A Study on Graphical Models

Debarshi Chakraborty

M.Stat 2nd Year
Indian Statistical Institute, Kolkata

March 10, 2023

Introduction

Graphical Model :A graph in which random variables are represented as nodes. The graph keeps track of the conditional independence relations that exist in the joint distribution of those random variables.

Broadly divided into two kinds :

- Undirected Graphical Models
- Directed Graphical Models

Advantages

- Intuitive way to visualize relationship between random variables.
- The number of parameters to be estimated can be reduced by exploiting conditional independence relations.
- Queries about conditional dependence relations can be answered just by looking at the graph.
- Understanding cause-effect relationships between different variables.
- Computation of conditional distributions becomes easier, especially in high dimensional setup and a sparse graph structure.

Markov Random Fields

- An undirected graph model with set of nodes $V = \{X_1, X_2, ..., X_p\}$ and set of edges $E \subseteq V \times V$, denoted by $G = (V, E)$.
- Edges are represented by ordered pair of nodes (X_i, X_j). Clearly in undirected graphs $(X_i, X_j) \in E \iff (X_j, X_i) \in E$.
- **Adjacency Matrix** : $A = ((a_{ij}))$ is defined by $a_{ij} = \mathbf{1}\{(X_i, X_j) \in E\}$.
- X_i and X_j is said to **separated** by $C \subseteq V$ if all the paths from X_i to X_j intersect C. Let A, B, C be subsets of V. C is said to separate A and B if it separates X_i and X_j for all $X_i \in A$ and $X_j \in B$.
- **Pairwise Markov Property** relative to graph G : For any $(X_i, X_j) \notin E$, we have $X_i \perp X_j | V - \{X_i, X_j\}$.
- A probability distribution P is said to **factorize** according to a graph G if for all complete subsets $A \subseteq V$, there exists non negative functions $\psi_A(.)$ depending on $\mathbf{x}$ only through $\mathbf{x_A}$, such that P has density $f(.)$ of the form $f(\mathbf{x}) \propto \prod_{\mathbf{A} \subseteq \mathbf{V}} \psi_{\mathbf{A}}(\mathbf{x})$.
- The functions $\psi_A(.)$(*potential functions*) are not unique.

Markov Random Fields

Theorem

(Hammersley and Clifford) A probability distribution P with a positive and continuous density f satisfies the pairwise Markov property relative to an undirected graph G if and only if P factorizes according to G.

- The converse of this theorem can also holds. The continuity assumption can be relaxed while positivity is essential.
- It can be shown that knowledge of pairwise conditional independence (pairwise Markov property) is enough to infer about higher order conditional independence relations (namely local and global Markov properties).
- We define $X_i \perp X_j | \, V - \{X_i, X_j\} \iff$ X$_i$ and X_j are separated by $V - \{X_i, X_j\}$ $\iff$ X$_i$ and X_j are not adjacent.
- Goal is to identify which edges are absent in the graph.
- Once the graph is identified, we will be able to write the joint distribution in a factorized form using the above theorem.

Example in Gaussian Setup

- **Model** : $X = (X_1, X_2, ..., X_p)^T \sim N_p(0, \Sigma)$.
- **Data** : $\begin{pmatrix} x_{11} & x_{12} & . & . & .x_{1p} \\ x_{21} & x_{22} & . & . & .x_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ x_{n1} & x_{n2} & . & . & .x_{np} \end{pmatrix}$
- **Alternative Parametrization** : $f(x) \propto exp[-\frac{1}{2}x^T \Theta x]$ where $\Theta = \Sigma^{-1}$.
- Breaking up the exponent gives two kinds of terms: $\theta_{ii}\ x_i^2$ and $\theta_{ij}\ x_i x_j$.
- These are our $\psi_A(.)$ s in this particular case for all complete subsets A of $\{X_1, ..., X_p\}$.
- This is an advantage for the multivariate normal that any p-variate density can be decomposed into linear and quadratic terms of the observations.
- Parameter of interest is the precision matrix.
- Why Θ instead of Σ ?

Example in Gaussian Setup

Theorem

Let $(X_1, ..., X_n) \sim N_p(\mu, \Sigma)$ and let $\Theta = \Sigma^{-1}$. Then $\theta_{ij} = 0$ if and only if X_i and X_j are conditionally independent given the rest.

