Graphical models

Mateusz Małowiecki

November 11, 2021

Reminder: Graphs

- Consists of vertices and edges
- Vertices are adjacent if there is an edge between them
- Path - sequence of vertices $x_1, x_2, ..., x_n$, such that x_i and x_{i+1} are adjacent
- Complete graph - graph in which all pairs of different vertices are adjacent
- H is subgraph of G iff $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$
- Sparse graphs - graphs with relatively small number of edges

Graphs - convention

- Vertices represent random variables
- Edges mean random variables depedency
- We will talk only about directed graphs
- Edges in graph are parametrized by value

Challenges

- Model selection
- Estimation of the edge parameters from data
- Computation of marginal vertex probabilities and exceptations, from their jonit distribution

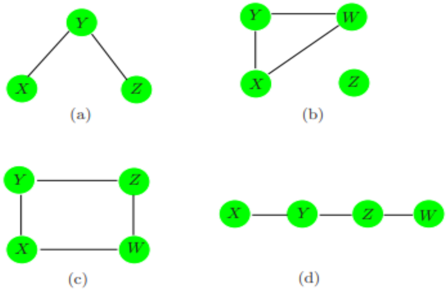
Challenges

- Model selection
- Estimation of the edge parameters from data
- Computation of marginal vertex probabilities and exceptations, from their jonit distribution

Markov graphs

Global properties

- No edge joining X and $Y \Leftrightarrow X \perp Y \mid rest$
- If A, B and C are subgraphs of G, and if every path between A and B intersects a node in C, then we say that C separate A and B
- If C separates A and B, then $A \perp B \mid C$
- We will separate the graph into cliques.



Density function

Probability density function over graph G, can be represented as:

$$f(x) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \tag{1}$$

where C is the set of maximal cliques and ψ_c are clique potentials and

$$Z = \sum_{x \in X} f(x) \tag{2}$$

Density function

Probability density function over graph G, can be represented as:

$$f(x) = \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \tag{3}$$

where C is the set of maximal cliques and ψ_c are clique potentials and

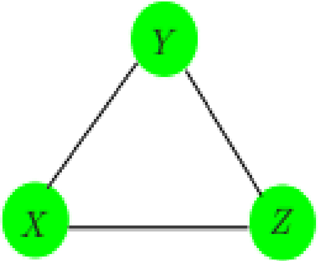
$$Z = \sum_{x \in X} f(x) \tag{4}$$

Dependence structure

- Consider three-node clique. It could represent the dependence structure of the distributions:

$$f_2(x, y, z) = \frac{1}{Z} * \psi(x, y) * \psi(x, z) * \psi(y, z) \tag{5}$$

$$f_3(x, y, z) = \frac{1}{Z} * \psi(x, y, z) \tag{6}$$

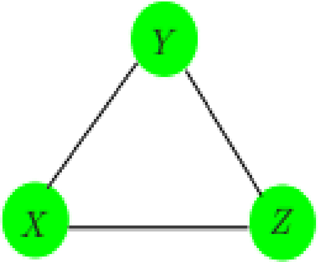


Undirected Graphical Models for Continuous Variables

- Consider three-node clique. It could represent the dependence structure of the distributions:

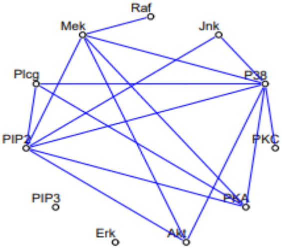
$$f_2(x,y,z) = \frac{1}{Z} * \psi(x,y) * \psi(x,z) * \psi(y,z) \tag{7}$$

$$f_3(x,y,z) = \frac{1}{Z} * \psi(x,y,z) \tag{8}$$



Undirected Graphical Models for Continuous Variables

- Markov graphs where all variables are continouos
- Multivariate Gaussian distribution



Undirected Graphical Models for Continuous Variables

Gaussian distribution properties

- All conditional distributions are also Gaussian
- The inverse covariance matrix($\theta = \Sigma^{-1}$) contains information about the partial covariances

