A Study on Graphical Models

Debarshi Chakraborty
M.Stat 2nd Year

January 31, 2024

Abstract

In this work, we first try to understand what are graphical models and why are they needed. Then we briefly discuss about the existing literature on one of the simplest cases which is when the nodes of the graph jointly follow a multivariate normal distribution. Next, we move on to study what methods can be applied when the data is not normally distributed, which is often the case in practice. To study non gaussian scenarios, we do a comprehensive literature review on nonparametric methods for continuous data and also study discrete graphical models. We also propose a different perspective of error control using multiple testing in the context of graph structure estimation. In most of the setups, we try to study the performance of the methods in different situations like for large sample size, for deviation from normality and for higher dimensions. We do comparisons between two nonparametric methods for continuous data using simulation. Also, we present some analysis of real data.

Contents

1 Introduction to Graphical Models **3**

1.1 Markov Random Fields . . . 3

1.2 Bayesian Networks . . . 5

1.3 Difference between Markov Random Fields and Bayesian Networks . . . 7

2 Gaussian Graphical Models **7**

2.1 Gaussian Markov Random Fields . . . 7

2.1.1 Estimation of parameters . . . 7

2.1.2 Estimation of graph structure . . . 8

2.2 Gaussian Bayesian Networks . . . 9

2.2.1 Estimation of parameters . . . 10

2.2.2 Estimation of graph structure . . . 10

3 Non Gaussian Graphical Models **12**

3.1 Nonparametric approaches for continuous data . . . 12

3.1.1 The Nonparanormal . . . 12

3.1.2 Forest Density Estimation . . . 13

3.1.3 Structure Learning using Distance Correlation . . . 14

3.2 Graphical Models for discrete data . . . 15

3.2.1 Estimation of parameters . . . 16

3.2.2 Estimation of graph structure . . . 17

3.2.3 Methods in high dimensional setup . . . 17

4 A Different Perspective for Error Control **19**

4.1 Usual notion of Error Control . . . 19

4.2 Which error is more serious in Graphical Models? . . . 19

4.3 A Multiple Testing Method for Error Control . . . 19

5 Simulation and Experiment Results **20**

5.1 Comparison between nonparametric methods for continuous data . . . 20

5.2 Application on discrete data . . . 21

5.2.1 Data Description . . . 21

5.2.2 Question of interest . . . 21

5.2.3 Graph structure estimation . . . 22

5.3 Application on continuous data . . . 22

5.3.1 Data Description . . . 22

5.3.2 Question of Interest . . . 22

5.3.3 Graph Structure Estimation . . . 22

1 Introduction to Graphical Models

A *graphical model* in statistics is a graph in which random variables are represented as nodes and an associated family of probability distributions, where the graph keeps track of the independence relations existing between the collection of random variables. Edges may be undirected, which deal with symmetric dependence structure, while directed edges represent cause effect relationships. These independences may come from a prior domain knowledge or may be derived from the data at hand. Advantages of the graphical representation include ease of comprehension, particularly of complicated patterns, reduction of computational burden and ease of comparing probabilities.

Two topics in which we are interested while studying graphical models are representation and learning. Representation involves how to represent a joint probability distribution in a compact manner using a graph, learning corresponds to estimating both the structure of a graph and parameters involved in the model. The primary reasons for studying graphical models can be summarized in the following manner.

1. Graphs are a very intuitive way to visualize relationship between random variables.

2. The number of parameters to be estimated in order to specify a joint probability distribution can be reduced by exploiting the conditional independence properties. This is a huge advantage in high dimensional setup.

3. Queries about conditional dependence relations existing between a set of random variables can be answered just by looking at the graph (or equivalently its adjacency matrix), which is much more convenient than looking at the functional form of the joint distribution.

4. Deciphering cause-effect relationships existing between different variables (this can be done using directed graphs).

5. Once we know the adjacency matrix or equivalently the pairwise, local and global Markov properties that exist in a joint distribution, computation of conditional distributions become much easier, especially when we have high dimensional data and a sparse graph structure.

Graphical models are broadly divided into two types based on the type of the graph used, namely *directed graph* and *undirected graph.* In this study, we include both.

1.1 Markov Random Fields

Here we consider an undirected graph whose nodes are random variables, known as undirected graphical models or Markov Random Fields. We introduce the concepts of adjacency matrix, separation and factorization in this case, how do we represent the edges of a graph mathematically, different properties on conditional dependence that hold in a graph, relation between them and finally see an example.

The *adjacency matrix* A of an undirected graph G with set of nodes $V = \{X_1, X_2, ...X_n\}$ and set of edges E is defined by $A_{ij} = \mathbb{1}\{(X_i, X_j) \in E\}$.

Separation in Undirected Graphs : A subset $C \subseteq V$ is said to separate X_i and X_j if all the paths from

X_i to X_j intersect C. This concept can be generalized to sets also. 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 $X_j \in B$.

Markov properties relative to an undirected graph G :

- Pairwise Markov Property (P) : For any $(X_i, X_j) \notin E$, we have $X_i \perp X_j | V - \{X_i, X_j\}$
- Local Markov Property (L) : For any $X_i \in V$, $\{X_i \perp V - cl(X_i)\}|\ ne(X_i)$ (where $cl(X_i)$ denotes the closure of the node X_i, $ne(X_i)$ denotes the set of neighbours of X_i.)
- Global Markov Property (G) : For disjoint subsets $A,\ B,\ C$ of V, such that C separates A and B in G, $A \perp B|\ C$ or an equivalent notation is used as $X_A \perp X_B|\ X_C$.

Theorem 1. *If a probability distribution P on D has a continuous and positive density with respect to a product measure μ, then (G) $\iff$ (L) $\iff$ (P), where $\{X_1, X_2, ..., X_n\}$ takes values on D.*

Remark : The right sided implications hold for any probability distribution.

A probability distribution P on D is said to factorize according to a graph G if for all complete subsets $A \subseteq V$, there exists non negative functions $\psi_A(.)$ that depend on $\mathbf{x}$ only through $\mathbf{x_A}$ and there exists a product measure μ on D, such that P has density $f(.)$ with respect to μ where $f(.)$ has the form $f(\mathbf{x}) \propto \prod_{\mathbf{A} \subseteq \mathbf{V}} \psi_{\mathbf{A}}(\mathbf{x})$. Remark : The functions $\psi_A(.)$, also referred to as *potential functions* are not unique.

Theorem 2. *For any undirected graph and any probability distribution on D it holds that*

$$(F) \implies (G) \implies (L) \implies (P)$$

Theorem 3. *(Hammersley and Clifford) A probability distribution P with a positive and continuous density f with respect to a product measure μ satisfies the pairwise Markov property relative to an undirected graph G if and only if P factorizes according to G.*

Remark by *Lauritzen [Lau96]* : Actually when P has a positive continuous density, it can be shown that $(P) \implies (F)$. The continuity condition can considerably be relaxed (*Koster, 1994*), whereas positivity is essential.

