A Study on Graphical Models

Debarshi Chakraborty

MD2105

M.Stat 2nd Year
Indian Statistical Institute, Kolkata

May 26, 2023

Recap of previous talk

In the session of March 13, 2023 we discussed

- Definition of Graphical Models
- Advantages of using Graphical Models
- Types of graphical models : Directed and Undirected
- Semantics of graphical model for both Bayesian Networks and Markov Random Fields
- Setup and estimation method of Gaussian Graphical Models
- Issues in estimation for high dimensional scenario

Recap of previous talk

- Gaussian Markov Random Field : Graphical LASSO
- Gaussian Bayesian Network : PC Algorithm
- Two nonparametric methods for continuous data : Nonparanormal and Forest Density Estimation
- Comparison between Graphical LASSO and Nonparanormal for non gaussian data (artificially created)

Overview of today's discussion

- A multiple testing method for graph structure estimation for continuous case
- Comparison with Nonparanormal (using artificial data)
- Structure estimation for discrete graphical models , both large sample and high dimensional setup
- A different perspective for error control while estimating graphical models
- Applications on dataset

Measure of Independence : Distance Correlation

- Let X and Y be two random variables
- We define the distance covariance ν between X and Y using the weighted L_2 norm between their joint characteristic function and product of individual characteristic functions
- $\nu^2(X,Y) = ||\phi_{X,Y}(t,s) - \phi_X(t)\phi_Y(s)||_w^2$ with appropriately chosen weight functions
- The distance correlation $R(X,Y)$ between X and Y is hence given by

$$R^2(X,Y) = \frac{\nu^2(X,Y)}{\sqrt{\nu^2(X,X)\nu^2(Y,Y)}} I(\nu^2(X,X)\nu^2(Y,Y) > 0)$$

Theorem

If $E(|X| + |Y|) < \infty$ then $0 \leqslant R(X,Y) \leqslant 1$ and $R(X,Y) = 0 \iff X \perp\!\!\!\perp Y$

Measure of Independence : Distance Correlation

- If we have a sample of size n then the sample version of distance covariance is given by

$$\nu_n^2(X, Y) = S_1 + S_2 - 2S_3$$

where

$S_1 = \frac{1}{n^2} \sum_{k,l=1}^{n} |X_k - X_l||Y_k - Y_l|$
$\mathrm{S}_2 = \frac{1}{n^2} \sum_{k,l=1}^{n} |X_k - X_l| \frac{1}{n^2} \sum_{k,l=1}^{n} |Y_k - Y_l|$
$\mathrm{S}_3 = \frac{1}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} |X_k - X_l||Y_k - Y_m|$

Theorem

Let $T(X, Y, \alpha, n)$ denote the test that rejects the hypothesis of independence when $\frac{n\nu_n^2(X,Y)}{S_2} > (\Phi^{-1}(1 - \alpha/2))^2$ and let α_n denote the level of significance of the test. Then if $E(|X| + |Y|) < \infty)$ we have $\lim_{n\to\infty} \alpha_n = \alpha \, \forall \, 0 \leqslant \alpha \leqslant 0.215$

We will use the above a asymptotic result for hypothesis testing

Using Distance Correlation to test for conditional independence

- Run LASSO regression on the two linear models
$$X_i^{(k)} = \beta_{1,ij}^T Z^{(k)} + \epsilon_i^{(k)} \text{ and } X_j^{(k)} = \beta_{2,ij}^T Z^{(k)} + \epsilon_j^{(k)} \ \forall \ k = 1, 2, ..., n$$
where $Z^{(k)} = X_{(-i,-j)}^{(k)}$
- Estimate the error vectors as $\hat{\epsilon}_i = X_i - Z\hat{\beta}_{1,ij}$ and $\hat{\epsilon}_j = X_j - Z\hat{\beta}_{2,ij}$
- Calculate the empirical distance covariance between $\hat{\epsilon}_i$ and $\hat{\epsilon}_j$ as
$$\nu_n^2(X_i, X_j) = S_1(X_i, X_j) + S_2(X_i, X_j) - 2S_3(X_i, X_j)$$
- The test statistic is given by $T_{ij} = n\nu_n^2(X_i, X_j)/S_2(X_i, X_j)$
- We reject the null hypothesis if $T_{ij} > (\Phi^{-1}(1 - \alpha/2))^2$
- Repeat the above steps for all $\{(i, j) : 1 \leqslant i \leqslant j \leqslant d\}$
- To achieve FDR control use an appropriate set of cutoffs such as Bonferroni's method, Benjamini Hochberg method, Holm's method, etc
- Basically we are testing for $H_0 : X_i \perp X_j | \, \mathcal{L}(Z)$ where $\mathcal{L}(Z)$ is the linear space spanned by all the variables except X_i and X_j