Example

Estimation of the Parameters when the Graph Structure is Know

- Suppose that we have complete graph (clique)
- We have N multivariate normal realizations $x_i, i = 1, ..., N$ with population mean μ and covariance Σ. Let

$$S = \frac{1}{N} * \sum_{i=1}^{N} (x_i - \overline{x}) * (x_i - \overline{x})^T \tag{9}$$

  where $\overline{x}$ is the sample mean vector.
- The log-likelihood of data can be written as:

$$l(\Theta) = log(det(\Theta)) - trace(S * \Theta) \tag{10}$$

- $l(\Theta)$ is a convex function of Θ. Maximum likelihood estimate of Σ is simply S

Equality-constrained convex optimization problem

- We assume that graphs are not complete(some edges are missing)
- We would like to maximize (10) under the constraints that some pre-defined subset of the parameters are zero.
- A number of methods have been proposed for solving it
- We outline a simple alternate approach

Alternate approach

- Idea is based on linear regression
- We want to estimate values $\theta_{i,j}$ for given i
- We use model-based estimate of the cross-product matrix of the predictors when we perform our regressions

Alternate approach - details

- We add Lagrange constants for all missing edges:

$$l_C(\Theta) = log(det(\Theta)) - trace(S * \Theta) - \sum_{(j,k) \notin E} \gamma_{jk} * \theta_{jk} \quad (11)$$

- Gradient equation for maximizing (11) can be written as:

$$\Theta^{-1} - S - \Gamma = 0 \quad (12)$$

  where Γ is a matrix of Lagrange parameters.
- We will solve for Θ and its inverse $W = \Theta^{-1}$ one row and column at a time.

Alternate approach - details ctd.

- For simplicity let's focus on the last row and column. Then the upper right block of equation (12) can be written as

$$w_{12} - s_{12} - \gamma_{12} = 0 \tag{13}$$

- Let's say that matrices W and Θ are written as:
$\begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} * \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0^T & 1 \end{pmatrix}$
- We can see that:

$$W_{11} * \theta_{12} + w_{12} * \theta_{22} = 0 \iff w_{12} = \frac{-W_{11} * \theta_{12}}{\theta_{22}} \tag{14}$$

- Let's say $\beta = -\theta_{12}/\theta_{22}$

Alternate approach - details ctd.

- Let's say $\beta = -\theta_{12}/\theta_{22}$
- Then
$$W_{11} * \beta - s_{12} - \gamma_{12} = 0 \tag{15}$$
- These can be interpreted as the $p-1$ estimating equations for the regression of X_p on the other predictors.
- To solve (15) we will use subset regression.

Subset regression

- Suppose there are $p - q$ nonzero elements in γ_{12}
- By removing these $p - q$ elements (and also reducing β to β^* by removing its p - q zero elements. We get system of q equations:
$$W_{11}^* * \beta^* - s_{12}^* = 0 \tag{16}$$
with solution $\beta^* = (W_{11}^*)^{-1} * s_{12}^*$, which filled with zeros gives us solution β
- We can also calculate θ_{12}, because
$$\frac{1}{\theta_{22}} = w_{22} - w_{12}^T * \beta \tag{17}$$

Algorithm 17.1 *A Modified Regression Algorithm for Estimation of an Undirected Gaussian Graphical Model with Known Structure.*

1. Initialize $\mathbf{W} = \mathbf{S}$.

2. Repeat for $j = 1, 2, \ldots, p, 1, 2, \ldots, p, \ldots$ until convergence:

   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.

   (b) Solve $\mathbf{W}_{11}^{*}\beta^{*} - s_{12}^{*} = 0$ for the unconstrained edge parameters β^{*}, using the reduced system of equations as in (17.19). Obtain $\hat{\beta}$ by padding $\hat{\beta}^{*}$ with zeros in the appropriate positions.

   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$

3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = s_{22} - w_{12}^{T}\hat{\beta}$.

Estimation of the graph structure

- In most cases we do not know which edges to omit from our graph.
- We would like to discover this from the data itself.
- We will use Lasso regression.

Estimation of the graph structure ctd.

- Let's consider maximizing the penalized log-likelihood:

$$log(det(\Theta)) - trace(S * \Theta) - \lambda * ||\Theta||_1 \qquad (18)$$

- We can adapt the lasso to give the exact maximizer of the penalized log-likelihood

Estimation of the graph structure ctd.