- So, instead of estimating $\frac{p(p+1)}{2}$ parameters of Σ, our goal is to estimate only the non zero entries of Θ.
- **Two step approach** : Estimate the graph/ adjacency matrix and then estimate Θ subject to the constraint that some particular entries of Θ are zeroes.
- **Step 1** : Use multiple testing. The partial correlation between X_i and X_j is given by $\rho^p_{ij} = -\frac{\theta_{ij}}{\sqrt{\theta_{ii}\theta_{jj}}}$. A natural idea is to test

$$H_0 : \rho^p_{ij} = 0 \text{ against } H_1 : \rho^p_{ij} \neq 0$$

for each of the $\binom{p}{2}$ pairs. We use $r^p_{ij} = -\frac{\hat{\theta}_{ij}}{\sqrt{\hat{\theta}_{ii}\hat{\theta}_{jj}}}$ as our test statistics and as an estimate of Θ we can use S^{-1}. Under H_0

$$t_{ij} = \sqrt{n-p}\frac{r^p_{ij}}{\sqrt{1-(r^p_{ij})^2}} \sim t_{n-p}$$

Example in Gaussian Setup

- **Step 2** : Use a modified regression algorithm for estimation of an undirected GGM with known structure. The constrained log likelihood is given by
$$logL(\Theta) = log|\Theta| - tr(S\Theta) + \textstyle\sum_{(j,k)\notin E} \gamma_{jk}\theta_{jk}$$
- The gradient equation for maximising the log likelihood
$$\Theta^{-1} - S - \Gamma = 0$$
where Γ is the matrix for Lagrange parameters with nonzero values for all pairs with edges absent
- At a particular iteration, denote the estimated covariance matrix by W and the estimated precision matrix by Θ. Partition the two matrices as
$$\begin{bmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{bmatrix} \begin{bmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{bmatrix} = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0^T} & 1 \end{bmatrix}$$
- Some elementary matrix algebra and properties of multivariate normal distribution yield the following :
- $w_{12} = -W_{11}\frac{\theta_{12}}{\theta_{22}} = W_{11}\beta$
- $W_{11}\beta - s_{12} - \gamma_{12} = 0$
- $\frac{1}{\theta_{22}} = w_{22} - w_{12}^T\beta, \; w_{22} = s_{22}$

Example in Gaussian Setup : Iterative Regression Algorithm

1. Initialise $W = S$.
2. Repeat for $j = 1, 2, ..., p,\ j = 1, 2, ..., p, ...$ until convergence
   - Partition W into part 1 : all but the j^{th} row and column and part 2 : j^{th} row and column
   - Solve the reduced system of equations (only those where $\gamma_{12} = 0$)
     $$W^*_{11}\beta* - s^*_{12} = 0$$
     to get $\hat{\beta*}$. Obtain $\hat{\beta}$ by padding zeroes in appropriate positions of $\hat{\beta*}$.
   - Update $w_{12} = W_{11}\hat{\beta}$
3. In the final cycle (for each j), solve for $\hat{\theta_{12}} = -\hat{\beta}\,\hat{\theta_{22}}$ with $\frac{1}{\hat{\theta_{22}}} = w_{22} - w_{12}^T\hat{\beta}$

The second step i.e. the estimation step works well even when $p \geqslant n + 1$, but multiple testing fails. In such scenario, **one step approach** is used which directly estimate only the nonzero entries of Θ. The algorithm is same as the above, just modifying the initialization by $W = S + \lambda I$ and in second part of step 2, a LASSO regression is used instead of usual linear regression. The resulting method is known as **Graphical LASSO**.

Bayesian Networks

- A **directed acyclic** graph whose nodes are random variables.
- A *directed edge* is represented by $X_i \rightarrow X_j$ or the ordered pair (X_i, X_j).
- Note that, unlike MRFs, in DAGs we have $(X_i, X_j) \in E \implies (X_j, X_i) \notin E$, since cycles are not allowed.
- **Adjacency matrix** A=$((a_{ij}))$ of a DAG G with set of nodes $V = \{X_1, X_2, ...X_p\}$ and set of directed edges E is defined by $a_{ij} = 1\{(X_i, X_j) \in E\}$.
- **Separation in DAGs**: A chain π from X_i to X_j in a DAG G is said to be *blocked* by a subset S if it contains a vertex $X_k \in \pi$ such that "$X_k \in S$ and arrows of π do not meet head to head at X_k" OR "$X_k \notin S$ nor has any descendants in S and arrows of π do meet head to head at X_k." *Two subsets A and B are said to be* **d-separated** *by S if all chains from A to B are blocked by S.*
- **Factorization**:Let G be a BN graph over $X_1, ..., X_p$. We say that a distribution P with density $f(.)$ over the same space factorizes according to G if P can be expressed as a product

$$f(x_1, ..., x_n) = \prod_{\alpha \in V} f(x_\alpha | Pa^G_{x_\alpha})$$

Bayesian Networks

Theorem

A positive probability distribution P factorizes with respect to a DAG G iff

$$X_i \perp Nd(X_i) | Pa(X_i)$$

where Nd(A) and Pa(A) respectively denote the set of non descendants and parents of A.

