Information-theoretic bounds on model selection for Gaussian Markov random fields

Wei Wang*, Martin J. Wainwright†,*, and Kannan Ramchandran*
Department of Electrical Engineering and Computer Sciences*,
and Department of Statistics†
UC Berkeley, Berkeley, CA 94720
{wangwei, wainwrig, kannanr}@eecs.berkeley.edu

***Abstract*—The problem of graphical model selection is to estimate the graph structure of an unknown Markov random field based on observed samples from the graphical model. For Gaussian Markov random fields, this problem is closely related to the problem of estimating the inverse covariance matrix of the underlying Gaussian distribution. This paper focuses on the information-theoretic limitations of Gaussian graphical model selection and inverse covariance estimation in the high-dimensional setting, in which the graph size p and maximum node degree d are allowed to grow as a function of the sample size n. Our first result establishes a set of necessary conditions on $n(p, d)$ for any recovery method to consistently estimate the underlying graph. Our second result provides necessary conditions for any decoder to produce an estimate $\widehat{\Theta}$ of the true inverse covariance matrix Θ satisfying $\|\widehat{\Theta} - \Theta\| < \delta$ in the elementwise ℓ_∞-norm (which implies analogous results in the Frobenius norm as well). Combined with previously known sufficient conditions for polynomial-time algorithms, these results yield sharp characterizations in several regimes of interest.**

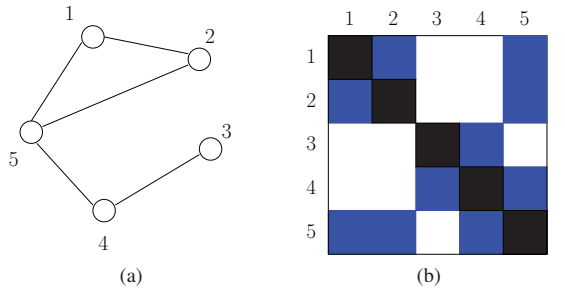
I. Introduction

Markov random fields or undirected graphical models are families of multivariate probability distributions whose factorization and conditional independence properties are characterized by the structure of an underlying graph [1]. Graphical model selection refers to the problem of estimating the graph structure based on observed samples from a Markov random field. This problem arises in a wide variety of settings, including statistical image analysis, natural language processing, and computational biology. In many applications, this problem is of interest under high-dimensional scaling, meaning both the graph size p and the number of samples n are large. Classical methods, such as those based directly on the sample covariance, are known (via random matrix theory [2]) to break down when p/n does not go to zero. Consequently, in the high-dimensional regime where $p \gg n$, additional structure is required in order to obtain consistent estimators. Accordingly, a line of recent work has focused on developing computationally efficient methods to solve this problem by imposing sparsity on the underlying graph. In particular, methods based on ℓ_1-regularization (e.g. [3], [4], [5], [6], [7]) have been shown to yield consistent estimators for Gaussian graphical models, or the associated inverse covariance matrices.

Complementary in nature to such achievable results are the information-theoretic limits associated with any procedure for graphical model selection. Such analysis can serve two purposes. First, it can demonstrate when known polynomial-time algorithms achieve the information-theoretic bounds. Second, it can reveal regimes in which there exists a gap between the performance of current methods and the fundamental limits. With this motivation, some previous work ([8], [9]) has studied both necessary and sufficient conditions for graphical model selection in discrete Markov random fields.

The focus of this paper is on the information-theoretic limits of Gaussian graphical model selection, in which the observed random vector has a multivariate Gaussian distribution. For Gaussian Markov random fields, by the Hammersley-Clifford theorem [1], the model selection problem is equivalent to estimating the off-diagonal sparsity pattern of the inverse covariance matrix. In this paper, we study the ensemble $\mathcal{G}_{d,p}$ of graphs on p vertices with maximum degree at most d, and derive two main results. Our first result is to derive conditions on the sample size n, graph size p, and maximum node degree d that are necessary for any method to correctly recover the underlying graph with probability of error going to zero. Our second result addresses the problem of estimating the inverse covariance matrix Θ, and establishes necessary conditions for any method to produce an estimate $\widehat{\Theta}$ satisfying $\|\widehat{\Theta} - \Theta\| < \delta$. Our results can be compared against known sufficient conditions for graph selection and inverse covariance estimation using ℓ_1-penalized maximum likelihood [7], and reveal regimes in which this polynomial-time algorithm achieves the information-theoretic scaling. One consequence of our results is conditions under which the scaling on the sample size $n = \Omega(d^2 \log p)$ is sharp.

