomas, 2006) is

$$I(A; B|C) = \frac{1}{2} \log \frac{\det(P_{A|C}) \det(P_{B|C})}{\det(P_{A\cup B|C})}. \qquad (4)$$

Computing the marginal covariance matrices needed in (4) via matrix inversion (or taking Schur complements) of $J_{A\cup B|C}$ generally requires $\mathcal{O}((Nd)^3)$ operations, even if one is computing *pairwise* MI (i.e., $|A| = |B| = 1$). For Gaussian trees, an efficient algorithm exist for reducing pairwise MI complexity to $\mathcal{O}(Nd^3)$, i.e., linear in the number of vertices (Levine and How, 2013). The main objective of this paper is providing a similar reduction in complexity for loopy Gaussian graphical models.

### 3.4 EMBEDDED TREES

The embedded trees (ET) algorithm was introduced in (Sudderth, 2002; Wainwright et al., 2000) to iteratively compute both conditional means and marginal error variances in Gaussian graphical models with cycles. Although the algorithm requires only the identification of subgraphs on which inference is tractable, and extensions to, for example, embedded polygons (Delouille et al., 2006) and embedded hypergraphs (Chandrasekaran et al., 2008) have been considered, this paper will focus for clarity of discussion on embedded trees.

Let $\mathbf{x} \sim \mathcal{N}^{-1}(\mathbf{0}, J)$ be a Gaussian distributed random vector Markov to an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that contains cycles. Consider an alternatively distributed random vector $\mathbf{x}_\mathcal{T} \sim \mathcal{N}^{-1}(\mathbf{0}, J_\mathcal{T})$ that is of the same dimension as $\mathbf{x}$ but is instead Markov to a cycle-free subgraph $\mathcal{G}_\mathcal{T} = (\mathcal{V}, \mathcal{E}_\mathcal{T})$ of $\mathcal{G}$ (in the sense that $\mathcal{E}_\mathcal{T} \subset \mathcal{E}$). The tree-shaped (and symmetric, positive definite) inverse covariance matrix $J_\mathcal{T}$ can be decomposed as $J_\mathcal{T} = J + K_\mathcal{T}$, where $K_\mathcal{T}$ is any symmetric *cutting matrix* that enforces the sparsity pattern of $J_\mathcal{T}$ by zeroing off-diagonal elements of $J$ corresponding to cut edges $\mathcal{E} \setminus \mathcal{E}_\mathcal{T}$. Since many cutting matrices $K_\mathcal{T}$ will result in a tree-shaped inverse covariance $J_\mathcal{T}$ Markov to $\mathcal{G}_\mathcal{T}$, attention will be restricted to so-called *regular* cutting matrices, whose nonzero elements are constrained to lie at the intersection of the rows and columns corresponding to the vertices incident to cut edges. Note that $K_\mathcal{T}$ can always be chosen such that $\operatorname{rank}(K_\mathcal{T})$ is at most $\mathcal{O}(Ed)$, where $E \triangleq |\mathcal{E} \setminus \mathcal{E}_\mathcal{T}|$ will be used to denote the number of cut edges.

#### 3.4.1 Conditional Means

Given an initial solution $\hat{\mathbf{x}}^{(0)}$ to (2), the single-tree Richardson iteration (Young, 1971) induced by embedded tree $\mathcal{G}_\mathcal{T}$ with cutting matrix $K_\mathcal{T}$ and associated inverse covariance $J_\mathcal{T} = J + K_\mathcal{T}$ is

$$\hat{\mathbf{x}}^{(n)} = J_\mathcal{T}^{-1} \left( K_\mathcal{T} \hat{\mathbf{x}}^{(n-1)} + \mathbf{h} \right). \qquad (5)$$

Thus, each update $\hat{\mathbf{x}}^{(n)}$ is the solution of a synthetic inference problem (2) with precision matrix $\widetilde{J} = J_\mathcal{T}$ and potential vector $\widetilde{\mathbf{h}} = K_\mathcal{T} \hat{\mathbf{x}}^{(n-1)} + \mathbf{h}$. This update requires a total of $\mathcal{O}(Nd^3 + Ed^2)$ operations, where $\mathcal{O}(Nd^3)$ is due to solving $\widetilde{J}\hat{\mathbf{x}}^{(n)} = \widetilde{\mathbf{h}}$ with a tree-shaped graph, and where $\mathcal{O}(Ed^2)$ with $E = |\mathcal{E} \setminus \mathcal{E}_\mathcal{T}|$ is due to forming $\widetilde{\mathbf{h}}$. In the case that $E$ is at most $\mathcal{O}(N)$, the overall complexity *per iteration* is $\mathcal{O}(Nd^3)$. Letting $\rho(D) \triangleq \max_{\lambda \in \{\lambda_i(D)\}} |\lambda|$ denote the spectral radius of a square matrix D, the asymptotic convergence rate of the single-tree iteration (5) is

$$\rho(J_\mathcal{T}^{-1} K_\mathcal{T}) = \rho(I - J_\mathcal{T}^{-1} J), \qquad (6)$$

with convergence to $\hat{\mathbf{x}}$ guaranteed (regardless of $\hat{\mathbf{x}}^{(0)}$) if and only if $\rho(J_\mathcal{T}^{-1} K_\mathcal{T}) < 1$. Inherent in (5) and (6) is a tradeoff in the choice of embedded structure between the tractability of solving $J_\mathcal{T}\hat{\mathbf{x}}^{(n)} = \widetilde{\mathbf{h}}$ and the approximation strength of $J_\mathcal{T} \approx J$ for fast convergence.