In our entire discussion, a central notion is the conditional independence between random variables, the graph (or equivalently the adjacency matrix) keeping track of the conditional independence relations. Since when we are in a setup where all the Markov properties are equivalent, it is reasonable to work with pairwise Markov properties. Thus, we define

$$X_i \perp X_j|\ V - \{X_i, X_j\} \iff \text{X}_i \text{ and } X_j \text{ are separated by } V - \{X_i, X_j\} \iff \text{X}_i \text{ and } X_j \text{ are not adjacent}$$

Hence, our objective now is to estimate the structure of the graph i.e. which edges are present and which are absent. Once we are able to do that, by the virtue of **Theorem 3**, we will be able to write down the joint density in a factorized form, which will facilitate easier computation of conditional distributions, reduction in number of parameters to be estimated, answering queries about conditional dependence relations existing between different variables, etc. As mentioned earlier, the parameters of interest will vary from problem to problem. We consider a very simple example for illustration.

Example 1. Suppose we have n observations from $(\mathrm{X}_1, X_2, ..., X_p) \sim N_p(\mu, \Sigma_p)$. We know the form of the multivariate normal density

$$f(x) \propto \; e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\mathbf{T}}\mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)}$$

with parameters μ and Σ. The density can be reparametrized in terms of the precision matrix $\Sigma^{-1} \; = \; \Theta$. Then the density can be written in the following form :

$$f(x) \propto e^{-\frac{1}{2}\mathbf{x^T\Theta x}+(\mathbf{\Theta}\mu)^{\mathbf{T}}\mathbf{x}}$$

Breaking up the exponent in this form of the multivariate normal density gives us three kinds of terms :

$$\Theta_{ii}\; x_i^2,\; (\Theta\mu)_i\; x_i \text{ and } \Theta_{ij}\; x_i x_j$$

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. Clearly, here our parameter of interests are the precision matrix Θ and the mean vector μ. A natural question arises why should we work with Θ instead of Σ ?

Theorem 4. *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.*

Hence, we can represent the conditional independence relations between different components of the random vector by the precision matrix. Note that, just replacing the non zero entries of the precision matrix by 1 yields our adjacency matrix. So, if we know / can estimate the adjacency matrix , we won't have to estimate the whole precision matrix Θ to specify the multivariate normal distribution. If we had gone for estimating Σ directly, we had to estimate all $\frac{p\,(p+1)}{2}$ unique entries of it. In a high dimensional setup and sparse graph structure, this reduction in number of parameters helps significantly, since with fixed amount of data it becomes difficult to estimate a large number of parameters efficiently.

1.2 Bayesian Networks

Here we consider a *directed acyclic* graph whose nodes are random variables, known as Bayesian Networks. We introduce the concepts of adjacency matrix, separation and factorization in this case, how do we represent the directed edges mathematically, different properties on conditional dependence that hold in a graph, relation between them and finally see an example. A *directed edge* from node X_i to node X_j is represented by an arrow $X_i \; \rightarrow \; X_j$ or simply the ordered pair (X_i, X_j). Note that, unlike undirected edges, this relation is not symmetric. In other words, if E denotes the set of directed edges of a graph, then $(X_i, X_j\;) \in \; E$ does not necessarily imply $(X_j, X_i\;) \in \; E$, which is the case for undirected edges. However, throughout our study we consider graphs with no loops and no multiple edges.

Adjacency matrix A of a DAG G with set of nodes $V \; = \; \{X_1, X_2, ...X_n\}$ and set of directed edges E is defined by $A_{ij} \; = \; \mathbb{1}\{(X_i, X_j) \in E\}$.

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.*

A probability distribution P admits a *recursive factorization* (denoted by (DF)) according to a graph G if there exists non negative functions $k^{\alpha}(.,.)$, $\alpha \in V$ defined on $D_\alpha \times D_{pa(\alpha)}$ such that $\int k(y_\alpha, x_{pa(\alpha)})\mu_\alpha(dy_\alpha) = 1$ and P has density f with respect to μ, where $f(x) = \prod_{\alpha \in V} k^{\alpha}(x_\alpha, x_{pa(\alpha)})$. Here D denotes the domain in which a vertex (usually an RV) can take values and $pa(\alpha)$ denotes the set of parents of the node α.

In the directed setup also, we have pairwise, local and global Markov properties similar to their undirected counterparts and it can be shown that under certain conditions (for example positivity of the probability distribution) it can be shown that those are equivalent.

Theorem 5. *Let G be a directed acyclic graph. For a probability distribution P on D which has density with respect to a product measure μ. Then (DF) $\iff$ (DG) $\iff$ (DL), where (DL) denotes directed local Markov property and (DG) denotes directed global Markov property.*

In our entire discussion, a central notion is the conditional independence between random variables, the graph (or equivalently the adjacency matrix) keeping track of the conditional independence relations. Since when we are in a setup where all the Markov properties are equivalent, it is reasonable to work with pairwise Markov properties. Thus, we define for any pair (X_i, X_j)

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

A directed acyclic graph which encodes this type of independence relations are also known as *Bayesian Networks.* Hence, our objective now is to estimate the structure of the graph 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 $X_i \to X_j$ means that the variable X_i has a causal effect on the variable X_j. Once we are able to do these two tasks, by the virtue of **Theorem 5**, we will be able to write down the joint density in a factorized form, which will facilitate easier computation of conditional distributions, reduction in number of parameters to be estimated, answering queries about conditional dependence relations existing between different variables, etc.

Example 2. 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.

1.3 Difference between Markov Random Fields and Bayesian Networks

Firstly, Bayesian Networks represent causal relations existing between random variables but Markov Random Fields do not, they represent only probabilistic dependence, which is symmetric in nature. Secondly, in Bayesian Networks, cyclic dependence structures are not allowed, this restriction is not there in the case of Markov Random Fields. Finally, in Bayesian Networks, the joint probability distribution is represented as the product of conditional distributions, which is not necessarily the case in Markov Random Fields where the joint density is represented normalised product of potential functions.

2 Gaussian Graphical Models

This is a particularly simple subclass of distributions which make very strong assumptions, still has been treated specially in Statistics literature over the years, graphical models are no exception. It has been studied much more extensively compared to other kinds of graphical models, especially in the undirected setup.

2.1 Gaussian Markov Random Fields

We consider the setup of *Example* 1, where we have seen our task boils down to determining which entries of Σ^{-1} are zero and estimating the "non-zero" entries of the same, which are our potential functions in this case.

We can opt for a two step approach, first estimate the adjacency matrix and then estimate the non-zero entries of the precision matrix.

2.1.1 Estimation of parameters

At first, suppose the adjacency matrix i.e. the structure of the graph is known to us, only the non-zero entries of $\Theta \ = \ \Sigma^{-1}$ are to be estimated. In a situation where the sample covariance matrix

$$S \ = \ \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

is full rank ($n \geqslant p+1$) and all edges are present then the log-likelihood is of the form

$$l_n(\Theta) \ = \ log|\Theta| \ - \ tr(S\Theta)$$

and thus we can simply obtain $\hat{\Theta} \ = \ S^{-1}$. But in practice, we will often have $p \geqslant n$ (which is very common in the age of high dimensional data) and a graph with many missing edges, where the above approach is not going to work, in fact the sample covariance matrix may not be invertible in the first place.