This paper is organized as follows. In Section II, we begin with some background and a precise formulation of the problem. Section III provides the statements of our main results and a discussion of their consequences. Section IV describes a general framework for deriving information-theoretic lower bounds and discusses several approaches for bounding the mutual information that arises in Fano's inequality. Subsections IV-B and IV-C are devoted to the proofs of the necessary conditions for graphical model selection and inverse covariance estimation. Given space constraints, this paper only provides statements and high-level proof ideas; we refer the reader to the technical report [10] for details. We conclude in Section V with a discussion of open directions.



(a) (b)

Figure 1. Illustration of Gaussian Markov random fields. (a) Given an undirected graph, associate a random variable X_i with each vertex i in the graph. A GMRF is the collection of Gaussian distributions over the vector X that respect the structure of the graph. (b) Sparsity pattern of the inverse covariance matrix Θ associated with the GMRF in (a).

II. Background and problem formulation

We begin with some background on Gaussian Markov random fields. We then formulate the graphical model selection problem, which for Gaussian models is directly related to estimation of the inverse covariance matrix. Our goal is to derive information-theoretic lower bounds on the number of samples required for recovery, which apply to any procedure regardless of its computational complexity.

A. Gaussian Markov random fields

Let $X = (X_1, \ldots, X_p)$ be a multivariate Gaussian random vector with zero mean and covariance matrix Σ. Accordingly, its density is determined completely by the inverse covariance matrix $\Theta = \Sigma^{-1}$, and has the form

$$\phi(x; 0, \Sigma) \quad = \quad \frac{1}{\sqrt{(2\pi)^p \det(\Theta^{-1})}} \exp\{-\frac{1}{2} x^T \Theta x\}. \quad (1)$$

For a given undirected graph $G = (V, E)$ with vertex set V and edge set $E \subset V \times V$, we associate a random variable X_i with each vertex $i \in V$. The Gaussian Markov random field associated with the graph G is the family of Gaussian distributions that respect the Markov properties of G. In particular, the off-diagonal sparsity pattern of the inverse covariance matrix Θ is specified by the edge structure of the graph, such that $\Theta_{ij} = 0$ if $(i, j) \notin E$ (see Figure 1).

Given i.i.d. samples from an unknown Markov random field, the problem of estimating the inverse covariance matrix Θ corresponds to recovering the graphical model instance, while the problem of estimating the underlying graph G corresponds to graphical model selection. We define the maximum degree of the graph as

$$d \quad := \quad \max_{i \in V} \Big| \{ j \in V \mid (i, j) \in E \} \Big|, \quad (2)$$

which is equal to the maximum number of non-zeros per row of the inverse covariance matrix Θ. Note that we are not including self-loops at each vertex in the degree count, corresponding to the diagonal entries Θ_{ii}. We often write $\Theta(G)$ to emphasize the graph-based structure of Θ.

B. Classes of graphical models

Let $\mathcal{G}_{p,d}$ be a family of undirected graphs on p vertices with edge sets that have degree at most d. For a given graph $G \in \mathcal{G}_{p,d}$, let $\Sigma(G)$ be the covariance matrix of a Gaussian Markov random field (GMRF) defined by the graph G. By definition, the inverse covariance matrix $\Theta(G)$ must have non-zeros only in positions corresponding to edges in E. In addition to graph structure, the difficulty of graphical model selection also depends on properties of the inverse covariance matrix entries. We measure the minimum value of each matrix $\Theta(G)$ by the function