The ET algorithm (Sudderth et al., 2004) is conceptualized as a nonstationary Richardson iteration with multiple matrix splittings of $J$. Let $\{G_{\mathcal{T}_n}\}_{n=1}^{\infty}$ be a sequence of embedded trees within $\mathcal{G}$, and let $\{K_{\mathcal{T}_n}\}_{n=1}^{\infty}$ be a sequence of cutting matrices such that $J_{\mathcal{T}_n} = J + K_{\mathcal{T}_n}$ is Markov to $\mathcal{G}_{\mathcal{T}_n}$ for $n = 1, \ldots, \infty$. The nonstationary Richardson update is then

$$\hat{\mathbf{x}}^{(n)} = J_{\mathcal{T}_n}^{-1}\left(K_{\mathcal{T}_n}\hat{\mathbf{x}}^{(n-1)} + \mathbf{h}\right), \tag{7}$$

with error $e^{(n)} \triangleq \hat{\mathbf{x}}^{(n)} - \hat{\mathbf{x}}$ that evolves according to

$$e^{(n)} = J_{\mathcal{T}_n}^{-1}K_{\mathcal{T}_n}e^{(n-1)}. \tag{8}$$

The criterion for convergence is when the normalized residual error $||K_{\mathcal{T}_n}(\hat{\mathbf{x}}^{(n)} - \hat{\mathbf{x}}^{(n-1)})||_2/||\mathbf{h}||_2$ falls below a specified tolerance $\epsilon > 0$. The sparsity of $K_{\mathcal{T}_n}$ permits the efficient computation of this residual.

When $\{G_{\mathcal{T}_n}, K_{\mathcal{T}_n}\}_{n=1}^{\infty}$ is periodic in $n$, a convergence rate analysis similar to (6) is given in (Sudderth et al., 2004). It is also demonstrated that using multiple embedded trees can significantly improve the convergence rate. Online adaptive selection of the embedded tree was explored in (Chandrasekaran et al., 2008) by scoring edges according to single-edge walk-sums and forming a maximum weight spanning tree in $\mathcal{O}(|\mathcal{E}|\log|N|)$.

#### 3.4.2 Marginal Variances

Given that $\mathrm{rank}(K_{\mathcal{T}}) \leq 2Ed$ (Sudderth, 2002), where $E = |\mathcal{E} \setminus \mathcal{E}_{\mathcal{T}}|$, an additive rank-one decomposition

$$K_{\mathcal{T}} = \sum_i w_i u_i u_i^T, \qquad u_i \in \mathbb{R}^{Nd} \tag{9}$$

can be substituted in the fixed-point equation (Sudderth et al., 2004)

$$P = J_{\mathcal{T}}^{-1} + J_{\mathcal{T}}^{-1}K_{\mathcal{T}}P, \tag{10}$$

yielding

$$P = J_{\mathcal{T}}^{-1} + \sum_i w_i (J_{\mathcal{T}}^{-1}u_i)(Pu_i)^T. \tag{11}$$

Solving for the vertex-marginal covariances $P_i = (P)_{i,i}, i \in \mathcal{V}$, which are the block-diagonal entries of $P$, requires:

- solving for the block-diagonal entries of $J_{\mathcal{T}}^{-1}$, with one-time complexity $\mathcal{O}(Nd^3)$ via GaBP;
- solving the synthetic inference problems $J_{\mathcal{T}}z_i = u_i$, for all $\mathcal{O}(Ed)$ vectors $u_i$ of $K_{\mathcal{T}}$ in (9), with one-time total complexity $\mathcal{O}(Nd^3 \cdot Ed) = \mathcal{O}(NEd^4)$ via GaBP;
- solving the synthetic inference problems $Jz_i = u_i$, for all $\mathcal{O}(Ed)$ vectors $u_i$ of $K_{\mathcal{T}}$ in (9), with *per iteration* total complexity of $\mathcal{O}(NEd^4)$ operations via ET conditional means (7);
- and assembling the above components via (11).

Note that there exists a decomposition, alternative to (9), of $K_{\mathcal{T}}$ into $\mathcal{O}(Wd)$ rank-one matrices using a cardinality-$W$ vertex cover of $\mathcal{E} \setminus \mathcal{E}_{\mathcal{T}}$ (where $W \leq E$ for any minimal vertex cover); this alternative decomposition requires solving a symmetric quadratic eigenvalue problem (Sudderth et al., 2004).