Hence, we would now like to maximise the likelihood under the constraint that some pre-defined subset of parameters are 0. This is an equality constrained optimization problem. A number of methods have been proposed to solve it, which exploit the simplifications that arise from decomposing a graph into maximal cliques. But there exists a simpler approach using regression which exploits sparsity in a different manner. The usefulness of this approach will be more apparent while dealing with the problem of estimating the graph structure. It turns out that if we partition $(X_1, X_2, ..., X_p) \ = \ (Z, Y)$ then we have (*Mardia et al., 1979)*

$$Y|Z = z \ \sim \ N(\mu_Y + (z - \mu_z)^T \Sigma_{ZZ}^{-1}\sigma_{ZY}, \ \sigma_{YY} - \sigma_{ZY}^T \Sigma_{ZZ}^{-1}\sigma_{ZY})$$

where

$$\Sigma = \begin{bmatrix} \Sigma_{ZZ} & \sigma_{ZY} \\ \sigma_{ZY}^T & \sigma_{YY} \end{bmatrix}$$

Partitioning the precision matrix Θ in a similar way we get the regression coefficient of Y on Z to be $\beta = -\frac{\theta_{ZY}}{\theta_{YY}}$. It is to be noted that the dependence of Y on Z is in the mean term alone and zero elements in β or equivalently θ_{ZY} implies conditional independence of the corresponding elements of Z and Y given the rest. Thus, the dependence structure can be learnt through multiple linear regression. For detailed method, refer *Algorithm 17.1, Hastie, Tibshirani, and Friedman [HTF09].*

2.1.2 Estimation of graph structure

Now we consider a more realistic situation where the graph structure is unknown to us. Hence, we have to somehow make a good estimate of which edges are present and which are not, essentially which pairs random variables are conditionally independent given the others and which pairs are not. Once we get the structure, then we can estimate the relevant parameters. So, a very simple, natural and intuitive way is to start with *multiple testing* i.e. test for each of the $\binom{p}{2}$ pairs of nodes that for which the conditional independence relation hold. Here, since we want to test for conditional independence, we can work with partial correlations, which coincides with the conditional correlation if the random variables are jointly distributed as the multivariate normal. 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

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

If the test is accepted, we put the corresponding entry of the adjacency matrix to be 0 and if rejected, we put 1. The distributions of r^p_{ij} has the same form as the distribution of the correlation coefficient r_{ij}, with some changes in parameters. (*Theorem 4.3.5, Anderson [And03]*)
Another approach is due to *Edwards, D (2000)*, which instead of testing all the pairs at once, adopts a backward elimination approach to detect which edges are present. The null and alternative hypotheses remain same as in the previous approach. We start with the graph with all edges present G_0 (say). For each of the $\binom{p}{2}$ pairs, a likelihood ratio test is carried out instead of a t-test. If L denote the likelihood ratio test statistic then $-2\,ln(L)$ follows a $\mathcal{X}^2$ distribution with appropriate degrees of freedom asymptotically under H_0. Based on these test, the edge $\{X_i, X_j\}$ with the largest deviance or equivalently the largest p-value is removed. Now considering the new graph (say G_1) as our null model, we repeat the same procedure till no edges are removed. But unlike in the previous case, here the distributions of the elements of the final precision matrix we obtain is very difficult to find since the estimates are based on successive M.L.E s.

There are some issues with the above mentioned approaches. When $n - p$ is small, the estimate r^p_{ij} is very unstable, for n < p the sample covariance matrix is not invertible, for $n < p$ the first test statistics cannot be computed, for n < p situations, controlling the $P(type\ I\ error)$ becomes difficult.

Some improvements have been proposed (*Drton and Pearlman (2004), (2007)*) to control the family wise error rate in order to control the overall type I error, based on asymptotic properties of the sample correlation coefficient, Sidak's Inequality and Fisher's variance stabilizing transformation, still the problems in the situation p > n persist.

The most recent method that was proposed by a number of authors is a one step procedure of estimating both the structure and the parameters of the graphical model simultaneously, is based on the L_1 (lasso) regularization. *Meinshausen and Bühlmann [MB06]* tried to estimate the non zero entries of Θ. To do this, they fit a lasso regression using each variable as the response and the others as predictors. The component θ_{ij} is then estimated to be nonzero if either the estimated co efficient of X_i on X_j is non zero or vice versa (alternatively they use an AND rule). They show that asymptotically this procedure consistently estimates the set of nonzero elements of Θ. *Friedman, Hastie, and Tibshirani [FHT07]* take a more systematic approach with the lasso penalty, following the procedure of estimation using multiple linear regression, just in this case instead of the log likelihood they try to maximise the penalized log likelihood

$$l_n(\Theta) \;=\; log|\Theta| \;-\; tr(S\Theta) \;+\; \lambda||\Theta||_1$$

where $||\Theta||_1$ is the L_1 norm, the sum of the absolute values of the elements of Σ^{-1}. The negative of the penalised log likelihood is a convex function of Θ. It turns out that one can simply replace the modified regression step of *Algorithm 17.1*, *Hastie, Tibshirani, and Friedman [HTF09]* by a modified lasso regression step. Details method discussed in *Algorithm 17.2* of the same reference. The resulting method is known as *Graphical Lasso*. The algorithm is extremely fast, and can solve a moderately sparse problem with 1000 nodes in less than a minute. It is easy to modify the algorithm to have edge-specific penalty parameters.

2.2 Gaussian Bayesian Networks

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 β_o, $\beta_1, ...,$ β_k and σ^2 such that

$$Y|X = x = \beta_o + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. A directed (acyclic) graphical model is called a Gaussian Bayesian network if all the variables are continuous and every variable has a linear gaussian model of its parents. In a Gaussian Bayesian network, each variable X_i is modeled as a linear function of its parents plus normally distributed random noise. One important result is that there is one-to-one correspondence between a multivariate Gaussian distribution and a Gaussian Bayesian network.

Theorem 6. *Let Y have a linear gaussian model of its parents $X_1, ..., X_k : Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$. We also assume that $(X_1, ..., X_k)$ are jointly multivariate normal $N_k(\mu, \mathbf{\Sigma})$. Then the joint distribution of $(Y, X_1, ..., X_k)$ is multivariate normal with $Cov(Y, X_i) = \sum_{j=1}^{k} \beta_j \Sigma_{ij}$, $V(Y) = \sigma^2 + \beta^T \Sigma \beta$ and $E(Y) = \beta_0 + \beta^T \mu$, where $\beta = (\beta_1, ..., \beta_k)$.*

The converse of this theorem is also true, which states that any multivariate normal distribution can be converted to a Gaussian Bayesian network.