$$\lambda^*(\Theta(G)) \quad := \quad \min_{(s,t) \in E} \frac{|\Theta_{st}|}{\sqrt{\Theta_{ss}\Theta_{tt}}}, \quad (3)$$

so that it is invariant to rescaling of the data. We study the class $\mathcal{G}_{p,d}(\lambda)$ of Gaussian Markov random fields parameterized by a lower bound λ on the minimum value, defined as the set of probability distributions $\phi_{\Theta(G)} = \phi(0, \Sigma(G))$ where the underlying graph $G \in \mathcal{G}_{p,d}$, the inverse covariance matrix satisfies $\Theta_{st} = 0$ if $(s, t) \notin E$, and $\lambda^*(\Theta(G)) \geq \lambda$.

C. Decoders and error metrics

Suppose we are given n i.i.d. vector samples $X_1^n = \big(X^{(1)}, \ldots, X^{(n)}\big) \in \mathbb{R}^{n \times p}$ from an unknown distribution $\phi_{\Theta(G)}$ in the class $\mathcal{G}_{p,d}(\lambda)$. Graphical model selection refers to the problem of estimating the underlying graph G based on the observations X_1^n. A decoder $\psi : \mathbb{R}^{n \times p} \to \mathcal{G}_{p,d}$ maps the observations X_1^n to an estimated graph $\widehat{G} = \psi(X_1^n)$. We define the error metric between the estimate $\widehat{G}$ and the true underlying graph G using the 0-1 loss function $\mathbb{I}[\psi(X_1^n) \neq G]$. For any decoder ψ, we define the maximal probability of error over the class $\mathcal{G}_{p,d}(\lambda)$ as

$$p_{err}(\psi) \quad := \quad \max_{\phi_{\Theta(G)} \in \mathcal{G}_{p,d}(\lambda)} \mathbb{P}_{\Theta(G)}\big[\psi(X_1^n) \neq G\big], \quad (4)$$

where the error probability

$$\mathbb{P}_{\Theta(G)}\big[\psi(X_1^n) \neq G\big] = \mathbb{E}_{\Theta(G)}\big[\mathbb{I}[\psi(X_1^n) \neq G]\big]$$

is taken with respect to the product distribution $\mathbb{P}_{\Theta(G)}(\cdot) = \phi(\cdot; 0, \Sigma(G))^n$ over n i.i.d. samples.

In contrast to graphical model selection (in which the goal is to recover the support set of $\Theta(G)$), the goal of inverse covariance estimation is to estimate the numerical values of the inverse covariance matrix. More precisely, a decoder $\bar{\psi} : \mathbb{R}^{n \times p} \to \mathcal{G}_{p,d}(\lambda)$ maps the samples X_1^n to an estimate $\widehat{\Theta} = \bar{\psi}(X_1^n)$. We measure the error between the estimate $\widehat{\Theta}$ and the true inverse covariance matrix Θ using the elementwise ℓ_∞-norm $\|\widehat{\Theta} - \Theta\|_\infty := \max_{st} |\widehat{\Theta}_{st} - \Theta_{st}|$, and define the probability of error $\mathbb{P}_{\Theta(G)}\big[\|\widehat{\Theta} - \Theta\|_\infty \geq \delta/2\big]$. The maximal probability of error over the model class $\mathcal{G}_{p,d}(\lambda)$ is then defined as

$$p_{err}(\bar{\psi}) \quad := \quad \max_{\phi_{\Theta(G)} \in \mathcal{G}_{p,d}(\lambda)} \mathbb{P}_{\Theta(G)}\big[\|\widehat{\Theta} - \Theta\|_\infty \geq \delta/2\big]. (5)$$

Although the error metrics for graphical model selection and inverse covariance estimation are closely related, neither

recovery guarantee is strictly stronger than the other. In particular, it is possible to recover the true graph (i.e. $\widehat{G} = G$) even when $\|\widehat{\Theta} - \Theta\|_\infty \geq \delta/2$, since the graph structure is determined only by which entries are zero. Conversely, it is also possible to recover an estimate satisfying $\|\widehat{\Theta} - \Theta\|_\infty < \delta/2$ and still fail to recover the true graph, if for instance there is a non-zero edge weight less than $\delta/2$.