- We define

$$X_i \perp X_j | \; V - \{X_i, X_j\} \iff \text{X}_i \text{ and } X_j \text{ are d-separated by } V - \{X_i, X_j\}$$

- Goal is to estimate the structure of the DAG i.e. which edges are present and which are absent.
- An extra task for the directed case involves estimating the direction of the arrows, since a DAG provides a visual representation of causal relationships among a set of random variables.
- Each arrow in a DAG represents a causal effect, for example $X_i \rightarrow X_j$ means that the variable X_i has a causal effect on the variable X_j.
- Once the graph is estimated, using the above theorem, we will be able to write down the joint density in a factorized form.

Bayesian Networks : A simple example

- Suppose we have n observations from (X_1, X_2, X_3, X_4) where each X_i is a Bernoulli random variable.
- If we know the adjacency matrix to be

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

  then we can write down the joint distribution as $f(x_1, x_2, x_3, x_4) = f(x_4/x_2)\, f(x_3/x_2)\, f(x_2/x_1)\, f(x_1)$.
- Each of the terms in the product is the density of a Bernoulli random variable. Thus, to represent the joint distribution, we have to estimate 4 parameters, whereas if we wanted to write down the whole distribution explicitly, we had to estimate 15 independent parameters. This gain in terms of the reduction in number of parameters become enhanced as dimension increases.

Example in Gaussian Setup : Parameter Estimation

- Let Y be a continuous variable in a DAG with parents $X_1, ..., X_k$. We say that Y has a *Linear Gaussian* model of its parents if there are parameters $\beta_o,\ \beta_1, ...,\ \beta_k$ and σ^2 such that

$$Y|X = x \sim N(\beta_o + \beta_1 x_1 + ... + \beta_k x_k, \sigma^2)$$

- A DAG is called a Gaussian Bayesian network if all the variables are continuous and every variable has a linear gaussian model of its parents.
- There exists a one to one correspondence between a gaussian BN and a multivariate normal distribution.
- **Goal** : Estimate the structure of the underlying DAG and the parameters of the distribution.
- Once the underlying DAG G is estimated, we can use M.L.E to estimate the parameters. Decompose the log likelihood according to the graph structure

$$l(\theta) = logL(\theta) = \sum_{j=1}^{d} log(\prod_{i=1}^{n} p(x_{ij}|\ pa(x_j); \theta_j)) = \sum_{j=1}^{d} logL_j(\theta_j) = \sum_{j=1}^{d} l_j(\theta_j)$$

- Maximize the contribution to the log-likelihood of each node independently.

Example in Gaussian Setup : Structure Learning

- Identifying the exact DAG G is not possible, only the Markov equivalence class of G can be identified.

Theorem

Two DAGs G1 and G2 are Markov equivalent if and only if (i) skeleton(G1) = skeleton(G2) and (ii) G1 and G2 have the same unshielded colliders (i.e. any two nodes pointing to the same collider are not connected).

- Due to this theorem, we can get hold of the Markov equivalence class of G if we can identify the set of undirected edges and unshielded colliders.
- **PC Algorithm** : Operates in 2 steps
  1. Identifying the skeleton (invloves testing for conditional independence).
  2. Identifying complete partially directed acyclic graph.
- Only point where gaussian assumption helps is the testing part, otherwise a general method for structure learning in Bayesian Networks.
- Thus, for studying non gaussian models, main challenge is to test for conditional independence.