Advantages and Limitations

Advantages

- Individual testing method is mathematically stronger then nonparanormal
- No restriction on graph structure
- Error control can be achieved

Limitations

Cannot be used with small sample size

Graphical Models for Discrete Data

- Here we consider all variables to be categorical, in other works, we study multi way contingency tables
- For simplicity, let us start with 3 binary RVs (X_1, X_2, X_3)
- Joint density can be written as
$$P(x_1, x_2, x_3) = p(0,0,0)^{(1-x_1)(1-x_2)(1-x_3)} ......p(1,1,1)^{x_1 x_2 x_3}$$
- The log linear expansion for this $2 \times 2 \times 2$ contingency table is
$$logP(x_1, x_2, x_3)$$
$$=\mathrm{u}_0 + u_1 x_1 + u_2 x_2 + u_3 x_3 + u_{12} x_1 x_2 + u_{13} x_1 x_3 + u_{23} x_2 x_3 + u_{123} x_1 x_2 x_3$$
- $X_2 \perp\!\!\!\perp X_3 | X_1 \iff u_{23} = 0$ and $u_{123} = 0$
- This idea can be generalised for contingency tables for non binary data i.e. we can denote conditional independence by setting some "u"terms to be 0

Log Linear Model and Conditional Independence

- Let us have k random variables $X_1, X_2, ..., X_k$ with the i^{th} random variable taking d_i values numbered $0, 1, 2, ..., d_i - 1$. the idea is to make the "u"terms functions of x rather than constants
- The *log linear expansion* of the cross classified multinomial distribution P_K is
$$\log P_K(x) = \sum_{a \subseteq K} u_a(x_a)$$
where the sum is taken over all possible subsets a of $1, 2, ..., k$ and where the u terms satisfy the constraints that $u_a(x_a) = 0$ whenever $x_i = 0$ and $i \in a$.

Theorem

If $(X_a, X_b, X_c)^T$ be a partition of the random vector $(X_1, X_2, ...X_k)^T$ then $X_b \perp\!\!\!\perp X_c | X_a$ if and only if all u terms containing co-ordinates from both b and c in the log linear expansion are equal to 0.

- Note that, we will restrict ourselves to models where "presence/absence of all lower order interaction terms " $\iff$ "presence/absence of the corresponding higher order interaction term"

Parameter Estimation and Structure Learning

- Maximum Likelihood Estimation is used, in many cases it is possible to derive the exact form of the MLE analytically
- We may not get closed form estimates always, or even if we get, it might be very inconvenient to write down the log likelihood
- A method called **Iterative Proportional Fitting** is used
- If no closed form estimates exist, the algorithm will converge to the MLE and if indeed closed form estimates exist, it will converge to that estimate in one iteration
- For structure learning, the basic idea is finding the best log linear model and find which "u"terms are absent there , hence remove those edges accordingly from the graph

Structure Learning

- To start with we assume $n >> p$ since we will be using an asymptotic result
- Let L_0 denote the log likelihood of the saturated model i.e.

$$L_0 = \sum_x n(x) \log \frac{n(x)}{N}$$

- Let L_M denote the log likelihood of any other model i.e.