### 3.5 COMPETING METHODS

Other methodologies have been proposed to perform inference in loopy graphs. Loopy belief propagation (LBP) is simply parallel belief propagation performed on graphs with cycles; if it converges, it does so to the correct mean but, in general, to incorrect variances (Weiss and Freeman, 2001). Extended message passing augments the original BP messages and provides for convergence to the correct variances, but its complexity is $\mathcal{O}(NL^2)$ in the scalar case, where $L$ is the number of vertices incident to any cut edge, and it requires the full message schedule to be executed to produce an estimate (Plarre and Kumar, 2004). Linear response algorithms can be used to compute pairwise marginal distributions for nonadjacent pairs of vertices, but at a complexity of $\mathcal{O}(N|\mathcal{E}|d^3)$, which may be excessive for large $N$ and $|\mathcal{E}| = \mathcal{O}(N)$ given that conditional MI requires only very specific pairwise marginals (Welling and Teh, 2004).

It is also possible to perform efficient inference if it is known that removal of a subset of $\mathcal{V}$, called a feedback vertex set (FVS), will induce a tree-shaped subgraph (Liu et al., 2012). The resulting belief propagation-like inference algorithm, called feedback message passing, has complexity $\mathcal{O}(Nk^2)$ for scalar networks, where $k$ is the size of the FVS. If the topological structure of the graphical model is well known *a priori*, or if the graph is learned by an algorithm oriented towards forming FVSs (Liu and Willsky, 2013), then identification of an FVS is straightforward. However, if the graphical model is provided without such identification, it may be computationally expensive to form an FVS of reasonable size, whereas finding a spanning tree (as the ET algorithm does) is comparatively simple.

Various graph sparsification methods have been pursued to find useful substructures that can precondition linear systems of equations (e.g., support graph theory (Bern et al., 2006)). Notably, Spielman and Teng (2011) present a spectral sparsification method for the graph Laplacian (which has scalar edge weights) that permits the solution of diagonally dominant linear systems in near-linear time. In contrast, this paper analyzes the ET sparsifier, which operates on edges with potentially *vectoral* weights and does *not* assume diagonal dominance (which would, for example, guarantee the convergence of LBP (Weiss and Freeman, 2001)).

# 4 ET MUTUAL INFORMATION QUANTIFICATION (ET-MIQ)

This section describes the application of embedded trees to efficient iterative computation of nonlocal mutual information measures on loopy Gaussian graphs.

It is typically intractable to enumerate all possible selection sets $\mathcal{A} \in 2^{\mathcal{V}}$ and evaluate the resulting MI objective $I(\mathcal{R}; \mathcal{A})$. Often, one balances tractability with performance by using suboptimal selection heuristics with either a priori or online-computable performance bounds (Krause and Guestrin, 2005; Levine and How, 2013). Starting from an empty selection $\mathcal{A} \leftarrow \emptyset$, the greedy heuristic

$$\begin{aligned} a &\leftarrow \underset{\{y \in \mathcal{S} \setminus \mathcal{A} \,:\, c(y) \leq \beta - c(\mathcal{A})\}}{\operatorname{argmax}} I(\mathcal{R}; y | \mathcal{A}) \\ \mathcal{A} &\leftarrow \mathcal{A} \cup \{a\} \end{aligned} \tag{12}$$

selects one unselected observable variable with the highest marginal increase in objective and continues to do so until the budget is expended. By comparison to (4), the MI evaluations needed to perform a greedy update are of the form

$$I(\mathcal{R}; y | \mathcal{A}) = \frac{1}{2} \log \frac{\det(P_{\mathcal{R}|\mathcal{A}}) \det(P_{\{y\}|\mathcal{A}})}{\det(P_{\mathcal{R} \cup \{y\}|\mathcal{A}})} \tag{13}$$

While inverse covariance matrices obey specific sparsity patterns, covariance matrices are generally dense. Thus two of the determinants in (13) require $\mathcal{O}(|\mathcal{R}|^3 d^3)$ operations to compute. If $|\mathcal{R}|$ is $\mathcal{O}(N)$ (e.g., the graph represents a regular pattern, a constant fraction of which is to be inferred), then such determinants would be intractable for large $N$. One instead fixes some ordering $R$ over the elements of $\mathcal{R}$, denoting by $r_k$ its $k$th element, $R_k = \cup_{i=1}^{k} \{r_i\}$ its first $k$ elements, and appeal to the chain rule of mutual information:

$$\begin{aligned} I(\mathcal{R}; y \mid \mathcal{A}) = I(r_1; y \mid \mathcal{A}) + I(r_2; y \mid \mathcal{A} \cup R_1) + \ldots \\ + I(r_{|\mathcal{R}|}; y \mid \mathcal{A} \cup R_{|\mathcal{R}|-1}). \end{aligned} \tag{14}$$

The advantage of this expansion is twofold. Each term in the summation is a pairwise mutual information term. Given an efficient method for computing marginal covariance matrices (the focus of the remainder of this section), the determinants in (4) can be evaluated in $\mathcal{O}(d^3)$ operations. More pressingly, conditioning in an undirected graphical model removes paths from the graph (by the global Markov property), potentially simplifying the structure over which one must perform the quantification. Therefore, the chain rule converts the problem of evaluating a set mutual information measure $I(\mathcal{R}; y|\mathcal{A})$ into $|\mathcal{R}|$ separate pairwise MI computations that *decrease* in difficulty as the conditioning set expands.