Theorem 7. *Let P be the joint distribution of d-dimensional multivariate Gaussian random vector $\boldsymbol{X} = (X_1, ..., X_d)$. For any ordering of the variables $X_{\tau(1)}, ..., X_{\tau(d)}$, we can construct a DAG G such that P is satisfies the Markov properties with respect to G, and where $X_\tau(i)$ is a linear Gaussian model of its parents $pa(X_{\tau(i)}) \subset \{X_{\tau(1)}, ..., X_{\tau(i-1)}\}$ for all i.*

Suppose we are given n independent observations from $\mathbf{X}$=$(X_1, X_2, ..., X_d) \sim p_\theta(x)$. In this context also, two questions may arise. Firstly, given a DAG G which represents the conditional dependencies of the distribution $p_\theta(x)$, how do we estimate its parameters? Secondly, how to estimate the underlying DAG G? In case of Bayesian Networks, we mention the estimation procedures in general, since same procedures are followed for any Bayesian Network.

2.2.1 Estimation of parameters

Once G is given, the task of estimating the parameters of the joint distribution can be greatly simplified by the application of the Markov property. Let $\theta = (\theta_1, ..., \theta_d)$ be the set of parameters, the joint distribution $p_\theta(x)$ can be written as

$$p(x;\theta) = \prod_{j=1}^{d} p(x_j | \, pa(x_j); \theta_j)$$

Given n data points $x_1, x_2, ..., x_n$, the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} p(x_i;\theta) = \prod_{i=1}^{n} \prod_{j=1}^{d} p(x_{ij} | \, pa(x_j); \theta_j)$$

where x_{ij} is the value of X_j for the i^{th} data point and θ_j are the parameters for the j^{th} conditional density. We can then estimate the parameters by maximum likelihood.It is easy to see that the log-likelihood decomposes according to the graph structure:

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

Therefore we can maximize the contribution to the log-likelihood of each node independently. When there is not enough information from the data points, we could also regularize the log-likelihood to avoid overfitting. In many applications, observed data may not include the values of some of the variables in the DAG. We refer to these variables as hidden variables. If Z denotes be the hidden variables, the log-likelihood can be written as

$$l(\theta) = \sum_{i=1}^{n} logp(x_i | \, \theta) = \sum_{i=1}^{n} \int_{z_i} p(x_i, z_i, \theta) dz_i$$

With hidden variables, the log-likelihood is no longer decomposable as in and maximizing the log-likelihood is often difficult. This can be approached using the EM algorithm.

2.2.2 Estimation of graph structure

Estimating a DAG from data is very challenging due to the enormous size of the space of DAGs. Existing methods can be roughly divided into two categories: *(i) constraint-based methods* and *(ii) score-based methods*. Constraint-based methods use statistical tests to learn conditional independence relationships (called constraints in this setting) from the data and prune the graph-searching space using the obtained constraints. In contrast, score-based algorithms assign each candidate DAG a score reflecting its goodness of fit, which is then taken as an objective function to be optimized. We discuss a constraint based method : the PC

Algorithm in brief. Let P be a probability distribution. We assume that there exists a DAG G such that $I(P) = I(G)$, where $I(A)$ = set of independence relations in A. Unfortunately, we cannot identify the exact DAG G, only the Markov equivalence class of G can be identified.

Theorem 8. *(Verma and Pearl, 1990) 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.

A *partially directed acyclic graph* (PDAG) is an acyclic graph with both directed and undirected edges i.e. one cannot trace a cycle by following the directions of directed edges and/or any direction of undirected edges. PDAG K is said to be a *complete partially directed acyclic graph* (CPDAG) with respect to the equivalence class of G if (i) Skeleton(K) = Skeleton(G) and (ii) K contains a directed edge $X \to Y$ if and only if any DAG in the equivalence class contains $X \to Y$. Thus, all graphs in the equivalence class agree with its CPDAG on the directed edges and for any undirected edge of the CPDAG, there exists at least two different DAGs in the equivalence class which disagree on the direction of that edge. Hence, it is enough to identify the CPDAG of the Markov equivalence class of G.
The PC Algorithm operates in two steps : (i) Skeleton identification (ii) CPDAG identification. Let adj(C, i) represents the adjacency nodes of X_i in an undirected graph C. The outer loop $k = 0, 1..., d$ indexes the size of the separating sets. $\mathcal{A}$ denotes the class of separating sets and $\mathcal{A}_{ij}$ denotes the separating set of X_i and X_j. The algorithm is as follows :

1. Step 1 : Skeleton Identification
   - Input : n iid observations from $V = (X_1, X_2, ..., X_d)$
   - Initialize C = complete undirected graph on V and $\mathcal{A}_{ij} = \phi \; \forall \; X_i, \; X_j \in V$
   - Repeat for any two adjacent nodes X_i, X_j such that $|adj(C, i) - j| \geqslant k$, for every $A \in adj(C, i) - \{j\}$ with $|A| = k$, remove the edge $\{X_i, X_j\}$ from C and set $\mathcal{A}_{ij} = A$ if $H_0 : X_i \perp X_j | A$ is accepted
   - Repeat the previous step for $k = 0, 1, 2, ..., d$
   - Output : Estimated skeleton C and the class of separating sets $\mathcal{A}$
2. Step 2 : CPDAG Identification
   - Input : Estimated skeleton C and the class of separating sets $\mathcal{A}$
   - Initialize $K = C$
   - For every pair of non adjacent nodes X_i and X_j with common neighbour X_k, replace X_i–X_k–X_j by $X_i \to X_k \leftarrow X_j$ if $K \notin \mathcal{A}_{ij}$
   - Find a subgraph in K for which at least one of unshielded collider rule, acyclicity rule or hybrid rule apply, and the add corresponding directions
   - Repeat previous step until convergence
   - Output : Identified CPDAG K

For Gaussian Bayesian Networks, the test for $H_0 : X_i \perp X_j \mid A$ is much simpler compared to other cases.

3 Non Gaussian Graphical Models

3.1 Nonparametric approaches for continuous data

The assumption of multivariate normality is too strict to hold most of the times. As a consequence, non gaussian graphical models arise more often. A large class of examples can be found or constructed where all the vertices of the graph are discrete random variables. One can refer to *Chapter 3* of *Koller and Friedman [KF09]* or *Chapter 4* of *Lauritzen [Lau96]* for several such examples. *Chapter 8* of the former presents a generalized discussion on the exponential family. However, the most general discussion in the continuous case is done by *Lafferty, Liu, and Wasserman [LLW12]*, which focuses on undirected graphical models. It represents two approaches of graphical modelling. One of them is a semiparametric extension of the gaussian graphical model allowing arbitrary graph structure, while the other is fully nonparametric but imposes some restrictions on the structure of the graph. Regarding DAGs, we have seen that the only point where the normality assumption comes into play is while testing for conditional independence, the rest of the method does not require any parametric assumptions. Hence for Bayesian Networks, our main challenge is testing for conditional independence between random variables. Here, in this section we first discuss the two approaches introduced by *Lafferty, Liu, and Wasserman [LLW12]* briefly.

