Structure Estimation in Gaussian DAG

Debarshi Chakraborty

debchak@iastate.edu

Department of Statistics
Iowa State University

October 24, 2023

Topics to be covered

- Brief introduction to Gaussian Graphical Models.
- Problem of structure estimation.
- Connection between methods of undirected graphical models and DAGs.
- Discussion of two basic methods .
- What can be done in high dimensional setup?
- Further scope of study.

Why the Gaussian obsession?

- Statistician's favourite for quite a few obvious reasons.
- What properties will help us?
- Independence $\iff$ No correlation.
- Conditional Independence $\iff$ No partial correlation.
- Conditional distributions are also gaussian.
- $E(X_1|X_2, ...X_p)$ gives exactly linear regression equation.
- Whole conditional dependence structure is encoded in a model parameter itself, namely the precision matrix denoted by $\Theta = \Sigma^{-1}$.
- ...and many more, basically the theory is easy to develop for this particular case, then generalise it slowly.

Gaussian Directed Acyclic Graphs

- **Directed Acyclic Graph (DAG)** : A graph $G = (V, E)$ where all edges are directed and it does not contain any cycle.
- **Linear Gaussian Model** : Let Y be a continuous random variable in a DAG with parents $X_1, X_2, ...X_k$. We say that Y has a linear Gaussian model of its parents if there exists $\beta_0, \beta_1, ..., \beta_k$ such that

  $$Y = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \epsilon$$

  where $\epsilon \sim N(0, \sigma^2)$.
- **Gaussian DAG** : A DAG where every node represents a continuous R.V and each node follows a linear gaussian model of its parents.
- Under some minimal conditions , 1-1 correspondence between a Gaussian DAG and a Multivariate Normal Distribution can be proved.

Available data and Estimation problem

- **Data** is available in the form of an $n \times p$ data matrix containing n observations on each of the p variables.

$$\begin{pmatrix} X_{11} & X_{12} & ... & ... & X_{1p} \\ X_{21} & X_{22} & ... & ... & X_{2p} \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \\ X_{n1} & X_{n2} & ... & ... & X_{np} \end{pmatrix}$$

- **Goal** is to estimate the structure of the graph i.e. which edges are present / absent and also the directions (maybe) ?

Motivation of methods from undirected graphical models

- Easy to understand since in undirected graphs absence of an edge between X_i and $X_j \iff X_i \perp\!\!\!\perp X_j |$ others.
- Moreover, if the nodes are multivariate normal, then we have

$$\Sigma^{-1}_{ij} = 0 \iff X_i \perp\!\!\!\perp X_j | \text{ others}$$

- So, entire problem boils down to estimating Σ^{-1}.
- Let us try to solve this simplified problem in different situations.
- Then use these methods as a basis to come up with methods for estimating the structure of gaussian DAGs.

Situation 1 : $n >> p$

- Compute $\hat{\Sigma}$ by maximum likelihood estimation.
- Get $\hat{\Theta}$ by invariance property of MLE.

Situation 2 : $n > p$ or $p > n$

- Maximum likelihood estimates are no longer reliable.
- Carry out **multiple testing**.
- For multivariate gaussian, theoritically nice tests are available.
- We can just use **partial correlations** to carry out the tests.
- Issues ? Maybe computationally expensive.

How to use in DAGs?

- A well known method for structure estimation in DAGs is the **PC Algorithm**.
- It operates in two steps : (i) Identifying the skeleton and (ii) Identifying the complete partially DAG (CPDAG).
- For step (i), multiple testing is used.
- Let $A \subset V \cap \{X_i, X_j\}^c$, goal is to find A such that $X_i \perp\!\!\!\perp X_j | A$.
- We carry out tests $H_0 : X_i \perp\!\!\!\perp X_j | A$ for $|A| = 1, 2, ..., p - 2$ till the null hypothesis is accepted.
- In this way, we find the "separating sets"for each such pair of nodes, denote that by $\mathcal{A}_{ij}$.
- In step 2, we proceed to identify the CPDAG using some constraints and semantics used in DAG literature
- Example for every pair of non adjacent nodes 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 $X_i \rightarrow X_k \leftarrow X_j$ if $X_k \notin \mathcal{A}_{ij}$.

Situation 3 : $p >> n$

- A seminal work by Friedman, Hastie and Tibshirani : **graphical lasso**.
- Understand only the main idea.
- Note that, we only need to discover whether $\theta_{ij} = 0$ or not.
- Regress X_i on all others, this is exact linear regression equation.
- We have $\beta_j = 0 \iff \theta_{ij} = 0$, exploit this relation.
- Impose an l_1 penalty on the coefficients, some β's will be automatically shrinked to zero
- Update both Σ and Θ, repeat the process until convergence
- For detailed algorithm, see the paper
- Actually the idea is much simpler and intuitive to implement in DAGs

How to use in DAGs?

- **Suppose,** the ordering of $X_1, ..., X_p$ is given.
- For any X_i, we know the set of its potential parents say $B_i = \{X_{i1}, X_{i2}, ..., X_{ij}\}$.
- Regress X_i on the variables in B_i.
- Instead of ordinary linear regression, fit a lasso regression.
- The variables affecting X_i and consequently the directed edges will be selected automatically.
- **Limitation** : the first line of this slide is often a too good assumption to have.

How to use in DAGs?

- A two stage adaptive lasso procedure has been developed in literature.
- Here it is not assumed that the ordering is known.
- usual linear gaussian notation : $X_i = \sum_{j \in pa(X_i)} \rho_{ij} X_j + Z_i$.
- Matrix notation : $X = \Lambda Z$.
- Let A denote the adjacency matrix of $X_1, X_2, ..., X_p$ such that a_{ij} = partial covariance between X_i and X_j given all other variables.
- WLOG assume all variables are centered and scaled
- $X = AX + Z$.
- Hence $\Lambda = (I - A)^{-1}$
- Ordering of variables is encoded in Λ, that needs to be estimated.
- In step 1, neighbourhood selection is done by the approach of Meinshausen and Buhlmann.
- Step 2 involves an optimization problem with l_1 penalties on a_{ij}'s.

Further studies

- The second step in the adaptive lasso algorithm is basically a variable selection problem of regression.
- It is not necessary to use lasso, any variable selection method (maybe knockoffs?) can be used.
- Methods like PC algorithm in nonparametric setup?
- Need nonparametric tests for independence (actually conditional independence)
- Distance Correlation, Chatterjee Correlation (maybe?)
- Computational cost?? Let's not talk about it at the moment!!!!

References

1. Introductory Read to DAGs
2. Graphical Lasso Paper
3. Penalized Likelihood Method for DAGs Paper
4. Two Stage Adaptive Lasso Paper

Conclusion

THANK YOU

QUESTIONS?