It suffices to describe how to compute *one* of the $|\mathcal{R}|$ terms in the summation (14); the template will be repeated for the other $|\mathcal{R}|-1$ terms, but with a modified conditioning set. In the remainder of this section, it is shown how to efficiently compute $I(r; y|C)$ for all $y \in \mathcal{S} \setminus C$ provided some $r \in \mathcal{R}$ and conditioning set $C \subset \mathcal{V} \setminus \{r\}$. Since conditioning on $C$ can be performed by selecting the appropriate submatrix of $J$ corresponding to $\mathcal{V} \setminus C$, it is assumed for clarity of presentation and without loss of generality[2] that either $C = \emptyset$ or that one is always working with a $J$ resulting from a larger $J'$ that has been conditioned on $C$. The resulting MI terms, in further simplification of (13), are of the form

$$I(r; y) = \frac{1}{2} \log \frac{\det(P_{\{r\}}) \det(P_{\{y\}})}{\det(P_{\{r,y\}})}, \tag{15}$$

where $P_{\{r\}} = (P)_{r,r}$ and $P_{\{y\}} = (P)_{y,y}$ are the $d \times d$ marginal covariances on the diagonal, and where $P_{\{r,y\}}$ is the $2d \times 2d$ block submatrix of the (symmetric) covariance $P = J^{-1}$:

$$P_{\{r,y\}} = \begin{bmatrix} (P)_{r,r} & (P)_{r,y} \\ (P)_{y,r} & (P)_{y,y} \end{bmatrix}.$$

In addition to the marginal covariances on the diagonal, the $d \times d$ off-diagonal cross-covariance term $(P)_{r,y} = (P)_{y,r}^T$ is needed to complete $P_{\{r,y\}}$. If it were possible to efficiently estimate the $d$ columns of $P$ corresponding to $r$, all such cross-covariance terms $(P)_{r,y}, \forall y \in \mathcal{S}$, would be available. Therefore, let $P$ be partitioned into columns $\{p_i\}_{i=1}^{Nd}$ and assume without loss of generality that $r$ corresponds to $p_1, \ldots, p_d$. Let $e_i$ be the $i$th $Nd$-dimensional axis vector (with a 1 in the $i$th position). Then $p_i \equiv P e_i, i = 1, \ldots, d$, can be estimated using the synthetic inference problem

$$J p_i = e_i. \tag{16}$$

Thus, by comparison to (2) and (7), the first $d$ columns of $P$ can be estimated with a complexity of $\mathcal{O}(Nd^4)$ per ET iteration.

Using the results of Section 3.4.2, the marginal variances can be estimated in $\mathcal{O}(NEd^4)$ per iteration, where $E = |\mathcal{E} \setminus \mathcal{E}_{\mathcal{T}_n}|$ is the number of cut edges. One can subsequently form each matrix $P_{\{r,y\}}$, $y \in \mathcal{S}$ in $\mathcal{O}(d^2)$ and take its determinant in $\mathcal{O}(d^3)$. Since $|\mathcal{S}| < N$, the ET-MIQ procedure outlined in the section can be used to iteratively estimate the set $\{I(r; y)\}_{y \in \mathcal{S}}$ with total complexity $\mathcal{O}(NEd^4)$ operations per iteration. Returning to the greedy selection of (12) and the chain rule of (14), given a subset $\mathcal{A} \subset \mathcal{S}$ of previous selections, the set of marginal gains $\{I(\mathcal{R}; y|\mathcal{A})\}_{y \in \mathcal{S} \setminus \mathcal{A}}$ can be estimated in $\mathcal{O}(N|\mathcal{R}|Ed^4)$ operations per iteration.

[2] Alternatively, the unconditioned $J$ can be used by treating conditioned vertices as blocked (not passing messages) and by zeroing the elements of $\mathbf{h}$ and $\hat{\mathbf{x}}^{(n)}$ in (5) corresponding to $C$.

# 5 EXPERIMENTS

## 5.1 ALTERNATIVE METHODS

In order to demonstrate the comparative performance of the ET-MIQ procedure of Section 4, alternative methods for computing mutual information in Gaussian graphs – two based on matrix inversion, and one based exclusively on estimating columns of $P$ – are briefly described.