Nonparametric Graphical Model : The Nonparanormal

- This approach is a semiparametric extension of the multivariate normal, allowing arbitrary graph structure.
- We say a random vector $X{=}(X_1, ..., X_p)^T$ has a nonparanormal distribution an write $X \sim NPN_p(\mu, \Sigma, f)$, in case there exists functions $\{f_j\}_{j=1}^{p}$ such that $(f_1(X_1), ..., f_p(X_p))^T \sim N_p(\mu, \Sigma)$.
- When the functions $\{f_j\}_{j=1}^{p}$ are monotone and differentiable, the joint pdf of X is given by
$$p(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(f(x)-\mu)^T\Sigma^{-1}(f(x)-\mu)] \prod_{j=1}^{p} |f_j^{'}(x_j)|$$
- To make the family identifiable we demand that the functions $\{f_j\}_{j=1}^{p}$ preserve the means and variances
$$\mu_j = E(X_j) = E(f_j(X_j)) \text{ and } \sigma_j^2 = V(X_j) = V(f_j(X_j))$$
- We choose $\{f_j\}_{j=1}^{p}$ such that they follow univariate normal. If $F_j(x)$ denote the marginal CDF of X_j then
$$f_j(x) = \mu_j + \sigma_j\Phi^{-1}(F_j(x))$$

Nonparametric Graphical Model : The Nonparanormal

- It can be shown that the conditional independence structure of remains invariant $(X_1, ..., X_p)^T$ under such transformation.
- A two step procedure is followed to estimate the graph :
  1. The functions $\{f_j\}_{j=1}^p$ and the parameters μ_j, σ_j are estimated from the data and the observations for each variable are replaced by their respective normal scores.
  2. Apply graphical lasso to the transformed data to estimate the undirected graph.
- **Limitation :** Although normality of the marginal distributions of $\{f_j(X_j)\}_{j=1}^p$ is justified, it is *assumed* that $(f_1(X_1), ..., f_p(X_p))^T$ follows multivariate normal. The validity of this assumption is not justified, thus the procedure remains incomplete.

Nonparametric Graphical Model : Forest Density Estimation

- In this approach, arbitrary nonparametric distributions are allowed the graph structure is restricted to a tree or forest.
- Let $p^*(x)$ be a probability density function on $\mathcal{R}^d$ an let $X_1, ..., X_n$ be i.i.d observations from $p^*(.)$.
- If F is a d-node undirected forest with set of nodes $V_F = \{X_1, ..., X_d\}$ and set of edges $E_F \subset \{X_1, ..., X_d\} \times \{X_1, ..., X_d\}$, the number of edges satisfies $|E_F| \leqslant d - 1$. We say that a probability density function $p(x)$ is supported by a forest F if the density can be written as
$$p_F(x) = \prod_{(x_i,x_j)\in E_F} \frac{p(x_i,x_j)}{p(x_i)p(x_j)} \prod_{x_k \in V_F} p(x_k)$$
- Let $\mathcal{F}_d$ be the family of forests with d nodes, and let $\mathcal{P}_d$ be the corresponding family of densities
$$q^* = argmin_{q\in\mathcal{P}_d} D(p^*, q)$$
where $D(p, q)$ is the KL divergence.

Theorem

There exists $F^* \in \mathcal{F}_d$ such that
$$q^* = p_{F*} = \prod_{(x_i,x_j)\in E_{F*}} \frac{p*(x_i,x_j)}{p*(x_i)p*(x_j)} \prod_{x_k \in V_{F*}} p^*(x_k)$$

Nonparametric Graphical Model : Forest Density Estimation

- Maximise $\sum_{(X_i,X_j)\in E_{F*}} I(X_i, X_j)$ where

$$I(X_i, X_j) = \int p^*(x_i, x_j) log(\frac{p*(x_i,x_j)}{p*(x_i)p*(x_j)}) dx_i dx_j$$

- $p^*(.)$ is unknown. Use kernel density estimates to get an estimate of each $I(X_i, X_j)$.
- **Two step method** : At first, we divide our data into two sets randomly (say D_1 and D_2 of sizes n_1 and n_2 respectively)

  1. Using D_1, compute kernel density estimates of the univariate and bivariate marginals and calculate $\hat{I}_{n_1}(X_i, X_j), i \neq j$. Construct a full tree $\hat{F}_{n_1}^{(d-1)}$ with $d-1$ edges using Chow-Liu algorithm.
  2. Using D_2, prune the tree $\hat{F}_{n_1}^{(d-1)}$ to find a forest $\hat{F}_{n_1}^{(\hat{k})}$ with $\hat{k}$ edges.

- Once we obtain $\hat{F}_{n_1}^{(\hat{k})}$ in step 2, we can calculate $\hat{p}_{\hat{F}_{n_1}^{(\hat{k})}}$ according to the proposition mentioned above using the kernel density estimates obtained in step 1.
- **Limitation :** Cycles are not allowed, which is one of the advantages of studying Markov Random Fields over Bayesian Networks.