With this set-up, our goal is to derive necessary conditions on the sample size $n(p, d, \lambda)$ for any decoder to reliably recover the underlying graph (or estimate the inverse covariance matrix). We say that recovery is asymptotically reliable over the graphical model class $\mathcal{G}_{p,d}(\lambda)$ if $p_{err} \to 0$ as $n \to \infty$. Our analysis is high-dimensional in nature, in which the graph size p, maximum degree d, and minimum value λ are all allowed to scale arbitrarily as the number of samples n tends to infinity.

III. Main results and consequences

In this section, we state our main results on the information-theoretic limits of Gaussian graphical model selection and inverse covariance estimation, and then discuss some of their consequences.

A. Graphical model selection

We begin with a set of necessary conditions for graphical model selection, applicable to any recovery method regardless of its computational complexity.

Theorem 1. *Consider the class $\mathcal{G}_{p,d}(\lambda)$ of Gaussian Markov random fields with $\lambda \in [0, \frac{1}{2}]$. A necessary condition for asymptotically reliable graphical model selection over the class $\mathcal{G}_{p,d}(\lambda)$ is*

$$n > \max\left\{ \frac{\log\binom{p-d}{2} - 1}{4\lambda^2}, \frac{\log\binom{p}{d} - 1}{\frac{1}{2}\big(\log(1 + \frac{d\lambda}{1-\lambda}) - \frac{d\lambda}{1+(d-1)\lambda}\big)} \right\} \quad (6)$$

The proof of Theorem 1 (given in Section IV-B) constructs restricted ensembles of graphical models and then, viewing the observation process as a communication channel, applies Fano's inequality [11] in order to bound the probability of error. The bounds in Theorem 1 capture how the sample size must grow with graph size p and minimum value λ. In particular, in order for the sum of the edge weights in each neighborhood of the graph to stay bounded, the minimum value must scale as $\lambda = \Theta(\frac{1}{d})$. In this regime, the first bound in Theorem 1 implies that the sample size must scale as $n = \Omega(d^2 \log(p - d))$. For any constant $\lambda \in [0, 1/2]$, the second bound in Theorem 1 scales as $n = \Omega\big(\frac{d \log(p/d)}{\log(1+d\lambda)}\big)$. Moreover, it implies that $n = \Omega(d^{1-\epsilon} \log(\frac{p}{d}))$ for any $\epsilon > 0$.

The information-theoretic bounds in Theorem 1 can be compared with previous work on polynomial-time methods for consistent graph selection. In particular, Ravikumar et al. [7] showed that a sufficient condition for ℓ_1-regularized maximum likelihood to consistently estimate the underlying graph is $n = \Omega((d^2 + \lambda^{-2}) \log p)$. In the regime in which $\lambda = \Theta(\frac{1}{d})$, this scaling matches the information-theoretic bounds in Theorem 1, showing that a polynomial-time method achieves the optimal rates (up to constant factors).

B. Inverse covariance estimation

We now state some necessary conditions for the closely related problem of inverse covariance estimation. Recall that $\|A\|_\infty := \max_{ij} |A_{ij}|$ denotes the elementwise ℓ_∞-norm applied to a matrix.

Theorem 2. *Consider the class of Gaussian Markov random fields $\mathcal{G}_{p,d}(\lambda)$. If there exists an estimator such that $\mathbb{P}[\|\widehat{\Theta} - \Theta\|_\infty < \delta/2] \geq 1/2$ uniformly over choices from $\mathcal{G}_{p,d}(\lambda)$, then we must have*

$$n \quad > \quad \frac{\log\left(\frac{pd}{4}\right) - 2}{4\delta^2}. \quad (7)$$

The proof of Theorem 2, given in Section IV-C, is based on constructing restricted ensembles of graphical models with minimum separation δ, and then applying Fano's inequality [11] to bound the probability of decoding error in distinguishing between such models. Theorem 2 captures how the sample size must grow with the minimum separation between models δ. A consequence of Theorem 2 is that if the recovery error decays at rate $\delta = 1/d$, then the sample size must scale as $n > d^2\big(\log\left(\frac{pd}{4}\right) - 2\big)/4$. Furthermore, Theorem 2 implies that the same necessary condition holds for inverse covariance estimation with other error metrics as well. In particular, let $|\!|\!|A|\!|\!|_F := (\sum_{ij} A_{ij}^2)^{1/2}$ denote the Frobenius norm.