3.1.1 The Nonparanormal

We say a random vector X=$(X_1, ..., X_d)^T$ has a nonparanormal distribution an write $X \sim NPN_d(\mu, \Sigma, f)$, in case there exists functions $\{f_j\}_{j=1}^d$ such that $(f_1(X_1), ..., f_d(X_d))^T \sim N_d(\mu, \Sigma)$. When the functions $\{f_j\}_{j=1}^d$ are monotone and differentiable, the joint pdf of X is given by

$$p(x) = \tfrac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp[-\tfrac{1}{2}(f(x)-\mu)^T\Sigma^{-1}(f(x)-\mu)] \textstyle\prod_{j=1}^d |f_j^{'}(x_j)|$$

To make the family identifiable we demand that the functions $\{f_j\}_{j=1}^d$ preserve the means and variances i.e. $\mu_j = E(X_j) = E(f_j(X_j))$ and $\sigma_j^2 = V(X_j) = V(f_j(X_j))$. We choose $\{f_j\}_{j=1}^d$ such that they follow univariate normal. If $F_j(x)$ denote the marginal CDF of X_j then

$$f_j(x) = \mu_j + \sigma_j \Phi^{-1}(F_j(x))$$

Lemma 3.1 of *Lafferty, Liu, and Wasserman [LLW12]* show that the conditional independence structure of remains invariant $(X_1, ..., X_d)^T$ under such transformation.
A two step procedure is followed to estimate the graph :

1. The functions $\{f_j\}_{j=1}^d$ 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 $\{f_j(X_j)\}_{j=1}^d$ is justified, it is *assumed* that $(f_1(X_1), ..., f_d(X_d))^T$ follows multivariate normal. The validity of this assumption is not justified, thus the procedure remains incomplete.

3.1.2 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 (densities in the family is supported by some $F \in \mathcal{F}_d$). Define the oracle forest density

$$q^* = argmin_{q \in \mathcal{P}_d} D(p^*, q)$$

where $D(p,q) = \int p(x) log(\frac{p(x)}{q(x)})dx$ is the Kullback-Leibler divergence.
Proposition 4.1 of *Lafferty, Liu, and Wasserman [LLW12]* states that there exists $F^* \in \mathcal{F}_d$ such that

$$q^* = p_{F*} = \prod_{(x_i,x_j)\in E_{F*}} \frac{p^*(x_i,x_j)}{p^*(x_i)p^*(x_j)} \prod_{x_k \in V_{F*}} p^*(x_k)$$

If the true density $p^*(x)$ is known, by the above proposition our problem would be reduced to finding the best forest F_d^* such that $F_d^* = argmin_{F\in\mathcal{F}_d} D(p^*, p_F^*)$. Subsequent discussions show that this can be done by maximising $\sum_{(X_i,X_j)\in E_{F*}} I(X_i, X_j)$ where

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

Algorithm (Kruskal/ Chow Liu) is available to get a guaranteed solution to this problem. However, this procedure is of no practical use since the true density $p^*(.)$ is unknown. Hence, we use kernel density estimates to get an estimate of each $I(X_i, X_j)$.
Again, this is a two step method as follows :
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.

3.1.3 Structure Learning using Distance Correlation

Here also we use a completely nonparametric approach and also allowing arbitrary graph structure. The idea is based on **distance correlation** (Székely, Rizzo, and Bakirov [SRB07]), a measure of ”dependence” between two random vectors , with nice asymptotic properties.
We first visit the approach due to Fan, Feng, and Xia [FFX19], a part of their work is dedicated to undirected graphical models. We summarize that part below. To learn the structure of the undirected graph with set of vertices $V = X_1, X_2, ..., X_p$, we need to test $X_i \perp X_j | V - \{X_i, X_j\}$

We assume

$$X_i^{(k)} = \beta_{1,ij}^T Z^{(k)} + \epsilon_i^{(k)} \text{ and } X_j^{(k)} = \beta_{2,ij}^T Z^{(k)} + \epsilon_j^{(k)} \;\; \forall \; k = 1, 2, ..., n \;\text{ where } Z^{(k)} = X_{(-i,-j)}^{(k)}$$

We decide whether an edge between X_i and X_j should be drawn directly by testing $H_0 : X_i \perp X_j | \, \mathcal{L}(Z)$ where $\mathcal{L}(Z)$ is the linear space spanned by all the variables except X_i and X_j. More specifically, for each pair of nodes $\{(i, j) : 1 \leqslant i \leqslant j \leqslant d\}$, we test $H_{0,ij} : \epsilon_i \perp \epsilon_j$ and then summarize the testing results by a graph in which nodes represent variables in V and the edge between node X_i and node X_i is drawn only when $H_{0,ij}$ is rejected at level α.

The Procedure :

- Run LASSO regression on the above mentioned linear models and obtain the estimated co-efficient vectors $\beta_{1,ij}$ and $\beta_{2,ij}$
- 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

Another approach is to use **conditional distance correlation** introduced by Wang et al. [Wan+15] for each of the tests, then go for FDR control.

3.2 Graphical Models for discrete data

An obvious and large class of non gaussian graphical models are those where all the nodes of the graph are discrete random variables. Here , we will consider an example which we will use throughout to understand definitions, theorems, etc. Let us consider 3 random variables X_1, X_2, X_3. Each of them can take 2 values. We need to consider at least 3 variables in order to learn something about conditional independencies that are of particular interest in graphical models.

Example 3. We consider a famous data set giving information about the survival rate of 715 infants attending two clinics and the amount of care received by the mother, where the amount of care is classified as either **more** or **less**. Essentially, we are trying to study the $2 \times 2 \times 2$ contingency table which looks like :

n(clinic, care, survival) Clinic	Care	Survival	
		No	Yes
Clinic 1	less	3	76
	more	4	293
Clinic 2	less	17	197
	more	2	23

One may be interested in queries like, given a particular clinic, whether care and survival are independent or not. If we consider these as our random variables X_1, X_2, X_3 respectively, then the query can be expressed mathematically as whether $X_2 \perp X_3 | X_1$.

For our convenience, we start with log linear models for binary variables. Let us consider the above example only. The trivariate Bernoulli random vector $(X_1, X_2, X_3)^T$ can take value $(0,0,0), (0,0,1), ..., (1,1,1)$. The 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

$$logP(x_1, x_2, x_3) = u_0 + u_1 x_1 + u_2 x_2 + u_3 x_3 + u_{12} x_1 x_2 + u_{13} x_1 x_3 + u_{23} x_2 x_3 + u_{123} x_1 x_2 x_3$$

Note that, in this particular setup $X_2 \perp X_3 | X_1 \iff u_{23} = 0$ 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. To do that, we need a general definition of log linear expansion for non binary data at the first place.

