Model Selection in Gaussian Graphical Models from High-Dimensional Missing Data

Yuxin Chen (yxchen@stanford.edu)
June 3, 2011

1 Introduction

Inference in large systems/networks often calls for a precise/approximate characterization of the mutual dependencies of a large set of variables, which can be concisely captured by graphical models. However, the precise graphical models may be unknown a priori, and hence performing model selection (or structure learning) based on the information of a collection of sample/training data becomes an important step to understand the system structure and internal interactions. Faced with the increasingly large number of variables in complex systems, one natural question arises as to how to effectively extract information from the relatively small number of samples.

Due to the high dimensionality of the complex system, however, the samples we obtain may not capture the activities of all relevant factors. Oftentimes, each sample we collect may contain only partial information with a large set of variables missing. Recovering graphical model structure from samples containing partial data entries in a tractable way becomes more challenging. Recently, Chandrasekaran et. al. [CPW10] proposed an algorithm for structure learning with latent variables, under the assumption that the samples belong to a single missing data pattern with the number of latent variables being much smaller than the ambient dimension. In practice, however, sample data may belong to various missing data patterns instead of a single one, and the number of observed data entries in each sample may likely be much smaller than the ambient dimension. An $\ell_1$ regularized EM algorithm which alternates among different missing data patterns has been proposed by Staedler et. al. [SB10] to deal with this issue. Although the algorithm performs reasonably well in practice, there is no theoretical guarantee that this heuristics can find the exact model structure. All these motivate our investigation on how to recover graphical models with incomplete data that belong to multiple missing data patterns.

2 Problem Setup and Notational Convention

In this project, we primarily investigate structure learning of Gaussian graphical models with multiple missing data patterns, and focus only on high-dimensional regime. Specifically, we assume that $\mathbf{X} = (X_1, \cdots, X_p) \sim \mathcal{N}(\mathbf{0}, \Sigma^*)$ is a zero-mean $p$-dimensional Gaussian vector. Its associated Markov random field (undirected graphical model) is specified by an undirected graph with vertex set $V = \{1, 2, \cdots, p\}$ and edge set $E$, which encodes certain conditional independence assumptions. If we define the inverse covariance matrix $\mathbf{K} := (\Sigma^*)^{-1}$, then it is well known that the graphical model has an edge $e \in E$ between node $i$ and $j$ if and only if $(\mathbf{K})_{ij} \neq 0$. It is oftentimes reasonable...
to assume that $\mathbf{K}$ is a sparse matrix, which captures the conditional independence of the underlying model.

Our task is to estimate the edge set $E$ of the underlying graph from $n$ i.i.d. samples $\{x^{(1)}, \ldots, x^{(n)}\}$ drawn from the true distribution, although each sample contains only a subset of entries with the remaining variables missing. Specifically, we suppose that there exist $l$ unique missing data patterns that actually appear in our sample data, and for pattern $i$ we denote by sets $o_j \subseteq \{1, \ldots, p\}$ and $m_j = \{1, \ldots, p\} \setminus o_j$ the set of observed variables and the set of missing variables, respectively. The index set of observed entries belonging to pattern $j$ is denoted by $\mathcal{I}_j$, while $\mathcal{I}_j^c$ denotes the index set of missing entries. Each sample $x^{(i)}$ belongs to one of the $l$ missing data patterns each with equal probability $1/l$.

As mentioned above, $\mathbf{K}$ is assumed to be sparse, i.e. its support set $S \subseteq [p] \times [p]$ satisfies $k_s/p^2 \ll 1$, where $k_s := |S|$. For any index set $A \subseteq [p] \times [p]$, let $\mathcal{P}_A$ be the orthogonal projection onto the linear space of matrices supported on $A$, i.e.

$$\mathcal{P}_A(\mathbf{X}) = \begin{cases} \mathbf{X}_{ij}, & (i, j) \in A, \\ 0, & \text{otherwise}. \end{cases}$$

Besides, for a symmetric matrix $\mathbf{B}$, we define the operator $\mathcal{O}_\mathbf{B}$ such that $\mathcal{O}_\mathbf{B}(\mathbf{X}) = \mathbf{B}\mathbf{X}\mathbf{B}$. We consider the high-dimensional regime where both the number of samples $n$ and the sparsity $k_s$ are allowed to grow with $p$.

### 3 Approach: Convex Relaxation

All information of the Gaussian distribution is specified in the pairwise covariance matrix $\Sigma^*$. When the number of samples $n$ grows, we can obtain an increasingly precise estimate of the pairwise correlation, $\Sigma_{ij} = \mathbb{E}(X_iX_j)$ provided that $(i, j)$ appears in at least one of the missing data pattern. However, the missing data patterns we have available may not cover all pairwise information. Some pair of variables may not belong to the index set of any missing data pattern, leading to some unknown pairwise correlation.