Corollary 1. *A necessary condition for asymptotically reliable inverse covariance estimation, with recovery error at most $\delta/2$ measured in the Frobenius norm, is $n > \frac{\log\left(\frac{pd}{4}\right) - 2}{4\delta^2}$.*

The necessary condition in Theorem 2 can be compared to known sufficient conditions for ℓ_1-regularized maximum likelihood to consistently estimate the inverse covariance matrix. Ravikumar et al. [7] showed that if the sample size satisfies $n > c\, d^2 \log p$ for some constant $c > 0$, then with probability going to one, the ℓ_1-regularized maximum likelihood method produces an estimate $\widehat{\Theta}$ satisfying $\|\widehat{\Theta} - \Theta\|_\infty = O\left(\sqrt{\frac{\log p}{n}}\right)$. Consequently, the performance of the polynomial-time algorithm in [7] matches the scaling of the information-theoretic bound in Theorem 2.

IV. Proof sketches

In this section, we describe our general framework for deriving necessary conditions for consistent graphical model selection and inverse covariance estimation. Our methods are information-theoretic in nature, inspired by techniques that have been used to derive minimax bounds in nonparametric estimation (e.g., [12], [13]).

A. Fano's method

Our general approach is to construct restricted ensembles of graphical models, and then use Fano's method to lower bound the probability of error in each restricted ensemble. Consider a restricted ensemble $\widetilde{\mathcal{G}}$ consisting of $M = |\widetilde{\mathcal{G}}|$ models, and let model index θ be chosen uniformly at random from

$\{1, \ldots, M\}$. Given the observations $\widetilde{X}_1^n \in \mathbb{R}^{n \times \nu}$, the decoder $\widetilde{\psi}$ estimates the underlying graph structure with maximal probability of decoding error defined as

$$p_{err}(\widetilde{\psi}) \;=\; \max_{j=1,\ldots,M} \mathbb{P}_{\widetilde{\Theta}(\widetilde{G}_j)}\Big[\widetilde{\psi}\big(\widetilde{X}_1^n\big) \neq \widetilde{G}_j\Big]. \qquad (8)$$

By Fano's inequality [11], the maximal probability of error over $\widetilde{\mathcal{G}}$ can be lower bounded as

$$p_{err}(\widetilde{\psi}) \;\geq\; 1 - \frac{I\big(\theta; \widetilde{X}_1^n\big) + 1}{\log M}. \qquad (9)$$

In order to make use of the Fano bound, the key is to design ensembles of models for which $\log M$ is large, while the mutual information $I\big(\theta; \widetilde{X}_1^n\big)$ is relatively small. Since it is typically difficult to evaluate the mutual information exactly, we discuss some upper bounds on it.

Entropy-based bound: Define the averaged covariance matrix

$$\bar{\Sigma} \;:=\; \frac{1}{M}\sum_{j=1}^{M} \widetilde{\Sigma}\big(\widetilde{G}_j\big). \qquad (10)$$

The mutual information is upper bounded by $I\big(\theta; \widetilde{X}_1^n\big) \leq \frac{n}{2} F\big(\widetilde{\mathcal{G}}\big)$, where

$$F\big(\widetilde{\mathcal{G}}\big) \;:=\; \log\det\bar{\Sigma} - \frac{1}{M}\sum_{j=1}^{M} \log\det\widetilde{\Sigma}\big(\widetilde{G}_j\big). \qquad (11)$$