$$L_M = \sum_x n(x) \log \hat{P}_M(x)$$

  The deviance between the model we want to test and the saturated model is given by

$$2(L_0 - L_1) = 2 \sum_x n(x) \log \frac{n(x)/N}{\hat{P}_M(x)}$$

- Let G denote our graphical model. M_0 and M_1 be two models such that $M_0 \subseteq M_1$ i.e. we can obtain M_0 from M_1 by setting additional constraints to 0 or equivalently removing some edges from M_1
- Test for

$$H_0 : G = M_0 \text{ against } H_1 : G = M_1$$

Structure Learning

- We use the test statistic

$$T = \text{deviance}(M_0)\text{-deviance}(M_1) = 2(L_1 - L_0)$$

- Under H_0, as $N \to \infty$, we have $T_N \sim \chi^2_{(m)}$ where m is equal to the number of additional restrictions in M_0 compared to M_1
- To choose between two models where one is not the subset of the other, the Akaike's Information Criterion (AIC) ca be used which is defined as

$$\text{AIC}(M) = \text{deviance}(M) + 2\text{dim}(M)$$

  where $\dim(M)$ is the number of parameters in the model M
- Only limitation is this method cannot be used when $n < p$
- In high dimensional case, we adopt a method which is similar to the graphical lasso algorithm

Structure Learning in High Dimensional Setup

- Again for the sake of simplicity, let us start with binary random variables
- Let us have n i.i.d observations from $(X_1, X_2, ..., X_p)^T$
- It is assumed the joint distribution of $X = (X_1, X_2, ..., X_p)^T$ can be written as
$$f(X_1, X_2, ..., X_p) \propto \exp(\textstyle\sum_i \theta_{ii} X_i + \sum_{i,j} \theta_{ij} X_i X_j)$$
where $\Theta = ((\theta_{ij}))$ is the $p \times p$ matrix specifying the conditional independence structure between the random variables
- $\theta_{ij} = 0 \iff X_i \perp X_j | V - \{X_i, X_j\}$
- The justification behind considering only the pairwise interaction effects is pointed out by *Ravikumar et al. (2010)*
- One difficulty in estimating Θ is that the constant term of the density is computationally intractable
- A strategy to overcome this difficulty is to use the pseudo-likelihood function to approximate the joint likelihood function associated with density

Structure Learning in High Dimensional Setup

- Let x_{ij} denote the i^{th} observation corresponding to X_j
- The pseudo likelihood is given by
$$\prod_{i=1}^{n} \prod_{j=1}^{p} \pi_{ij}^{x_{ij}} (1 - \pi_{ij})^{1-x_{ij}}$$
where $\pi_{ij} = P(x_{ij} = 1 | x_{ik}, k \neq j; \theta_{jk} \forall 1 \leqslant k \leqslant p) = \frac{exp(\theta_{jj} + \sum_{k \neq j} \theta_{jk} x_{ik})}{1 + exp(\theta_{jj} + \sum_{j \neq k} \theta_{jk} x_{ik})}$
- This gives rise to a logistic regression problem where the j^{th} variable is taken as the response and is regressed on the remaining variables, and hence decomposes the problem into p separate logistic regressions, which are simple to solve
- l_1 penalty can be used to achieve sparsity
- *Ravikumar et al(2010)* solved the following optimization problem $\forall j$
$$max_{\{\theta_{jk}\}_{k=1}^{p}} \sum_{i=1}^{n} [x_{ij}(\theta_{jj} + \sum_{k \neq j} \theta_{jk} x_{ik}) - log(1 + exp(\theta_{jj} + \sum_{j \neq k} \theta_{jk} x_{ik}))] - \lambda_j \sum_{j \neq k} |\theta_{jk}|$$
- Often $\theta_{jl} \neq \theta_{lj}$, thus their minimum or maximum is taken to aggregate the results obtained
- The above method is known as Neighbourhood Selection