Define the observed index set $\Omega$ of the correlation matrix such that

$$\Omega_{ij} = \begin{cases} 1, & \text{if } \exists s \in \{1, \ldots, l\} \text{ such that } (i, j) \in \mathcal{I}_s; \\ 0, & \text{otherwise}. \end{cases}$$

The observed correlation matrix is defined as $\Sigma_\Omega^* = \mathcal{P}_\Omega(\Sigma^*)$, with the hidden correlation matrix being $\mathbf{H}^* := \mathcal{P}_{\Omega^c}(\Sigma^*) = \Sigma^* - \Sigma_\Omega^*$. Correspondingly, we define the empirical observed correlation matrix $\Sigma_\Omega^n$ as

$$(\Sigma_\Omega^n)_{ij} = \begin{cases} \frac{\sum_{s=1}^n x^{(s)}_i x^{(s)}_j 1\{i, j \in \text{supp}(x^{(s)})\}}{\sum_{s=1}^n 1\{i, j \in \text{supp}(x^{(s)})\}}, & \text{if } (i, j) \in \Omega; \\ 0, & \text{otherwise}. \end{cases}$$

In order to encourage sparsity, the problem can be posed as the following matrix completion problem

$$\text{minimize} \quad \|\Sigma^{-1}\|_0$$

subject to $$(\Sigma)_{ij} - (\Sigma_\Omega^n)_{ij} < \epsilon_n, \quad \forall (i, j) \in \Omega,$$

Non-convex! (1)
where \(\epsilon_n\) depends on the maximum deviation between the empirical estimate and the true correlation. Apparently, this is a non-convex optimization problem, and cannot be solved efficiently in general. This motivates our investigation of convex relaxation as a surrogate.

In particular, the inverse covariance matrix can be rewritten as follows.

\[
K^* = (\Sigma_o^* + H^*)^{-1} = (\Sigma_o^* (I + \Sigma_o^{-1} H^*))^{-1} = (I + \Sigma_o^{-1} H^*)^{-1} \Sigma_o^{-1}
\]

\[
= \Sigma_o^{-1} - \Sigma_o^{-1} H^* \Sigma_o^{-1} + \Sigma_o^{-1} H^* (\Sigma_o^* + H^*)^{-1} H^* \Sigma_o^{-1}
\]

\[
= \Sigma_o^{-1} - \Sigma_o^{-1} (H^* - H^* K^* H^*) \Sigma_o^{-1}
\]

Define \(\Delta^n := \Sigma_o^n - \Sigma_o^o\). By simple manipulation, we can have

\[
\Sigma_o^n = \Sigma_o^o - \Delta^n = \Sigma_o^o K^* \Sigma_o^o + H^* - H^* K^* H^* - \Delta^n
\]

\[
= \Sigma_o^o K^* \Sigma_o^o + H^* + (\Delta^n K^* \Sigma_o^n + \Sigma_o^o K^* \Delta^n + \Delta^n K^* \Delta^n - \Delta^n) - H^* K^* H^*
\]

where \(K^*\) is sparse, the support of \(H^*\) is known. If \(\|H\|_2\) is relatively small (which may occur when \(|\Omega|\) is small compared with the ambient dimension \(p^2\)), then \(H^* K^* H^*\) can be treated as noise with small magnitude. In addition, when the sample complexity \(n\) is large enough, \(\Delta^n\) can be reasonably small, which can again serve as noise. Treating \(W\) as an additive noise, this problem can be posed as how to find a sparse representation \((K^*\) and \(H^*)\) [DH01] to approximate \(\Sigma_o^n\). Based on it, we propose the following convex program:

\[
(CVX) \text{ minimize } f(K, H) := \lambda_n \|K\|_1 + \frac{1}{2} \|\Sigma_o^o K \Sigma_o^o + H - \Sigma_o^n\|_F^2
\]

subject to \(\text{ supp}(H) \subseteq \Omega^c\) \(\quad (2)\)

where \(\lambda_n\) is the regularization parameter. The \(\ell_1\) regularized term is applied here in order to encourage sparsity of \(K\). In this project, we will attempt to analyze the theoretical performance of this algorithm, namely, under what conditions will we be able to recover the exact support of \(K^*\).

\section{Theoretical Guarantee}

\subsection{Main Result}

This paper focuses on the theoretical aspect of our theorem. Specifically, we derive some sufficient conditions under which the algorithm succeeds in finding the true support. Encouragingly, most of the conditions are mild ones which can oftentimes be naturally satisfied.

The following theorem gives a formal statement of our sufficient conditions that ensure consistency of the recovered support.