### 5.1.1 Naïve Inversion

Whenever a mutual information term of the form $I(A;B|C)$ is needed, the `NaïveInversion` procedure conditions $J$ on $C$ and computes the marginal covariance matrices $P_{A\cup B|C}$, $P_{A|C}$, and $P_{B|C}$ of (4) using standard matrix inversion, which is $\mathcal{O}(N^3d^3)$. A greedy selection update, which requires computing marginal information gain scores $\{I(\mathcal{R};y|\mathcal{A})\}_{y\in\mathcal{S}\setminus\mathcal{A}}$, thereby requires $\mathcal{O}(N^3|\mathcal{S}|d^3)$ operations using this procedure.

### 5.1.2 Block Inversion

Intuitively, the `NaïveInversion` procedure appears wasteful even for an inversion-based method, as it repeats many of the marginalization operations needed to form $\{I(\mathcal{R};y|\mathcal{A})\}_{y\in\mathcal{S}\setminus\mathcal{A}}$. The `BlockInversion` procedure attempts to rectify this. Given a previous selection set $\mathcal{A}$, `BlockInversion` conditions $J$ on $\mathcal{A}$ and marginalizes out nuisances $\mathcal{U}\setminus\mathcal{R}$ (along with infeasible observation selections $\{y\in\mathcal{S}\setminus\mathcal{A} \mid c(y) > \beta - c(\mathcal{A})\}$) using Schur complements. The complexity of this approach, for each greedy update, is $\mathcal{O}(|\mathcal{S}|^4+|\mathcal{R}||\mathcal{S}|^3+|\mathcal{R}|^3|\mathcal{S}|+N^3)$. `BlockInversion` has the same worst-case asymptotic complexity of $\mathcal{O}(N^3|\mathcal{S}|d^3)$ as `NaïveInversion` but may achieve a significant reduction in computation depending on how $|\mathcal{R}|$ and $|\mathcal{S}|$ scale with $N$.

### 5.1.3 ColumnET

The `ColumnET` procedure uses nonstationary embedded tree estimation of specific columns of $P$ to compute all information measures. That is to say, no marginal error variance terms are computed (cf. Section 3.4.2). Given a previous selection set $\mathcal{A}$, and an ordering $R$ over $\mathcal{R}$, the columns of $P_{\cdot|\mathcal{A}\cup R_{k-1}}$ corresponding to $\{r_k\}\cup\mathcal{S}\setminus\mathcal{A}$ are estimated via (7) and (16). The complexity of a greedy update using `ColumnET` is $\mathcal{O}(N|\mathcal{R}||\mathcal{S}|d^4)$ operations per ET iteration.

## 5.2 "HOOP-TREE" EXAMPLES

To investigate the performance benefits of ET-MIQ, we consider a subclass of scalar ($d=1$) loopy graphs containing $m$ simple cycles (achordal "hoops") of length $l$, where cycles may share vertices but no two cycles may share edges. The structure of this graph resembles a macro-tree over hoop subcomponents (a "hoop-tree"; see Figure 1). Any embedded tree on this graph must only cut $m$ edges ($E=m$), one for each $l$-cycle. This class of graphs is useful for benchmarking purposes, as it permits randomization without requiring the subsequent enumeration of loops via topological analysis, which may be computationally expensive and thus inefficient for testing.

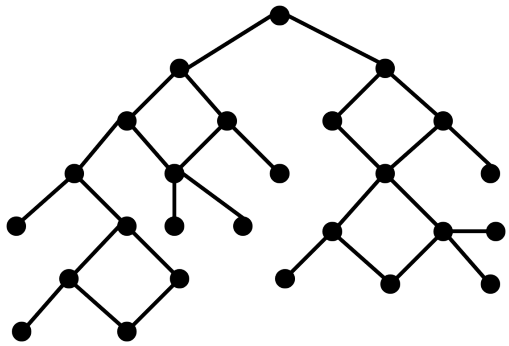

Figure 1: Example of a hoop-tree with 4-vertex cycles.

For each problem *instance*, we generate a random hoop-tree $\mathcal{G}=(\mathcal{V},\mathcal{E})$ of size $|\mathcal{V}|=N$. To generate a corresponding inverse covariance $J$, we sample $(J)_{i,j}\sim\texttt{uniform}([-1,1])$ for each $\{i,j\}\in\mathcal{E}$, and sample $(J)_{i,i}\sim\texttt{Rayleigh}(1)$, with the diagonal rescaled to enforce the positive definiteness of $J$. We then randomly label vertices in $\mathcal{V}$ as belonging to $\mathcal{S}$ or $\mathcal{U}$ (or neither), set a budget $\beta\propto|\mathcal{S}|$, and sample an integer-valued additive cost function $c(\cdot)$ such that $c(s)\sim\texttt{uniform}([1,\gamma\beta])$ for some $\gamma\in[0,1]$ and all $s\in\mathcal{S}$, and such that $c(\mathcal{A})=\sum_{a\in\mathcal{A}}c(a)$ for all $\mathcal{A}\subseteq\mathcal{S}$.