- The analog of the gradient equation:

$$\Theta^{-1} - S - \lambda * Sign(\Theta) = 0 \tag{19}$$

- We assume that if $\theta_{jk} = 0$ then $Sign(\theta_{jk}) \in \{-1, 1\}$ and otherwise it is singum function.
- Next we can have analogue of (15):

$$W_{11} * \beta - s_{12} + \lambda * sign(\beta) = 0 \tag{20}$$

- This system is equivalent to the estimating equations for a lasso regression

Lasso recall

- Recall that with outcome variables y and predictor matrix Z, lasso minimizes

$$\frac{1}{2} * (y - Z * \beta)^T * (y - Z * \beta) + \lambda * ||\beta||_1 \tag{21}$$

- Gradient of this expression is:

$$Z^T * Z * \beta - Z^T * y + \lambda * Sign(\beta) = 0 \tag{22}$$

If we replace $Z^T * y$ with s_{12} and $Z^T * Z$ with W_{11} we have analog of (20)

Algorithm 17.2 *Graphical Lasso.*

1. Initialize $\mathbf{W} = \mathbf{S} + \lambda\mathbf{I}$. The diagonal of $\mathbf{W}$ remains unchanged in what follows.

2. Repeat for $j = 1, 2, \ldots p, 1, 2, \ldots p, \ldots$ until convergence:

   (a) Partition the matrix $\mathbf{W}$ into part 1: all but the jth row and column, and part 2: the jth row and column.

   (b) Solve the estimating equations $\mathbf{W}_{11}\beta - s_{12} + \lambda \cdot \mathrm{Sign}(\beta) = 0$ using the cyclical coordinate-descent algorithm (17.26) for the modified lasso.

   (c) Update $w_{12} = \mathbf{W}_{11}\hat{\beta}$

3. In the final cycle (for each j) solve for $\hat{\theta}_{12} = -\hat{\beta} \cdot \hat{\theta}_{22}$, with $1/\hat{\theta}_{22} = w_{22} - w_{12}^T\hat{\beta}$.

Coordinate descent method

- Let $V = W_{11}$. The update has form:

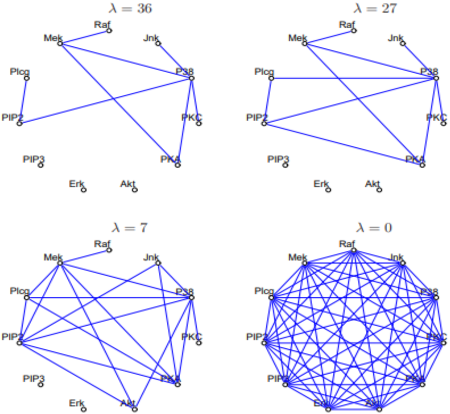
$$\beta_j \leftarrow \frac{S(s_{12j} - \sum_{k \neq j} V_{kj} * \beta_k, \lambda)}{V_{jj}} \tag{23}$$

where S is the soft-threshold operator

$$S(x, t) = sign(x) * (|x| - t)_+ \tag{24}$$

- The procedure cycles through the predictors until convergence.

Grpahical example



Undirected Graphical Models for Discrete Variables

Undirected Graphical Models for Discrete Variables

- Markov networks with all variables being discrete
- The most common are pairwise Markov networks with binary variables
- Sometimes called Ising models or Boltzmann machines

Undirected Graphical Models for Discrete Variables ctd.

- Let X_j be binary valued variable at node j. The Ising model for their joint probabilities:

$$P(X, \Theta) = exp\Big[\sum_{j \neq k} \theta_{jk} * X_j * X_k - \Phi(\Theta)\Big] \tag{25}$$

where $X \in \{0,1\}^p$ and $\Phi(\Theta)$ is the log of the partition function

$$\Phi(\Theta) = log \sum_{x \in \{0,1\}^p} \Big[exp\Big(\sum_{(j,k) \in E} \theta_{jk} * x_j * x_k\Big)\Big] \tag{26}$$

- The Ising model implies a logistic form for each node conditional on the others. Let's say that X_{-j} means all of the nodes except j.