Structure Learning in High Dimensional Setup

- An alternative is to solve the following optimization problem
- Maximise $\sum_{j=1}^{p}\sum_{i=1}^{n}[x_{ij}(\theta_{jj}+\sum_{k\neq j}\theta_{jk}x_{ik})-log(1+exp(\theta_{jj}+\sum_{j\neq k}\theta_{jk}x_{ik}))]-\lambda\sum_{j<l}|\theta_{jl}|$ subject to $\theta_{jl}=\theta_{lj}$
- An iterative algorithm is used to solve the above
- If the variables which we consider as response in the logistic regressions have more than two categories, the method remains same except that ordinary logistic regression is replaced by a penalised multi class logistic regression

Error control in Graphical Model

- Usually in Statistics, the hypothesis testing problem is set up in such a way that the type 1 error is considered to be more serious than the type 2 error
- This is the reason why we try to control the type 1 error at a particular desired level and then achieve as much power as possible
- Consequently in multivariate setup we are interested to control the family wise error rate (FWER) or the false discovery rate (FDR)
- Now let us ask ourselves which type of error is more serious while we are estimating the structure of a graph
- To answer this, let us analyse the consequences of the two types of errors that may occur

Which error is more serious in Graphical Models?

- To estimate the structure of a graph we test hypotheses of the form
$$H_0 : X_i \perp\!\!\!\perp X_j \,|\, X_{-(i,j)} \text{ vs } H_1 : \text{not } H_0$$
- One of the objectives of graphical models was to write down a joint density in a factorized form which will further facilitate reduction in number of estimable parameters and ease of computing conditional probabilities
- For instance consider 3 node in the graph X, Y, Z
- **Type 1 error** : H_0 is true but we reject it i.e. we are “missing out the information”that $f(x, y|\, z) = f(x|\, z)f(y|\, z)$, thus estimation of density will be a bit difficult
- **Type 2 error** : H_0 is false but we accept it i.e. even when the conditional independence between X and Y given Z does not hold we write $f(x, y|\, z) = f(x|\, z)f(y|\, z)$, which is completely “incorrect”. This will lead us to wrong conclusions

Which error is more serious in Graphical Models?

- Hence, our primary objective should be controlling the number of cases where H_0 is actually false but we accept it
- Here we can consider “Accepting H_0 given H_0 is true”to be a “discovery”, since we want to get hold of as many conditional independence relations as possible

$\downarrow Null Decision \rightarrow$	Accept	Reject	Total
True	U	V	n_0
False	T	S	n_1
Total	A	R	n

- Clearly, here we should try to control $P(T \geqslant 1)$= GFWER (say)
- $P(accept\ H_0/H_0\ true)$= GPOWER (say) can be considered as the counterpart of Power in this scenario, since we want to “discover”as many independence relations as possible

Multiple Testing Method for controlling GFWER

- There are two ways to tackle the above problem
- We can swap H_0 and H_1 and then go for FWER/FDR control, but finding the distribution of the test statistics under the hypothesis of dependence is difficult
- So we can do each individual test at a comparatively high level of significance and then aggregate the results
- The cutoffs for each test should be chosen in such a way that $P(T \geqslant 1) \leqslant \alpha$ or $E(\frac{T}{A} I(A > 0)) \leqslant \alpha$

Theorem

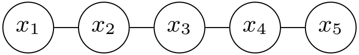
In the above setup,if we use a step up method with cut offs $\alpha_i = 1 - \frac{\alpha}{n}$ *or* $\alpha_i = 1 - \frac{\alpha}{i}\ \forall i$ *then* $P(T \geqslant 1) \leqslant \alpha$

- Note that, the cutoffs are quite similar with step down method of Bonferroni and Holm for FWER control, which may become very conservative (not able to accept H_0) very often, thus improvements based on this idea is possible

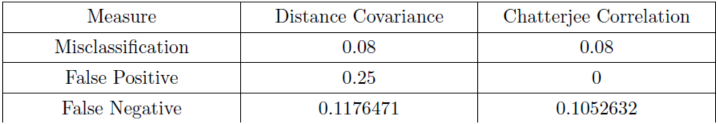
Limitation and a possible alternative

- It may happen that we need to do an individual test at a level higher than 0.215, in that case the asymptotic test using distance correlation won't work
- Alternative is to test using *Chatterjee Correlation Coefficient (CCC)*, which has one of the simplest limit laws among all existing measures of independence
- It is often more powerful than other competing independence measures in certain scenarios, especially when one variable is not a smooth function of the other
- The above is just a hypothesis, mathematics needs to be worked out
- However , we have some small amount of evidence why it might work
- We artificially create the same data which has been used in the paper of estimation using projected distance covariance method