Let $\mathcal{G}_{\mathcal{T}_1}$ and $K_{\mathcal{T}_1}$ be the embedded subtree and associated regular cutting matrix formed by cutting the edge of each $l$-cycle with the highest absolute precision parameter $|(J)_{i,j}|$. Guided by the empirical results of Sudderth et al. (2004), the second embedded tree $\mathcal{G}_{\mathcal{T}_2}$ is selected such that in every $l$-cycle, $K_{\mathcal{T}_2}$ cuts the edge farthest from the corresponding cut edge in the $G_{\mathcal{T}_1}$ (modulo some tie-breaking for odd $l$).

Figure 2 summarizes a comparison of ET-MIQ against `NaïveInversion`, `BlockInversion`, and `ColumnET` in terms of the mean runtime to complete a full greedy selection. Random networks of size $N$ were generated, with $|\mathcal{R}|=5$ and $|\mathcal{S}|=0.3N$. The alternative methods were suppressed when they began to take prohibitively long to simulate (e.g., $N=1200$ for `BlockInversion` and `ColumnET`).

The runtime of ET-MIQ, which vastly outperforms the alternative methods for this problem class, appears to grow superlinearly, but subquadratically, in $N$ (approximately, bounded by $o(N^{1.7})$). The growth rate is a confluence of three factors: the $\mathcal{O}(N|\mathcal{R}|Ed^4)$ complexity per Richardson iteration of updating $\{I(\mathcal{R};y|\mathcal{A})\}_{y\in\mathcal{S}\setminus\mathcal{A}}$; the number

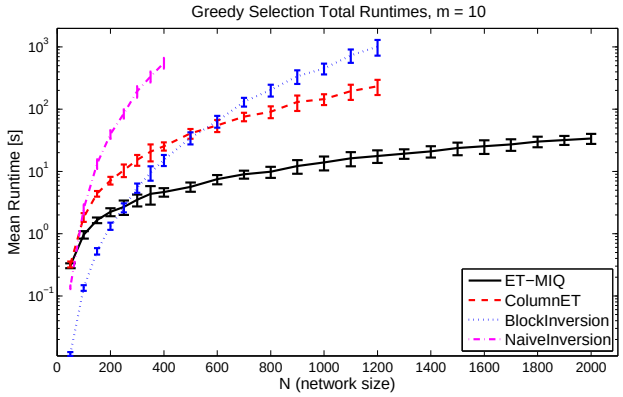


Figure 2: Mean runtime of the full greedy selection as a function of the network size $N$ for randomized loopy graphs with $m = 10$ simple cycles of length $l = 4$.

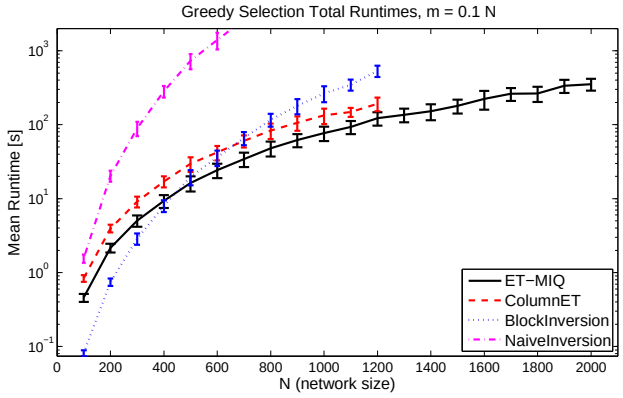


Figure 3: Mean runtime of the full greedy selection as a function of the network size $N$ for randomized loopy graphs with $m = 0.1N$ simple cycles of length $l = 4$. As predicted, in the case where $m = \mathcal{O}(N)$, the ET-based algorithms have the same asymptotic complexity; ET-MIQ has a lower constant factor.

of Richardson iterations until the normalized residual error converges to a fixed tolerance of $\epsilon = 10^{-10}$; and the growth rate of $|\mathcal{S}|$ as a function of $N$, which indirectly affects the runtime through the budget $\beta$ by permitting larger selection sets, and hence more rounds of greedy selection. To better disambiguate the second and third factors, we studied how the number of Richardson iterations to convergence (for a random input $\mathbf{h}$; cf. (2)) varies as a function of $N$ and found no significant correlation in the case where $m$ is constant (not a function of $N$). The median iteration count was 7, with standard deviation of 0.6 and range 5-9 iterations.

