High-Dimensional Statistics: A Non-Asymptotic Viewpoint

Chapter 11: High dimensional graphical models

Notes by Renxiong Liu

In this chapter, we discuss various types of problems in high-dimensional statistics that arise in the context of graphical models.

11.1 Introduction

In this section, we mainly focus on the statistical inference on the undirected graphical models, or the Markov random fields. Denote by $G = (V, E)$ the undirect graph, where $V = \{1, 2, \dots, d\}$ denotes the set of nodes and $E = \{(i, j)\}$ denotes the edges of graph G. To link the graph with probalistic distribution, we assign each node with a random variable X_i, $i \in V$ and denote by $X = (X_1, \dots, X_d)$ the random vector represented by this graph.

Given such setting, our primary interests are connections between the structure of joint distributions, and the structure of the underlying graph. Two ways are considered in this section: one is based on the factorization and the other comes from conditional independence properties.

11.1.1 Factorization

We first introduce some basic concepts in graphical model. A clique C is a subset of vertices that are all joined by edges, i.e., $(i, j) \in E$ for $i, j \in C$. A maximal clique is a clique that is not a subset of any other cliques. An example is shown in Figure 11.1.

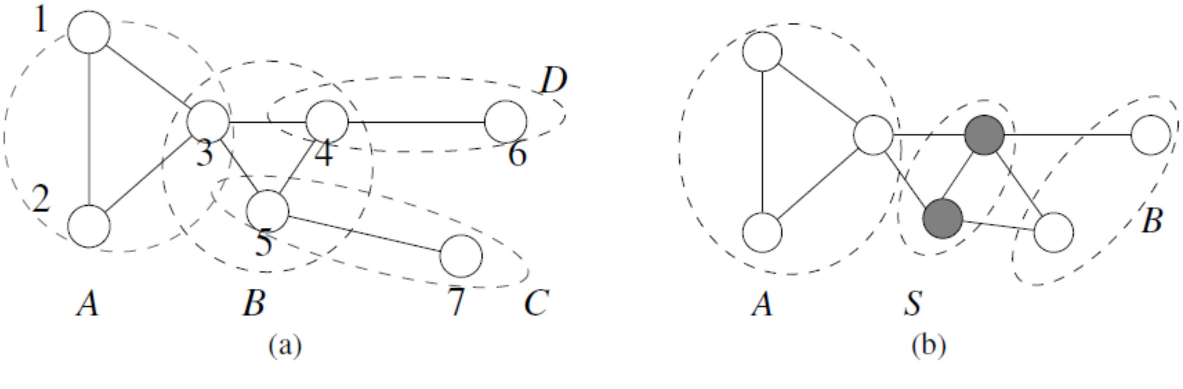

Figure 11.1: Illustrations of clique and maximal clique

Let's denote $\mathcal{C}$ the set of cliques in graph G. For each clique $C \in \mathcal{C}$, we denote by ψ_C a nonnegative function over $X_C = (X_i : i \in C)$, which is also referred as the clique compatibility function in some paper. We have the following definition for factorization.

Definition 11.1. We say random vector $(X_1, \ldots, X_d)$ factorizes by the graph G, if the joint density function p can be represented by

$$p(x_1, \ldots, x_d) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C) \tag{11.1}$$

for some collection of clique compatibility functions $\psi_C : \mathcal{X}_C \to \mathbb{R}^+$.

11.1.2 Conditional independence

We now trun into the other method to connect the structure of distributions and structure of graphical model. In this section, most statements are based on the notion of a vertex cutset S , which is a subset of vertices whose removal from the graph will break it into two or more disjoint pieces. Formally speaking, denote $V \setminus S$ the nodes after removing set S, $E(V \setminus S)$ the residual edge set

$$E(V \setminus S) = \{(i,j) \mid (i,j) \in E, \text{for } (i,j) \in V \setminus S\} \tag{11.2}$$

We denote $G(V \setminus S) = (V \setminus S, E(V \setminus S))$ the graph after deleting sets S. If $G(V \setminus S)$ consists of two or more pieces of disconnected components, then we call S the vertex cutset.

Given the definition of vertex cutset, we now define the conditional independence over graph. For any $A \subseteq V$, let X_A be the random vectors associated with vertices A, we have the following definition for conditional independence.

Definition 11.2. A random vector $X = (X_1, \ldots, X_d)$ is Markov with respect to a graph G, if for all vertex cutsets S that breaks the graph into disjoint pieces A and B, the conditional independence statement $X_A \perp X_B \mid X_S$ holds.