KL-based bound: Let $\mathbb{P}_j = f\big(\widetilde{X}_1^n \big| \theta = j\big) = \phi\big(0, \widetilde{\Sigma}\big(\widetilde{G}_j\big)\big)^n$ for $j = 1, \ldots, M$. An alternative bound on the mutual information is given by

$$I\big(\theta; \widetilde{X}_1^n\big) \;\leq\; \mathbb{E}_\theta[D(\mathbb{P}_\theta \| \mathbb{Q})] \qquad (12)$$

for any distribution $\mathbb{Q}$ over $\widetilde{X}_1^n$. Setting $\mathbb{Q} = \phi(0, I_{\nu\times\nu})^n$, the KL distance can be expressed as

$$D(\mathbb{P}_j \| \mathbb{Q}) = \frac{n}{2}\Big\{ \log\det\widetilde{\Theta}\big(\widetilde{G}_j\big) + \mathrm{trace}\big(\widetilde{\Sigma}\big(\widetilde{G}_j\big)\big) - \nu \Big\}. \qquad (13)$$

Note that we are assuming $\log_{\mathrm{e}}$ throughout this paper.

B. Analysis of graphical model selection

We now briefly outline the proofs of the necessary conditions in Theorem 1 on the sample size n as a function of the number of vertices p, maximum degree d and minimum value λ. We obtain two necessary conditions, which can be seen as end points of an entire family of bounds, by analyzing ensembles of graphs in which a subset S of up to d nodes form a clique (i.e. fully connected subset), and the remaining nodes are all isolated.

1) Restricted ensemble A: We begin by deriving the first bound in Theorem 1, which captures how the sample size must grow with the minimum value λ. Consider a family of graphs on p vertices, in which each edge set $E(S,T) = \{(s,t) \,|\, s,t \in S \text{ or } s,t \in T\}$ defines a clique over a subset S of size 2, and another clique over a disjoint subset T of size d. For a given graph $G = (V, E(S,T))$ and a parameter $a \geq 0$, we define the inverse covariance matrix $\Theta(G) := I + a\mathbf{1}_S\mathbf{1}_S^T + a\mathbf{1}_T\mathbf{1}_T^T$, where $\mathbf{1}_S$ and $\mathbf{1}_T$ are the indicator vectors of sets S and T, respectively. The covariance matrix can then be computed as

$$\Sigma(G) \;=\; I - \frac{a}{1+2a}\mathbf{1}_S\mathbf{1}_S^T - \frac{a}{1+da}\mathbf{1}_T\mathbf{1}_T^T. \qquad (14)$$

The resulting class of graphical models is a subset of $\mathcal{G}_{p,d}(\lambda)$ if $\lambda^*(\Theta(G)) = \frac{a}{1+a} \geq \lambda$.

Suppose the decoder is given the indices of the d vertices in T, and the parameter value a. Estimating the underlying graph structure G now amounts to finding the remaining pair of nodes in S, out of $\binom{p-d}{2}$ possibilities. More precisely, given (T, a), the decoder can extract the submatrix of observations $\widetilde{X}_1^n := (X_1^n)_{T^C} \in \mathbb{R}^{n\times(p-d)}$. When the original observations are sampled i.i.d. from the distribution $X^{(i)} \sim N(0, \Sigma)$, the modified observations are distributed according to $\widetilde{X}^{(i)} \sim N(0, \Sigma_{T^CT^C})$. Since the modified covariance matrix is of the form

$$\widetilde{\Sigma}\big(\widetilde{G}\big) \;:=\; \Sigma_{T^CT^C} \;=\; I - \frac{a}{1+2a}\mathbf{1}_S\mathbf{1}_S^T, \qquad (15)$$

the inverse covariance matrix becomes

$$\widetilde{\Theta}\big(\widetilde{G}\big) \;=\; \big(\widetilde{\Sigma}\big(\widetilde{G}\big)\big)^{-1} \;=\; I + a\mathbf{1}_S\mathbf{1}_S^T. \qquad (16)$$

Note that the underlying graph associated with $\widetilde{\Theta}\big(\widetilde{G}\big)$ is $\widetilde{G} := G \setminus T$ (i.e. the graph obtained by removing the vertices in set T and all edges connected to T from graph G). The remaining sub-problem is to determine, given the observations $\widetilde{X}_1^n$, the single edge graph on $(p-d)$ vertices.