We also considered the effect of letting $m$, the number of cycles in the graph, vary with $N$. A runtime comparison for $m = 0.1N$ is shown in Figure 3. Given that $E = m = \mathcal{O}(N)$ and $|\mathcal{R}| = \mathcal{O}(1)$, ET-MIQ has an asymptotic complexity of $\mathcal{O}(N|\mathcal{R}|Ed^4) = \mathcal{O}(N^2)$. Similarly, the complexity of `ColumnET` is $\mathcal{O}(N|\mathcal{R}||\mathcal{S}|d^4) = \mathcal{O}(N^2)$. Figure 3 confirms this agreement of asymptotic complexity, with ET-MIQ having a lower constant factor.

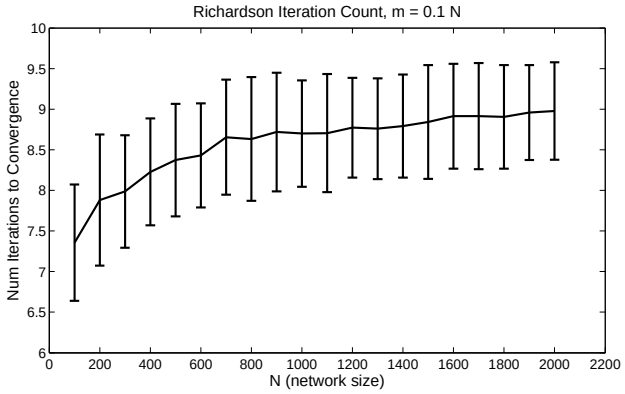


Figure 4: Number of Richardson iterations until convergence, for $m = 0.1N$ cycles.

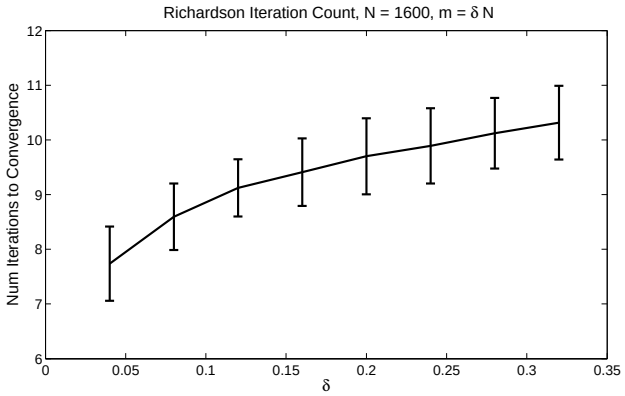


Figure 5: Number of Richardson iterations until convergence, for $N = 1600$ vertices, $m = \delta N$ cycles.

We repeated the convergence study for $m = 0.1N$ and varying $N \in [100, 2000]$ (see Figure 4). The mean iteration count appears to grow sublinearly in $N$; the *actual* increase in iteration count over $N$ is quite modest.

The relationship between the convergence and the problem structure was more clearly illustrated when we fixed a network size of $N = 1600$ and varied the number of 4-vertex cycles $m = \delta N$, for $\delta \in [0.04, 0.32]$ (see Figure 5). The cycle fraction $\delta$ is strongly correlated with the iteration count – and even slightly more correlated with its log – suggesting an approximately linear (and perhaps marginally sublinear) relationship with $\delta$, albeit with a very shallow slope.

## 6 DISCUSSION

This paper has presented a method of computing nonlocal mutual information in Gaussian graphical models containing both cycles and nuisances. The base computations are

iterative and performed using trees embedded in the graph. We assess the proposed algorithm, ET-MIQ, and its alternatives (cf. Sections 3.5 and 5.1) in terms of the asymptotic complexity of performing a greedy update. For ET-MIQ, per-iteration complexity is $\mathcal{O}(N|\mathcal{R}|Ed^4)$, where $N$ is the number of vertices in the network, $\mathcal{R} \subset \mathcal{V}$ is set of relevant latent variables that are of inferential interest, $E$ is the number of edges cut to form the embedded tree, and $d$ is the dimension of each random vector indexed by a vertex of the graph. Let $\kappa$ denote the expected number of Richardson iterations to convergence of ET-MIQ, which is a direct function of the eigenproperties of the loopy precision matrix and its embedded trees and an indirect function of the other instance-specific parameters (number of cycles, network size, etc.). The experimental results of Section 5 suggest that the proposed algorithm, ET-MIQ, achieves significant reduction in computation over inversion-based methods, which have a total complexity of $\mathcal{O}(N^3|\mathcal{S}|d^3)$, where $\mathcal{S} \subset \mathcal{V}$ is the set of observable vertices that one has the option of selecting to later realize.

Based on the asymptotic complexities, we expect ET-MIQ would continue to achieve a significant reduction in computation for large networks whenever $|\mathcal{R}|Ed\kappa = o(N^2|\mathcal{S}|)$. Typically, the vertex dimension $d$ is not a function of the network size. For dense networks ($|\mathcal{E}| = \mathcal{O}(N^2)$), we would not expect significant performance improvements using ET-MIQ; however, it is often the case that $\mathcal{E}$ is sparse in the sense that the number of cut edges $E = \mathcal{O}(N)$. With $|\mathcal{S}| = \mathcal{O}(N)$ (the number of available observations growing linearly in the network size), asymptotic benefits would be apparent for $|\mathcal{R}|\kappa = o(N^2)$. Since we suspect $\kappa$ grows sublinearly (and very modestly) in $N$, and whichever system utilizing the graphical model is free to choose $\mathcal{R}$, we expect that ET-MIQ would be beneficial for efficiently quantifying information in a wide class of active inference problems on Gaussian graphs.