\textbf{Lemma 1.} Suppose that the following assumptions hold

(1) The eigenvalues of \(\Sigma_o^o\) satisfies \(\lambda_{\min} \leq \lambda (\Sigma_o^o) \leq \lambda_{\max}\);

(2) The operator \(P_{\Omega^c} O_{\Sigma_o} P_S\) obeys the following restricted isometry property: there exists some \(\delta \geq 0\) independent of \(p\) such that \(\forall X\) supported on \(S\), we have

\[
\|P_S O_{\Sigma_o} P_{\Omega^c} O_{\Sigma_o} P_S (X)\|_F \geq \frac{1}{1 + \delta} \frac{\sqrt{k_{\Omega^c} k_S} \Lambda_{\min}}{p^2} \|X\|_F
\]
and
\[ \| P_{S \cap O} P_{\Omega} O S_0 P_S (X) \|_\infty \leq \frac{1}{3 \theta_{\min}} \| X \|_F \]

(3) \[ |K_{ij}^*| \geq \theta_{\min} \text{ if } (i, j) \in S; \]

(4) \[ \| H^* W^* H^* \|_F \leq \max \left\{ \frac{1}{3 \Lambda_{\max}}, \frac{1}{3 \Lambda_{\max} (1+\delta)} \right\} \].

In addition, if we have
\[ \lambda_n < \frac{1}{3 \theta_{\min}} \frac{\Lambda_{\min} \sqrt{k_{\Omega}}}{(1+\delta)p^2} \]
then for sufficiently large sample size \( n \), the solution of the convex program (2) recovers the exact support of \( K^* \).

**Explanations.** We provide some explanation of these conditions as follows.

(a) The first condition guarantees that \( \Sigma_0^* \) is non-degenerate and hence invertible.

(b) The second condition makes sure that this operator preserves signal energy when restricted to a small class of signals. In other words, this characterizes the *incoherence* between \( \Sigma_0 \) and \( S \).

(c) The third condition ensures that each entry has sufficient strength, otherwise it would be buried by noise or residual term.

(d) The fourth condition guarantees that the residual term has small magnitude and can hence be treated as noise.

The proof is deferred to supplemental materials [Che11] due to space limit, with a roadmap of proof given in the following subsection.

### 4.2 Roadmap of Theorem Proof

#### 4.2.1 Optimality and Uniqueness Condition

In order to examine whether a given matrix \( \hat{K} \) is the optimizer, one sufficient and necessary condition is to see whether the subdifferential of the objective function \( f(K, H) \) contains 0. However, this does not guarantee that \( K \) is the unique optimizer, and hence the optimization program may end up with a solution with a wrong support. In order to guarantee uniqueness, we need the following restricted isometry property, as stated in the following lemma.

**Lemma 2.** Suppose that there exists a primal-dual pair \( (\hat{K}, \hat{Z}) \) that satisfies

1. \( \hat{Z}_{ij} = \text{sign} \left( \hat{K}_{ij} \right) \) if \( (i, j) \in \Omega \) and \( |\hat{Z}_{ij}| < 1 \) otherwise;

2. \( \lambda_n \hat{Z} + \Sigma_0^n P_{\Omega} \left( \Sigma_0^n (\hat{K} - K^*) \Sigma_0^n + W \right) \Sigma_0^n = 0. \)

In addition, suppose that \( P_{\Omega} O \Sigma_0 \hat{P}_S \) obeys the following restricted isometry property: there exists some \( \epsilon > 0 \) such that
\[ \| P_{\Omega} O \Sigma_0 \hat{P}_S (X) \|_F \geq \epsilon \| X \|_F \quad \forall X = \hat{P}_S (X), \]
where \( \hat{S} \) denotes the support set of \( \hat{K} \). Then \( \hat{K} \) is the **unique** optimal solution of the convex program (2).

**Proof.** See supplemental materials [Che11].
4.2.2 Construction of A Primal-Dual Pair

Now, we define the following candidate solution \( \hat{K} \) when restricted to the matrices lying in the linear subspace supported on \( \operatorname{supp}(K^*) \). More specifically, we set \( \hat{K} \) to be the solution of the following program

\begin{equation}
\begin{aligned}
\text{minimize} \quad & f(K, H) = \lambda_n \|K\|_1 + \frac{1}{2} \|\Sigma_o^n K \Sigma_o^n + H - \Sigma_o^n\|_F^2 \\
\text{subject to} \quad & \operatorname{supp}(K) \subseteq S, \quad \operatorname{supp}(H) \subseteq \Omega^c
\end{aligned}
\end{equation}

It can be shown that the subgradient of \( f(K, H) \) with respect to \( K \) is

\[ \lambda_n \partial \|K\|_1 + \Sigma_o^n (\Sigma_o^n K \Sigma_o^n + H - \Sigma_o^n) \Sigma_o^n. \]

For any given \( K \), the minimizing \( H \) is given by

\[ H = \mathcal{P}_{\Omega^c}(-\Sigma_o^n K \Sigma_o^n + \Sigma_o^n), \]

which allows the subgradient with respect to \( K \) on this optimizing \( H \) to be expressed as

\[ \lambda_n \partial \|K\|_1 + \Sigma_o^n (\Sigma_o^n K \Sigma_o^n + H - \Sigma_o^n) \Sigma_o^n \]

by observing that \( \Sigma_0^n = \Sigma_o^n K^* \Sigma_o^n + H^* + W \) and that \( \mathcal{P}_{\Omega}(H^*) = 0 \).