$$Pr(X_j = 1 | X_{-j} = x_{-j}) = \frac{1}{1 + exp(-\theta_{j0} - \sum_{(j,k) \in E} \theta_{jk} * x_k)} \tag{27}$$

Estimation of the Parameters when the Graph Structure is Known

- Suppose we have observations $x_i \in \{0,1\}^p$
- The log-likelihood is

$$l(\theta) = \sum_{i=1}^{N} log(P(X_i = x_i, \Theta)) = \sum_{i=1}^{N}(\sum_{(j,k)\in E} \theta_{jk} * x_{ij} * x_{ik} - \Phi(\Theta)) \tag{28}$$

- Gradient of log-likelihood is:

$$\frac{\partial l(\Theta)}{\partial \theta_{jk}} = \sum_{i=1}^{N} x_{ij} * x_{ik} - N * \frac{\partial \Phi(\Theta)}{\partial \theta_{jk}} \tag{29}$$

where

$$\frac{\partial \Phi(\Theta)}{\partial \theta_{jk}} = \sum_{x\in\{0,1\}^p} x_j * x_k * p(x, \Theta) = E_\Theta(X_j X_k) \tag{30}$$

Estimation of the Parameters when the Graph Structure is Known ctd.

- If we set gradient to zero we will have

$$E(X_j X_k) - E_\Theta(X_j X_k) = 0 \tag{31}$$

Where

$$E(X_j X_k) = \frac{1}{N} * \sum_{i=1}^{N} x_{ij} * x_{ik} \tag{32}$$

- If p is not so big, we can use bunch of methods to solve it.

Poisson log-linear modeling

- We treat problem as regression problem.
- Vector y is the vector of 2^p counts in each of the possible distribution.
- Matrix Z has 2^p rows and up to $1 + p + p^2$ columns that characterize each of the distribution
- Cost is $O(p^4 * 2^p)$.

Poisson log-linear modeling

- We treat problem as regression problem.
- Vector y is the vector of 2^p counts in each of the possible distribution.
- Matrix Z has 2^p rows and up to $1 + p + p^2$ columns that characterize each of the distribution
- Cost is $O(p^4 * 2^p)$.

Estimation, when p is big

- Let's say $p > 30$. In this case we can approximate the gradient with methods:

Hidden Nodes

- We can make Markov's graph complexity better by including hidden nodes.
- Let X_H be the subset of hidden nodes and reminder X_V be the subset of visible nodes. Then the observed log-likelihood of the observed data is:

$$l(\Theta) = \sum_{i=1}^{N} log[Pr_{\Theta}(X_V = x_{iV})] =$$

$$\sum_{i=1}^{N} log\left[\sum_{x_H \in \chi_H} exp \sum_{(j,k) \in E} (\theta_{jk} * x_{ij} * x_{ik} - \Phi(\Theta))\right]$$

Hidden Nodes - ctd.

- The gradient is:

$$\frac{\partial l(\Theta)}{\partial \theta_{jk}} = E_V * E_\Theta(X_j X_k | X_v) - E_\Theta(X_j X_k) \tag{33}$$

- The value of the first term depends whether variables X_j and X_k are hidden or not. If both are visible E_V is mean of them. If one or both are hidden, they are first imputed given the visible data, and then averaged over the hidden variables.
- $E_\theta(X_j X_k | X_v)$ is given by formula:

$$E_\theta(X_j X_k | X_v = x_{iv}) = \begin{cases} x_{ij} * x_{ik} & \text{if } j, k \in \\ x_{ij} * Pr_\Theta(X_k = 1 | X_V = x_{iV}) & \text{if } j \in V, \\ Pr_\Theta(X_j = 1, X_k = 1 | X_V = x_{iV}) & \text{if } j, k \in \end{cases}$$

Estimation of the Graph Structure

- An approximate solution, analogous to the graphical Lasso for continued variables.
- We want to fit an L1-penalized logistic regression model to each node as a function of the other nodes, and then symmetrize the edge parameter estimates in some fashion
- The key difference between estimation of the discrete and continuous models is that in the continuous case, both Θ and its inverse will often be of interest, while discrete case only yields Θ

Restricted Boltzmann Machines

TODO: Think if this slide is neccessary