Machine Learning for Finance (FIN 570)
Graphical Model and Covariance Estimation

Instructor: Jaehyuk Choi

Peking University HSBC Business School, Shenzhen, China

2021-22 Module 1 (Fall 2021)

Background

- In the world of big data, variables (K) and relations (K^2) are exploding.
- We need models and intuitions to simplify and breakdown into smaller pieces.

Independence and product rules

Two probability events a and b are **independent**: the probability $p(a)$ is not influenced by outcome of event b (and vice-versa)

$$p(a|b) = p(a) \quad \text{and} \quad p(b|a) = p(b)$$

Therefore the joint probability is broken down into products:

$$p(a, b) = p(a|b)p(b) = p(b|a)p(a) = p(a)p(b)$$

For (fully connected) three random variables, a, b, c, the joint distribution is decomposed to

$$p(a, b, c) = p(a|b, c)p(b, c) = p(a|b, c)p(b|c)p(c)$$

For $\boldsymbol{x} = (x_1, \cdots, x_K)$ in general,

$$p(\boldsymbol{x}) = p(x_1|x_2 \cdots x_K)\, p(x_2|x_3 \cdots x_K) \cdots p(x_K)$$

Conditional Independence

Now, events a and b are **conditionally independent** (Wiki) on c when a and b are independent given c happens (or not):

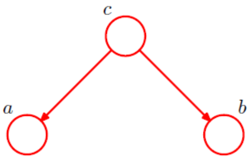
$$p(a|b,c) = p(a|c), \quad p(b|a,c) = p(b|c) \;\Rightarrow\; p(a,b|c) = p(a|c)p(b|c)$$

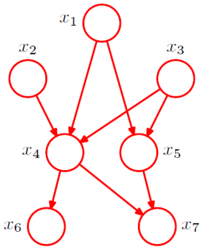
Example

If two people live in the same city, the probabilities that two people get home in time for dinner (event A and B respectively) are not independent. Assume that the only reason is the traffic condition: if a snow storm hits the city (event C) and traffic will be at a stand still, the events A and B become highly correlated. If a traffic condition is given (either good or bad), the events are independent conditionally on C. (Stackexchange)

Graphical Model

- Node (vertice) representing a random variable.
- Edge (link) between nodes represents probabilistic relation.
- GM can visualize the structure of variables in terms of conditional independence
- In **directed** GM (Bayesian belief network), relation is directional (arrow), suitable for model causal relation/chronic events.
- In directed acyclic graph (**DAG**), node c is is a parent of a if a depends on (or have influence from) c.





Product rule under DAG

In **DAG**, two nodes are independent on the rest if they are not in an ancestor-descendant relation.

Case 1: a and b are independent conditional on c

$$p(a, b, c) = p(a|b, c)p(b, c) = p(a|c)p(b|c)p(c)$$

Case 2:

$$p(\boldsymbol{x}) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5)$$

General case: If $\boldsymbol{x} = x_1, \cdots, x_K$ and $\boldsymbol{pa}_k$ is the parent nodes of x_k,

$$p(\boldsymbol{x}) = \prod_{k=1}^{K} p(x_k|\boldsymbol{pa}_k)$$

Linear model under DAG

Multivariate linear regression

$$y = \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon \quad \text{where} \quad \boldsymbol{\beta} = \Sigma_{\boldsymbol{x},\boldsymbol{x}}^{-1} \Sigma_{\boldsymbol{x},y} \quad (E(\boldsymbol{x}) = \mathbf{0})$$

- Σ is the covariance matrix (including y and $\boldsymbol{x}$).
- The coefficients $\boldsymbol{\beta}$ are determined such that the error ε is minimized in MSE.
- ε is not correlated with any $x_k \in \boldsymbol{x}$.

Regression in DAG

For $x_i \in \boldsymbol{x}$ and its parents $\boldsymbol{pa}_i$,

$$x_i = \boldsymbol{\beta}_i^T \boldsymbol{pa}_i + \varepsilon_i \quad \text{where} \quad \boldsymbol{\beta} = \Sigma_{\boldsymbol{pa}_i,\boldsymbol{pa}_i}^{-1} \Sigma_{\boldsymbol{pa}_i,i}$$

- x_i is *explained* by its parents.
- ε_i is correlated to neither $\boldsymbol{pa}_i$ nor any other component of $\boldsymbol{x}$.

Covariance estimation

Covariance estimation (Wiki) is essential in multivariate analysis, finance in particular.

- Multi-variate linear regression requires covariance. (Wiki)
- Mean-variance portfolio optimization. (Wiki)

However, there are challenges in estimating large covariance

- Sample covar is often not positive-definite, thus cannot be inverted. There are many tricks to avoid it (e.g. make the eigenvalues positive)
- Often, sample covar (one var correlated to all the rest) is not desirable.
- Example: hedging off-the-run treasury (e.g., 7y) bonds with on-the-run treasury bonds (e.g., 1y, 2y, 5y, 10y).

Covar Estimation in DAG

We express the error equation

$$\varepsilon_i = x_i - \boldsymbol{\beta}_i^T \boldsymbol{pa}_i \quad \text{or} \quad \boldsymbol{\varepsilon} = A\boldsymbol{x}, \quad \text{where}$$

$$A_{i,j} = \begin{cases} 1 & \text{if} \quad j = i \\ -(\Sigma^{-1}_{\boldsymbol{pa}_i,\boldsymbol{pa}_i} \Sigma_{\boldsymbol{pa}_i,i})_j & \text{if} \quad j \in \boldsymbol{pa}_i \\ 0 & \text{otherwise.} \end{cases}$$

Since we know $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ and

$$\mathrm{Var}(\varepsilon_i) = \Sigma_{i,i} - \Sigma_{i,\boldsymbol{pa}_i} \Sigma^{-1}_{\boldsymbol{pa}_i,\boldsymbol{pa}_i} \Sigma_{\boldsymbol{pa}_i,i},$$

the estimated covariance Σ' is solved as

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{diag}(\mathrm{Var}(\boldsymbol{\varepsilon})) = A\Sigma' A^T$$

$$\Sigma' = A^{-1}\mathrm{Cov}(\boldsymbol{\varepsilon})(A^{-1})^T$$

Python Demo

More ...

Further readings

- Bishop (PR&ML) Ch. 8
- High dimensional sparse covariance estimation via directed acyclic graphs by (Philipp Rutimann and Peter Buhlmann, 2009)

Further questions

- **Undirected** graph instead of directed graph?
- Given full covariance matrix, how to **estimate** graph which minimized error? (Friedman, Hastie, Tibshirani, 2007), (sklearn package)