Now, 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. However, in order to get rid of redundant parameters, we impose the constraint $u_a(x_a) = 0$ whenever $x_i = 0$ and $i \in a$ ($a \subseteq \{1, 2, ..., k\}$). In *Example 3*, we will have $u_1(0) = u_2(0) = u_3(0) = u_{12}(0,0) = u_{13}(0,0) = u_{23}(0,0) = u_{12}(0,1) = u_{12}(1,0) = u_{13}(0,1) = u_{13}(1,0) = u_{23}(0,1) = u_{23}(1,0) = u_{123}(0,0,0) = ... = u_{123}(1,1,0) = 0$. This ensures that the complete log linear expansion has as many "u" terms as there are cells in the $d_1 \times d_2 \times ... \times d_K$ contingency table.

Also, note that the setup is consistent with the simple 3 way table of binary random variables where the "u" terms were constants. The *log linear expansion* of the cross classified multinomial distribution P_K is

$$\log\, P_K(x) = \textstyle\sum_{a \subseteq K} u_a(x_a)$$

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 $i \in a$.
Now, we can denote the conditional independence relations between different random variables by the virtue of the following theorem.

Theorem 9. *If $(X_a, X_b, X_c)^T$ be a partition of the random vector $(X_1, X_2, ...X_k)^T$ then $X_b \perp X_c | X_a$ if and only if all u terms containing co-ordinates from both b and c in the log linear expansion are equal to 0.*

3.2.1 Estimation of parameters

Here, we start with as if the graph structure is known to us or we have already estimated the graph or equivalently the adjacency matrix. Say in *Example 3*, we know the graph structure to be like

$$X_2 \text{——} X_1 \text{——} X_3$$

i.e. given a particular clinic, survival and care are conditionally independent of each other. We will use maximum likelihood estimation to estimate the parameters (cell probabilities).

Theorem 10. *Let N be the total frequency of the contingency table. The maximum likelihood estimator of the graphical model satisfies $\hat{n}_a = N \times \hat{P}_a = n_a$ whenever the subset of vertices a in the graph form a clique.*

The proof is in the similar lines of derivation of MLE of cell probabilities of a two way contingency table with no additional constraints. Clearly, in the above example

$$\hat{P}(x_1) = \tfrac{n(x_1)}{N} \text{ and } \hat{P}(x_1, x_2) = \tfrac{n(x_1,x_2)}{N},\; \hat{P}(x_1, x_3) = \tfrac{n(x_1,x_3)}{N}$$

Then, using the conditional independence relation, the individual cell probabilities can be estimated as

$$\hat{P}(x_1, x_2, x_3) = \tfrac{\hat{P}(x_1,x_2)\hat{P}(x_1,x_3)}{\hat{P}(x_1)}$$

The estimates of the joint probabilities can be calculated from the following tables

n_{12}	Care	Care
Clinic	Less	More
Clinic 1	179	297
Clinic 2	214	25

n_{13}	Survival	Survival
Clinic	Yes	No
Clinic 1	7	469
Clinic 2	19	220

The maximum likelihood estimates may not always have a closed form estimate. Even if they admit of a closed form estimate, with increasing dimensions, it may be very complicated notationally to write down the mathematical form. But we know the idea exactly how to calculate the MLE for the entries of each clique from **Theorem 10**. Hence, we can do the estimation by an algorithm called **Iterative Proportional Fitting**, which is based on this idea. The advantage of using this algorithm is, we do not need to know beforehand whether there exists closed form estimated of MLE or not. If no closed form estimates exist, the algorithm will converge to the MLEs and if indeed closed form estimates exist, it will converge to that estimate in one iteration.

3.2.2 Estimation of graph structure

Given any graph structure, we know how to estimate the parameters of the contingency table.Now, we need to decide which is the best graph structure for our data. Here, we will use likelihood ratio test. For the time being, we will consider $n >> p$, since we are going to use an asymptotic result on the null distribution of the likelihood ratio test statistic. Let L_0 denote the log likelihood of the saturated model i.e.

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

and let L_M denote the log likelihood of any other model i.e.

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

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

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

Larger the model deviance, poorer the fit. 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. We want to test

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

We use the test statistic

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

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. This is a feasible method for testing whether some edges are present or absent in our graphical model. 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

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

where dim(M) is the number of parameters in the model M.

3.2.3 Methods in high dimensional setup

As we have seen in the previous section, an asymptotic result is used to estimate the structure of the graph, which naturally fails if we are working with finite sample size and a large number of variables ($n < p$). Here we discuss how to estimate the graph structure in such a situation.

Suppose we have p categorical variables $X_1, X_2, ..., X_p$ each taking $d_1, d_2, ...d_p$ values respectively. We consider the joint distribution of $X = (X_1, X_2, ..., X_p)^T$ as

$$f(X_1, X_2, ..., X_p) \propto \exp(\sum_i \theta_{ii} X_i + \sum_{i,j} \theta_{ij} X_i X_j)$$

where $\Theta = ((\theta_{ij}))$ is the $p \times p$ matrix specifying the conditional independence structure between the random variables. Here $\theta_{ij} = 0 \iff X_i \perp X_j | V - \{X_i, X_j\}$ i.e. their corresponding nodes are not adjacent in the graph. The justification behind considering only the pairwise interaction effects is pointed out by Ravikumar, Wainwright, and Lafferty [RWL10].

Assumption : We consider a restricted framework where we assume an interaction of order k is present if and only if all the interaction effects of order $2, 3..., k-1$ are present. It can be shown, under this restriction (also known as **hierarchical models**), discovering the presence of pairwise interaction effects is is enough to detect presence of higher order interaction effects.

Penalised Logistic Regression : We consider the conditional distribution of the variables X_k (without loss of generality) given the rest of the variables. At first, let us consider the simple case where X_k is a binary random variable. Some elementary calculations can be done to see the form of the conditional distribution as follows :

$$f(X_k|X_{-k}) \propto \frac{exp(\theta_{kk}X_k+\sum_{j\neq k}\theta_{kj}X_kX_j)}{1+exp(\theta_{kk}X_i+\sum_{j\neq k}\theta_{kj}X_kX_j)}$$

This gives rise to a logistic regression problem with X_i as the response variables and others as the predictors. If we have n i.i.d observations $D_n = X_1, X_2, ...X_n$, we estimate the parameter vector θ_{-k} by using a l_1 penalised logistic regression of X_k on the rest of the variables, denoted by X_{-k}. This leads us to solving the convex optimization problem of

$$\min_{\theta_{-k}}\{l(\theta, D_n) + \lambda||\theta_{-k}||_1\}$$

where $l(\theta, D_n) = -\frac{1}{n}\sum_{i=1}^{n}\log f(X_k^i|X_{-k}^i)$ and λ is a tuning parameter, typically depending on the sample size (n), the data dimension (p) and the maximum size of the neighbourhood (d). This can be implemented using a coordinate descent algorithm, for each of the nodes. Since each of the logistic regressions are fit separately, $\hat{\theta}_{ij}$ and $\hat{\theta}_{ji}$ will in general not be equal, especially when sample size is small. These estimates can be aggregated by taking either the minimum of the maximum of the two. Alternatively, maximisation of a different objective function is done with a common penalty parameter λ. For details, see Guo et al. [Guo+10].