Limitation and a possible alternative



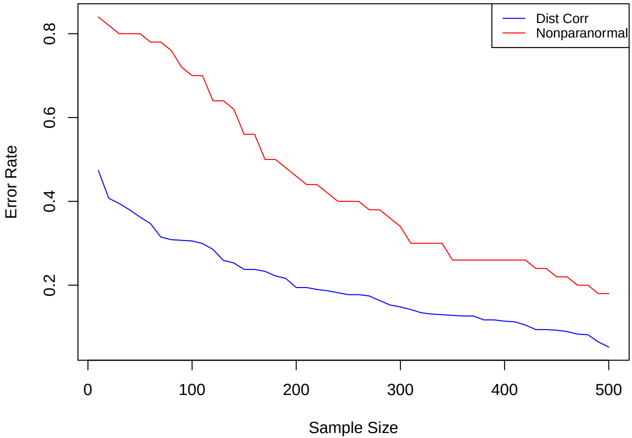
- Random sample is drawn from the above graph structure
- Use Multiple testing using Distance Covariance and Chatterjee Correlation (asymptotic tests)
- The results are as follows



Measure	Distance Covariance	Chatterjee Correlation
Misclassification	0.08	0.08
False Positive	0.25	0
False Negative	0.1176471	0.1052632

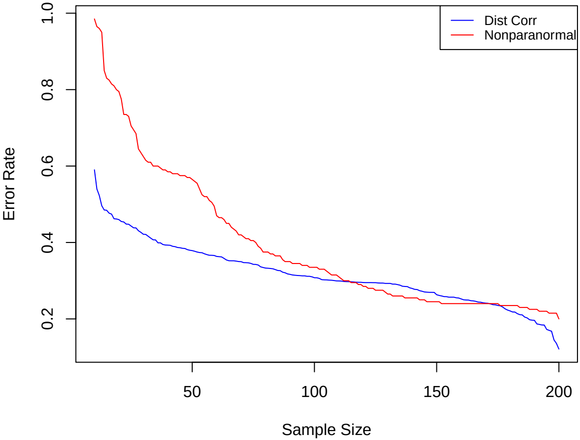
- At least in this example it performs as good as the distance covariance testing

Nonparanormal vs Dist Corr Method



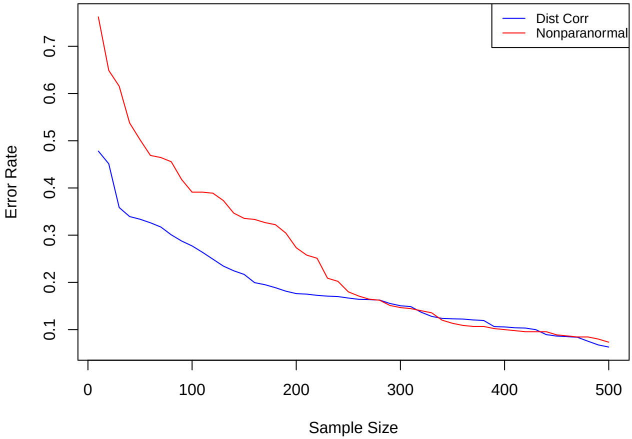
p=10

Nonparanormal vs Dist Corr Method



p=25

Nonparanormal vs Dist Corr Method



p=50

Application on Data 1

Data Description

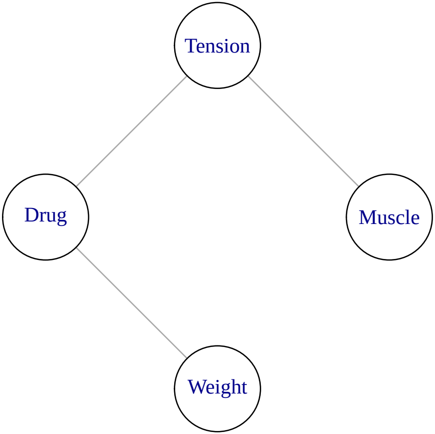
For each individual, the response to the following 4 binary variables are noted