Let $\widetilde{\mathcal{G}}$ denote the set of graphs on $(p-d)$ vertices with a single edge, and let $\widetilde{\mathcal{G}}(\lambda)$ denote the associated class of Gaussian Markov random fields with inverse covariance matrices defined as in (16). The proof then applies the Fano bound (9) over this restricted ensemble using the entropy-based bound on mutual information (11).

2) Restricted ensemble B: We now derive the second lower bound in Theorem 1 using an ensemble of d-clique graphs and the entropy-based bound on mutual information (11). Consider the ensemble of graphs consisting of edge sets $E(S) = \{(s,t) \,|\, s,t \in S\}$ with $|S| = d$. For a given edge set $E(S)$ and paramter $a \geq 0$, define the inverse covariance matrix $\Theta(G) := I + a\mathbf{1}_S\mathbf{1}_S^T$, and its associated covariance matrix

$$\Sigma(G) = (\Theta(G))^{-1} \;=\; I - \frac{a}{1+da}\mathbf{1}_S\mathbf{1}_S^T.$$

The cardinality of this restricted ensemble is $\binom{p}{d}$. The proof then follows by applying Fano's inequality (9) using the entropy-based bound (11).

C. Analysis for inverse covariance estimation

In this section, we provide the basic intuition underlying the proof of Theorem 2. We derive a set of necessary conditions for inverse covariance estimation using an ensemble of graphical models which share the same underlying graph, but vary by perturbing a single edge weight. These bounds capture the difficulty of distinguishing between models with inverse covariance matrices that are δ-close, e.g in the elementwise

ℓ_∞-norm. Note that for any two models $\Theta^{(i)}$ and $\Theta^{(j)}$ in our ensemble, since $\|\Theta^{(i)} - \Theta^{(j)}\|_\infty = \delta$ by construction, there does not exist a matrix $\widehat{\Theta}$ satisfying both $\|\widehat{\Theta} - \Theta^{(i)}\|_\infty < \delta/2$ and $\|\widehat{\Theta} - \Theta^{(j)}\|_\infty < \delta/2$. Consequently, we can apply Fano's inequality (9) to bound the probability of error in the restricted ensemble, and the problem is reduced to bounding the mutual information between the model index and the observations.

1) Alternate KL bound: We begin by stating a variant of the KL-based bound on mutual information in (13), using KL distances between all pairs of models in the class, instead of KL distances between each model and the standard Gaussian distribution.

Pairwise KL-based bound: We define the symmetrized Kullback-Leibler divergence,

$$S(\mathbb{P}_i \| \mathbb{P}_j) \quad := \quad D(\mathbb{P}_i \| \mathbb{P}_j) + D(\mathbb{P}_j \| \mathbb{P}_i). \tag{17}$$

By convexity of the KL divergence, we have the following bound on mutual information

$$I\big(\theta; \widetilde{X}_1^n\big) \quad \leq \quad \frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=i+1}^{M} S(\mathbb{P}_i \| \mathbb{P}_j). \tag{18}$$

For Gaussian Markov random fields, a straightforward calculation shows that the symmetrized KL distance is equal to

$$S(\mathbb{P}_i \| \mathbb{P}_j) = \frac{n}{2} \sum_{\ell=1}^{p} \sum_{m=1}^{p} \left(\Theta^{(i)}_{\ell m} - \Theta^{(j)}_{\ell m}\right) \left(\Sigma^{(j)}_{\ell m} - \Sigma^{(i)}_{\ell m}\right) \tag{19}$$