When restricted to the support set \( S \), the optimality condition can be written as

\[ \lambda_n \operatorname{sign}(K^*) - \mathcal{P}_S \mathcal{O}_{\Sigma_o} \mathcal{P}_{\Omega}(W) = -\mathcal{P}_S \mathcal{O}_{\Sigma_o} \mathcal{P}_{\Omega} \mathcal{O}_{\Sigma_o} \mathcal{P}_S \left( \hat{K} - K^* \right) \]

Suppose that \( \mathcal{P}_S \mathcal{P}_{\Sigma_o} \mathcal{P}_{\Omega} \mathcal{P}_{\Sigma_o} \mathcal{P}_S \) is injective when restricted to the set of matrices supported on \( S \). Then we can obtain

\[ \hat{K} - K^* = -\left( \mathcal{P}_S \mathcal{O}_{\Sigma_o} \mathcal{P}_{\Omega} \mathcal{O}_{\Sigma_o} \mathcal{P}_S \right)^{-1} \left( \lambda_n \operatorname{sign}(K^*) - \mathcal{P}_S \mathcal{O}_{\Sigma_o} \mathcal{P}_{\Omega}(W) \right) \]

If we want the restricted solution \( \hat{K} \) to be the true optimal, then we need \( \hat{K} \) to yield small residual outside the support \( S \), namely

\[ \|\mathcal{P}_S^c \mathcal{O}_{\Sigma_o} \mathcal{P}_{\Omega} (\Sigma_o (K - K^*) \Sigma_o - W)\|_\infty < 1 \]

When considering the magnitude of the residual term \( W \), we may need to control the size of the estimation error \( \Delta \) between the empirical covariance and the true covariance. This can be bounded through the following lemma.

**Lemma 3.** Let \( \lambda(\Sigma_0^n) \leq \Lambda_{\max} \). If \( n \geq \max \left\{ \frac{64lp\|\Sigma_1\|_2^2}{\delta^2}, 50l \log p \right\} \), then we have

\[ \operatorname{Prob}(\|\Delta\|_\infty > \delta) \leq 2 \exp \left( -\frac{n\delta^2}{100l \Lambda_{\max}^2} \right) \]

Our proof follows by combining the above facts. Interested readers are referred to supplemental materials [Che11] for details.
5 Numerical Example

In order to justify the correctness of algorithm, we provide a numerical example as follows. Suppose $K$ is a $30 \times 30$ matrix. We generate $K$ by random Jacobi rotation of a positive definite diagonal matrix. Although all conditions on $K$ can be easily justified in Matlab, we do not test it in order to see the generality and applicability of our algorithm. The regularization parameter $\lambda_n$ is set to 1, and $n$ is chosen to be $\max \left\{ \frac{64l}{p} \frac{\| \Sigma^* \|_2^2}{\delta^2}, 50 \log p \right\}$. All missing data patterns we consider here are pairwise patterns, i.e. each pattern contain only 2 entries. We define the density $\rho$ as $l/p^2$, where $l$ is the number of different patterns. This gives us in Fig. 1.

The blue curve illustrates the success probability in recovering the true support using our algorithm. When the sampled density $\rho$ is small, the success probability is almost 0. This begins to grow after we sample above half the pairs, and keeps increasing with the sampled density. When we have more than 90% pairs observed, the algorithm will find a consistent support with very high probability. Since we do not test whether the randomly generated matrix satisfies our sufficient conditions, it is very likely that it does not satisfy all the conditions, especially Condition (3). If Condition (3) is violated, the algorithm may end up finding a support which is strictly contained in the true support. The red curve illustrates exactly the probability of finding a support that is a subset of the true support. In this case, we further use the green curve to plot of the average portion of true positions being recovered. As can be seen, the algorithm always have a very high probability to find a support inside the true one, but the correct portion may grow with the sampled density $\rho$.

6 Concluding Remarks

This paper illustrates how to use convex relaxation approach to attack this non-convex structure learning problem. This also poses a more general question on inverse matrix completion problem, namely, given that a matrix inverse is sparse, how to recover it when we observe partial entries of the original matrix. Besides, using empirical pairwise covariance to estimate the covariance matrix is typically inefficient in sample complexity. It would be interesting to see how to apply the maximum likelihood function here to improve sample complexity.
References