In the case where the response variable can take 3 or more values, we can use multi class logistic regression with l_1 penalty instead of the simple logistic regression. The methodology remains same otherwise.

Advantage of considering pairwise interactions only is that we have no interaction terms between the predictors in the logistic regression (can be seen from the form of the conditional distribution), which would have increased the number of parameters exponentially.

4 A Different Perspective for Error Control

4.1 Usual notion of Error Control

Coventionally, in Statistics literature, we frame our null hypothesis in such a way where the type 1 error is considered to be more serious than the type 2 error. Thus, our approach is always controlling the type 1 error at first and then trying to 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).

4.2 Which error is more serious in Graphical Models?

To estimate the structure of a graph we test hypotheses of the form

$$H_0 : X_i \perp X_j |\, X_{-(i,j)} \text{ vs } H_1 : \text{not } H_0$$

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_1, X_2, X_3. We discuss the consequences of the two types of errors.

- **Type 1 error** : H_0 is true but we reject it i.e. we are "missing out the information" that $f(x_1, x_2 |\, x_3) = f(x_1 |\, x_3) f(x_2 |\, x_3)$, 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_1 and X_2 given X_3 does not hold we write $f(x_1, x_2 |\, x_3) = f(x_1 |\, x_3) f(x_2 |\, x_3)$, which is completely "incorrect". This will lead us to wrong conclusions.

Clearly, in this setup, the Type 1 error is less serious than the type 2 error. Hence, our primary objective should be controlling the number of cases where H_0 is actually false but we accept it. Unlike other problems in statistics, 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.

4.3 A Multiple Testing Method for Error Control

Let us consider n null hypotheses out of which n_0 are true and n_1 are false.

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

As we have discussed earlier, instead of controlling the FWER $= P(V \geqslant 1)$ we are interested to control $P(T \geqslant 1)$, let us call it GFWER. We have two ways to proceed, one way is to swap our null and alternate hypotheses and then do usual testing and control FWER, but in that case, for each of the tests, finding the null distribution is difficult. The method we propose is to do the individual tests as usual but using a set of cutoffs which will help to control $P(T \geqslant 1)$. Intuitively, if we want to control the GFWER at by α, we have to do each individual test using a much higher level than α. This will eventually increase the number of false positives but here our main motive is to control the number of true negatives.

Let $p_{(1)} \leqslant p_{(2)} \leqslant ... \leqslant p_{(n)}$ denote the ordered p-values of the tests and $\alpha_1, ..., \alpha_n$ denote the levels of those tests respectively. Also, let $p_{(i):n_1}$ denote the p-value corresponding to the i^{th} false null hypothesis $\forall i = 1, 2, ..., n_1$ and α_{i,n_1} denote the respective levels of significance. We consider a step up method.

GFWER
$= P(T \geqslant 1)$
$\leqslant P(p_{(n_1):n_1} > \alpha_{n_1,n_1})$
$\leqslant P(p_{(n_1):n_1} > \alpha_{n_1})$
$=P($at least one p value corresponding to a false null hypothesis $> \alpha_{n_1}$ $)$
$\leqslant \sum_{i=1}^{n_1} P(U_i > \alpha_{n_1})$ where $U_i \sim$ Uniform(0,1)
$=n_1(1 - \alpha_{n_1})$

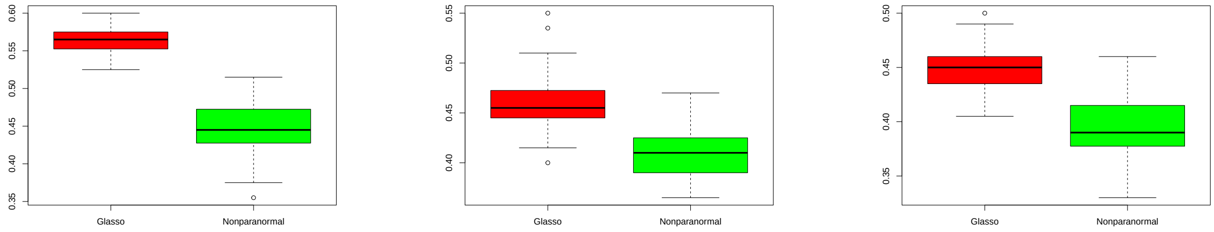
It is easy to see that if we use $\alpha_i = 1 - \frac{\alpha}{n}$ or $\alpha_i = 1 - \frac{\alpha}{i}$ $\forall i = 1, 2, ..., n$, then GFWER is bounded above by α. The first set of cutoffs is more conservative in nature i.e does not accept H_0 easily, it is analogous to the Bonferroni's method of controlling FWER, while the second set of cutoffs is a bit liberal than the previous one and is analogous to Holm's Method of controlling FWER.
Remark : In this setup, once we have achieved the error control as above, we will be interested to "discover" as many conditional independence relations as possible. Thus we can think $P($accept H_0 / H_0 is true) as our "power" of the tests here.

5 Simulation and Experiment Results

In this section we present some simulations of the theoritical methods that we have discussed so far and also applications on real data sets.

5.1 Comparison between nonparametric methods for continuous data

At first, we note that the Nonparanormal is a method which the authors "expect" to work well but it lacks strong mathematical justification. Thus, we should at least verify by simulation that it performs better than the graphical lasso when implemented on non gaussian data. We generate non gaussian data using some monotone functions mentioned in the paper by Lafferty, Liu, and Wasserman [LLW12]. To measure the goodness of performance of a method, we use "error rate" which is nothing but the proportion of incorrectly detected edges (presence of edge detected as absence or vice versa). We performed over 1000 simulations for each of the two methods and drawn the boxplots of the error rates generated, for a fixed dimension ($p = 20$). We give the boxplots side by side for 3 different sample sizes (Figure 1), from where it is clearly visible that the Nonparanormal performs better than the graphical lasso. Similar results are observed if we increase the dimension. Next we compare the two nonparametric methods. The forest density estimation puts a too much stringent restriction on the structure of the graph to be applied in practice. Thus, we compare between nonparanormal method and multiple testing using projected distance covariance. Here also we simulate from non gaussian data as previous and calculate the error rates and plot it for both the methods with increasing sample size.