- change in muscle tension (high/low)
- muscle type (type 1/ type 2)
- type of drug taken (drug 1/ drug 2)
- weight of muscle (high / low)

Question of Interest

Conditional independence relations between these variables. This is a very small example, often in medical diagnosis we can have multiple potential response variables. In this case, two potential response variables are muscle tension and weight of the muscle

Application on Data 1



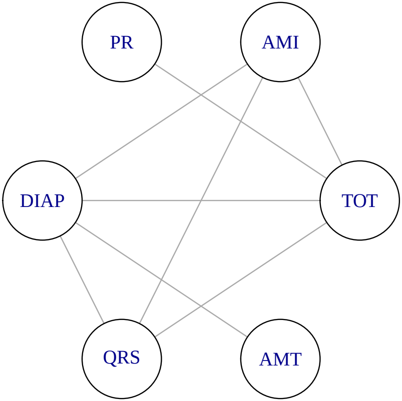
Application on Data 2

Data Description

Here we have another similar example with the following variables

- TOT : Total TCAD plasma level
- AMI : Amount of amitriptyline present in the TCAD plasma level
- AMT : amount of drug taken at time of overdose
- PR : PR wave measurement
- DIAP : Diastolic blood pressure
- QRS : QRS wave measurement

Application on Data 2



Application on Data 3

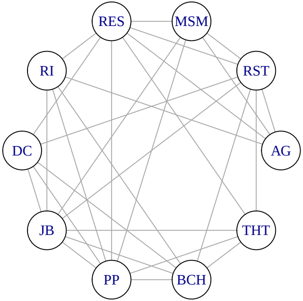
Data Description

Average rating given by 980 individuals to different categories of places they visit. Data is from *tripadvisor.com*.

Abbreviations

- AG = Art Gallery
- DC= Dance Club
- JB = Juice Bar
- RST = Restaurant
- MSM = Museum
- RES = Resort
- PP = Park/ Picnic Spot,
- BCH = Beach
- THT = Theatre
- RI = Religious Instituition

Application on Data 3



- Art Gallery - Museum / Dance Club - Resort / Restaurant - Juice Bar
- Religious Institute $\perp\!\!\!\perp$ Dance Club / Art Gallery $\perp\!\!\!\perp$ Dance Club

Application on Data 4

Data Description

We use a flow cytometry dataset with $p = 11$ proteins measured on $N = 7466$ cells. Each vertex corresponds to the expression level of a protein

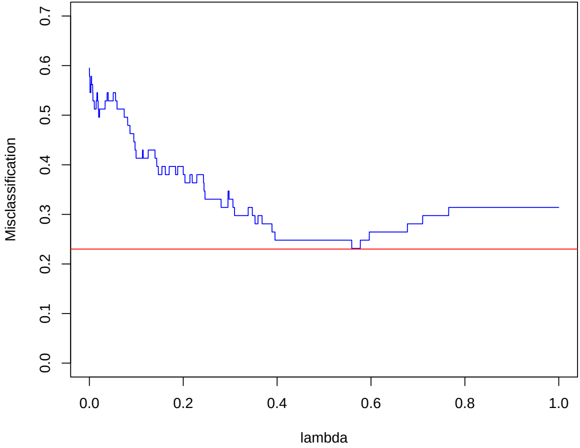
Previous work

Friedman et al used this data set to demonstrate the usage of graphical lasso, but did not arrive at any conclusion or interpretation, neither compared it to the output of the work where the data first appeared (*Sachs et al*)

Why use Nonparametric Method

The data is not multivariate normally distributed, even each of the variables are not univariate normal

Application on Data 4



End Notes

THANK YOU