2) Restricted ensemble C: We now use these methods to derive necessary conditions for inverse covariance estimation (stated in Theorem 2), which capture how the sample size must grow with the minimum separation between models δ. Consider a graph on p vertices consisting of $\lfloor \frac{p}{d+1} \rfloor$ cliques, where each clique is of size $(d+1)$. Let $N = \lfloor \frac{p}{d+1} \rfloor$, and let $\{S_1, \ldots, S_N\}$ denote the N cliques with $|S_i| = d+1$. We define the inverse covariance matrix associated with this graph as

$$\bar{\Theta} \quad := \quad I + a \sum_{i=1}^{N} \mathbf{1}_{S_i} \mathbf{1}_{S_i}^T, \tag{20}$$

for some parameter $a \geq 0$. From this base model, we generate an ensemble of Gaussian Markov random fields in which each model perturbs the weight associated with one edge. Thus the model obtained by perturbing the weight on edge (s,t) is defined by the inverse covariance matrix $\Theta^{(i)} := \bar{\Theta} + \delta(\mathbf{1}_{st}\mathbf{1}_{st}^T - I_{st})$ for some parameter $\delta \in (0, \frac{1}{2}]$. Note that we are using $(\mathbf{1}_{st}\mathbf{1}_{st}^T - I_{st})$ to denote the matrix with ones in locations (s,t) and (t,s), and zeros elsewhere. The resulting ensemble of graphical models has cardinality $M = \lfloor \frac{p}{d+1} \rfloor \binom{d+1}{2} \geq \frac{pd}{4}$. The proof then computes the KL-based bound on mutual information in (19) and applies Fano's inequality (9).

V. Discussion

In this paper, we have studied the information-theoretic limits of Gaussian graphical model selection and inverse covariance estimation in the high-dimensional setting. Our analysis yields a set of necessary conditions for consistent graph selection with any method, which matches the scaling of known sufficient conditions [7] for ℓ_1-regularized maximum likelihood in regimes in which the minimum value scales as $\lambda = \Theta(\frac{1}{d})$. The tightness of the bounds in other regimes of λ is an interesting open question. Furthermore, we derived a set of necessary conditions for inverse covariance estimation, which similarly matches the performance of polynomial-time recovery methods [7]. Our results consider recovery in the elementwise ℓ_∞ and Frobenius norms; the tightness of the necessary conditions for recovery in other norms is an interesting open question. At a high-level, our analysis is based on a general framework for deriving information-theoretic bounds in which we view the observation process as a communication channel, and may be applicable to other problems as well.

Acknowledgment

The work of WW and KR was supported by NSF grant CCF-0830788 and AFOSR grant FA9550-09-1-0120. The work of MJW was supported by NSF grants CAREER-CCF-0545862 and AFOSR-09NL184.

References

[1] S. L. Lauritzen, *Graphical Models*. Oxford: Oxford University Press, 1996.

[2] V. A. Marcenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Annals of Probability*, vol. 4, no. 1, pp. 457–483, 1967.

[3] M. Yuan and Y. Lin, "Model selection and estimation in the Gaussian graphical model," *Biometrika*, vol. 94, no. 1, pp. 19–35, 2007.

[4] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2007.

[5] A. d'Aspremont, O. Banerjee, and L. E. Ghaoui, "First order methods for sparse covariance selection," *SIAM Journal on Matrix Analysis and its Applications*, vol. 30, no. 1, pp. 56–66, 2008.

[6] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electronic Journal of Statistics*, vol. 2, pp. 494–515, 2008.

[7] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," Department of Statistics, UC Berkeley, Tech. Rep. 767, November 2008.

[8] N. Santhanam and M. J. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," in *International Symposium on Information Theory (ISIT)*, Toronto, Canada, July 2008.

[9] G. Bresler, E. Mossel, and A. Sly, "Reconstruction of markov random fields from samples: Some easy observations and algorithms," UC Berkeley, Tech. Rep. arXiv, 2008.

[10] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian markov random fields," Department of Statistics, UC Berkeley, Tech. Rep., May 2010.

[11] T. Cover and J. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[12] B. Yu, "Assouad, Fano and Le Cam," *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pp. 423–435, 1996.

[13] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, vol. 27, no. 5, pp. 1564–1599, 1999.