11.1.3 Hammersley-Clifford equivalence

Now we have introduced two methods that will be used for graphical model studies. We may wonder the relationship between the two definitions, which is established by the well-known Hammersley-Clifford theorem.

Theorem 11.1 (Hammersley-Clifford). *For a given undirected graph G and any random vector $X = (X_1, \ldots, X_d)$ with a strictly positive density p, the following two properties are equivalent.*

- *The random vector X can be factorzied according to graph G.*
- *The random vector X is Markov with respect to graph G.*

11.2 Estimations of Gaussian graphical model

We first present the standard model that will be considered in this section. Any non-degenerate Gaussian distribution with zero mean can be parameterized in terms of its precision matrix $\Theta^\star = \Sigma^{-1}$:

$$p(x_1, \ldots, x_n; \Theta) = \frac{1}{(\sqrt{2\pi})^d} \sqrt{\det(\Theta)} \cdot \exp(-\frac{1}{2} x^\top \Theta x) \tag{11.3}$$

which can be further factorized into

$$p(x_1, \ldots, x_n; \Theta) = \frac{1}{(\sqrt{2\pi})^d} \sqrt{\det(\Theta)} \cdot \exp(-\frac{1}{2} \sum_{(i,j) \in E} \Theta_{i,j} x_i x_j) \tag{11.4}$$

By Hammersley-Clifford Theorem, we have that $\Theta_{i,j} = 0$ if and only if $(i, j) \neq E$. In the remaining part of this section, we investigate graphical model selection, the goal of which is to recover edge set E of the underlying graph G. More specifically, denote $\hat{E}$ the estimated edges based on $\hat{\Theta}$, one merit we investigate is $\mathbb{P}(\hat{E} \neq E)$, which assesses whether or not we have recovered the true underlying edge set. A similar and related criterion is: given the tolerance parameter $\delta \in (0, 1)$, what is the probability to recover the fraction of $1 - \delta$ edges of underlying graph. Another more widely considered topic is on the inverse covariance matrix estimation itself, i.e. $\|\hat{\Theta} - \Theta^\star\|_2 = \sigma_{\max}(\hat{\Theta} - \Theta^\star)$ or $\|\hat{\Theta} - \Theta^\star\|_F$.

11.2.1 Graphical LASSO

A popular method to estimate the precision matrix $\Theta^\star$ is the well-known graphical LASSO. To proceed, we first introduce some notations that will be used throughout this section. For symmetric matrix A and B, we denote inner product $\langle A, B \rangle = \operatorname{tr}(A^\top B)$. Moreover, we define the following negative log-determinant function over space $\mathbb{S}^{d \times d}$:

$$-\log\det(\Theta) = \begin{cases} -\sum_{k=1}^{d} \sigma_k(\Theta) & \text{if } \Theta \succ 0 \\ \infty & \text{otherwise} \end{cases} \tag{11.5}$$

Now, given sample $\{x_i\}_{i=1}^n$ from distribution $N(0, \Theta^{-1})$, then up to constant, the log-likelihood function can be represented as

$$L_n(\Theta) = \langle \Theta, \hat{\Sigma} \rangle - \log\det\Theta \tag{11.6}$$

where $\hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} x_k x_k^\top$ is the sample covariance matrix.

When the sample covariance matrix is invertible, the unrestricted maximizer can be derived by taking derivative with respect to Θ in (11.6):

$$\hat{\Sigma} - \Theta_{\text{MLE}}^{-1} = 0 \Rightarrow \Theta_{\text{MLE}} = \hat{\Sigma}^{-1} \tag{11.7}$$

However, when $n < p$, $\hat{\Sigma}$ is always rank-deficient, meaning that MLE does not exist. In this scenario, regularization will take effects. Graphical LASSO impose an l_1-constraint on the entries of Θ, which leads to the following problem

$$\hat{\Theta} \in \arg\min_{\Theta \in \mathbb{S}^{d \times d}} \left\{ \langle \Theta, \hat{\Sigma} \rangle - \log\det\Theta + \lambda_n \|\Theta\|_{1,\text{off}} \right\} \tag{11.8}$$

where $\|\Theta\|_{1,\text{off}} = \sum_{i \neq j} |\Theta_{i,j}|$ is the sum of absolute value of the off-diagonal terms in precision matrix Θ. In some sense, l_1-constraint could force the penalized entry to be close to 0. Since the diagonal term of precision matrix is always posotive, adding penalty to them will naturally lead to the bias (heuristically).

We start our discussions by considering the bounds on Frobenius norm $\|\hat{\Theta} - \Theta\|_F$.