The methods described in this paper are exact in the sense that all mutual information measures are estimated to within a specified tolerance. If the computational cost of quantifying mutual information were constrained (e.g., in a distributed estimation framework with communication costs), it may be of interest to develop algorithms for allowing prioritized approximation depending on how sensitive the overall information reward is to these conditional mutual information terms. In addition to algorithms for adaptively selecting embedded trees to hasten convergence, Chandrasekaran et al. (2008) propose methods for choosing and updating only a subset of variables in each Richardson iteration. If, in an essentially dual problem to (1), the cost of sensor selections were to be minimized subject to a quota constraint on the minimum amount of collected information, the ability to truncate information quantification when a subset of the graph falls below an informativeness threshold would be of potential interest, which we intend to explore in future work.

**Acknowledgements**

The authors thank Dr. John W. Fisher III for helpful discussions during the preparation of this paper. This work was supported by DARPA Mathematics of Sensing, Exploitation and Execution (MSEE).

**References**

M. Bern, J. R. Gilbert, B. Hendrickson, N. Nguyen, and S. Toledo. Support-graph preconditioners. *SIAM Journal on Matrix Analysis and Applications*, 27(4):930–951, 2006.

W. F. Caselton and J. V. Zidek. Optimal monitoring network designs. *Statistics and Probability Letters*, 2(4): 223–227, 1984.

V. Chandrasekaran, J. K. Johnson, and A. S. Willsky. Estimation in Gaussian graphical models using tractable subgraphs: A walk-sum analysis. *IEEE Transactions on Signal Processing*, 56(5):1916–1930, May 2008.

M. J. Choi, V. Y. F. Tan, A. Anandkumar, and A. S. Willsky. Learning latent tree graphical models. *Journal of Machine Learning Research*, 12:1771–1812, May 2011.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, second edition, 2006.

V. Delouille, R. Neelamani, and R. G. Baraniuk. Robust distributed estimation using the embedded subgraphs algorithm. *IEEE Transactions on Signal Processing*, 54: 2998–3010, 2006.

D. Golovin and A. Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *Proc. Int. Conf. on Learning Theory (COLT)*, 2010.

C. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43: 684–691, 1995.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, jun 2004.

A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. Uncertainty in Artificial Intelligence (UAI)*, 2005.

A. Krause and C. Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, 35:557–591, 2009.

A. Krause, A. Singh, and C. Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284, 2008.

D. Levine and J. P. How. Sensor selection in high-dimensional Gaussian trees with nuisances. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 26, pages 2211–2219, 2013.

Y. Liu and A. S. Willsky. Learning Gaussian graphical models with observed or latent FVSs. In *Advances in Neural Information Processing Systems (NIPS)*, volume 26, 2013.

Y. Liu, V. Chandrasekaran, A. Anandkumar, and A. S. Willsky. Feedback message passing for inference in Gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(8):4135–4150, Aug 2012.

G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14:489–498, 1978.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufman, San Mateo, CA, 1988.

K. H. Plarre and P. R. Kumar. Extended message passing algorithm for inference in loopy Gaussian graphical models. *Ad Hoc Networks*, 2:153–169, 2004.

T. P. Speed and H. T. Kiiveri. Gaussian Markov distributions over finite graphs. *Annals of Statistics*, 14(1):138–150, mar 1986.

D. A. Spielman and S.-H. Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

E. B. Sudderth. Embedded trees: Estimation of Gaussian processes on graphs with cycles. Master's thesis, Massachusetts Institute of Technology, February 2002.

E. B. Sudderth, M. J. Wainwright, and A. S. Willsky. Embedded trees: Estimation of Gaussian processes on graphs with cycles. *IEEE Transactions on Signal Processing*, 52(11):3136–3150, November 2004.

M. J. Wainwright, E. B. Sudderth, and A. S. Willsky. Tree-based modeling and estimation of Gaussian processes on graphs with cycles. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 13. MIT Press, Nov 2000.

Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13(10):2173–2200, 2001.

M. Welling and Y. W. Teh. Linear response algorithms for approximate inference in graphical models. *Neural Computation*, 16(1):197–221, 2004.

J. L. Williams, J. W. Fisher III, and A. S. Willsky. Performance guarantees for information theoretic active inference. In M. Meila and X. Shen, editors, *Proc. Eleventh Int. Conf. on Artificial Intelligence and Statistics*, pages 616–623, 2007.

D. M. Young. *Iterative Solution of Large Linear Systems*. Academic, New York, 1971.