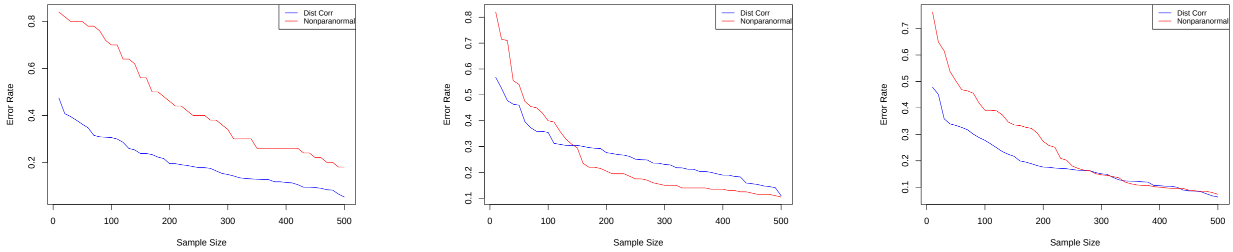


(a) $n = 200$

(b) $n = 500$

(c) $n = 1000$

Figure 1: Graphical Lasso vs Nonparanormal



(a) $p = 10$

(b) $p = 25$

(c) $p = 50$

Figure 2: Nonparanormal Method vs Distance Correlation Test Method

For lower dimensions (roughly up to p=15), the distance covariance test method is better than the nonparanormal but after that if we keep on increasing the dimension, there is no significant difference between the two methods. The performance of the testing method is much more stable than that of the nonparanormal, which performs slightly better or slightly worse than the multiple testing method depending on the data we generate.

5.2 Application on discrete data

5.2.1 Data Description

Now we consider a real data consisting of 4 variables. 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).* The data can be found here. It can be visualized as a multi way contingency table.

5.2.2 Question of interest

We want to find out which of the variables are conditionally independent given the rest. This can facilitate identifying important variable at one step when we have multiple potential response variables (namely muscle tension and weight of muscle in our case).

5.2.3 Graph structure estimation

Clearly, we want to estimate the structure of the graph which has the 4 variables mentioned above as its nodes. We use the methodology described in section 3.2.2 to find the optimum graph structure. Since we have 299 observations, which can be considered as ”large”, it is reasonable to apply the large sample approximation to the likelihood ratio test statistic. The adjacency matrix comes out to be

Variables	Muscle	Tension	Drug	Weight
Muscle	0	1	0	0
Tension	1	0	1	0
Drug	0	1	0	1
Weight	0	0	1	0

Hence, from this graph, we can make inferential statements like : change in muscle tension is independent of the weight of the muscle given the drug type and muscle type.

5.3 Application on continuous data

5.3.1 Data Description

The data is from a travel website. Details can be found here. For a particular individual, there are 10 random variables $X_1, X_2, ..., X_{10}$ denoting the average feedback given by the particular individuals for the i^{th} category. Categories are different places which people visit like picnic spots, restaurants, museums, art galleries, beaches, etc. Each rating is given in the scale of 0 to 4 where 0 denotes the minimum and 4 denotes the maximum level of satisfaction. We have data on 980 individuals. The data is taken from the machine learning repository of University of California, Irvine (Dua and Graff [DG17]).

5.3.2 Question of Interest

If we think from a business perspective, the conditional dependence structure of average rating given by people to different categories of places can help in improving suggestions to travellers by the website. Since the preference of a particular individual can be guessed from the average rating he/she gives to a particular type of place. We can make an idea about the general tendency of the public. Thus we try to construct an undirected graph with $X_1, ..., X_{10}$ as its nodes.

5.3.3 Graph Structure Estimation

We will use the method using projected distance correlation and use the Benjamini Hochberg Procedure for FDR control. Since we have a large number of observations (980), it is reasonable to use the large sample approximation of the distance correlation test statistic. The results are summarized in the adjacency matrix given below.

(The following short forms are used : 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)

CAT	AG	DC	JB	RST	MSM	RES	PP	BCH	THT	RI
AG	0	0	0	1	1	1	0	0	0	1
DC	0	0	1	1	0	1	1	1	0	0
JB	0	1	0	1	1	0	1	1	1	1
RST	1	1	1	0	1	1	0	1	1	0
MSM	1	0	1	1	0	1	1	0	0	0
RES	1	1	0	1	1	0	1	0	1	1
PP	0	1	1	0	1	0	0	1	1	1
BCH	0	1	1	1	0	0	1	0	1	1
THT	0	0	1	1	0	1	1	1	0	0
AG	1	0	1	0	0	1	1	1	0	0

Hence, from the above undirected graph, we can make statements like : the average rating of museum is conditionally dependent on the average rating of art gallery, juice bar, restaurant, resort and picnic spot given the others. Taking a cue from this information, we can check whether the dependence between the rating of museum is positive or negative with the others and hence decide whether to suggest a place or not to a particular individual who likes museums. Also, the rating given to a dance club is conditionally independent of the rating given to an art gallery or museum given the others. So, the undirected graph can serve as a first step in improving suggestions to visitors of the website.

****

References

[And03] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN: 9780471360919. URL: https://books.google.co.in/books?id=Cmm9QgAACAAJ.

[DG17] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[FFX19] Jianqing Fan, Yang Feng, and Lucy Xia. *A Projection Based Conditional Dependence Measure with Applications to High-dimensional Undirected Graphical Models*. 2019. arXiv: 1501.01617 [stat.ME].

[FHT07] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *Sparse inverse covariance estimation with the lasso*. 2007. DOI: 10.48550/ARXIV.0708.3517. URL: https://arxiv.org/abs/0708.3517.

[Guo+10] Jian Guo et al. "Joint Structure Estimation for Categorical Markov Networks". In: 2010.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. Springer, 2009. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009. ISBN: 9780262013192. URL: https://books.google.co.in/books?id=7dzpHCHzNQ4C.

[LLW12] John Lafferty, Han Liu, and Larry Wasserman. "Sparse Nonparametric Graphical Models". In: *Statistical Science* 27.4 (Nov. 2012). DOI: 10.1214/12-sts391. URL: https://doi.org/10.1214%2F12-sts391.

[Lau96] Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, 1996. ISBN: 0-19-852219-3.

[MB06] Nicolai Meinshausen and Peter Bühlmann. "High-dimensional graphs and variable selection with the Lasso". In: *The Annals of Statistics* 34.3 (June 2006). DOI: 10.1214/009053606000000281. URL: https://doi.org/10.1214%2F009053606000000281.

[RWL10] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. "High-dimensional Ising model selection using ℓ1-regularized logistic regression". In: *The Annals of Statistics* 38.3 (June 2010). DOI: 10.1214/09-aos691. URL: https://doi.org/10.1214/09-aos691.

[SRB07] Gá bor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. "Measuring and testing dependence by correlation of distances". In: *The Annals of Statistics* 35.6 (Dec. 2007). DOI: 10.1214/009053607000000505. URL: https://doi.org/10.1214%2F009053607000000505.

[VP13] Tom S. Verma and Judea Pearl. *On the Equivalence of Causal Models*. 2013. DOI: 10.48550/ARXIV.1304.1108. URL: https://arxiv.org/abs/1304.1108.

[Wan+15] Xueqin Wang et al. "Conditional Distance Correlation". In: *Journal of the American Statistical Association* 110.512 (Oct. 2015), pp. 1726–1734. DOI: 10.1080/01621459.2014.993081. URL: https://doi.org/10.1080/01621459.2014.993081.