NONPARAMETRIC GRAPHICAL MODELS — PNI Meeting

LECTURE #1 — OCTOBER 2, 2015

LECTURER: HAN LIU — SCRIBE: KIRAN VODRAHALLI

Contents

1 Introduction

My name is Han Liu, I am an assistant professor in ORFE. This is joint work with Ken Norman and Jeremy, Xiaoyan, and Angela. A nonparametric graphical model is a very powerful tool for analyzing complex data; these tools can be published in highly theoretical journals (statistics), but on the other spectrum, you can publish this in Nature, or Quantitative Finance. It is simple, theoretically interesting, and interesting to applications. It has been successfully applied on resting state fMRI data. We are trying to extend this approach to task-specific fMRI.

2 Graph Estimation Problem

This line of work is about infer conditional independence based on observational data. It is a very simple problem: n samples, d variables (n scans, d voxels, or regions, whatever you like). Given this data, we want to estimate an undirected graph $G = (V, E)$. Each node corresponds to a random variable, a column of this semantic matrix. If there is no edge, it just means that $X_j \perp X_k |$the rest. We hope that the graph has certain meanings for the underlying distributions. I call this high dimensional topological data analysis. First, it is different because it is high dimensional. We are talking about $50,000$ dimensions. The other thing, from topological data analysis perspective, we are analyzing the conditionally independence property of the high-dimensional distribution. This is a much more challenging setting. We call this graph Markov random field. If these two nodes have an edge, it just means they interact with each other. Given this simple setting of the graphical model, graph-estimation problem, why do we care about this?

Suppose you have a movie that a human is watching. Across time from $t = 1, \cdots, T$, with T large.

So we want to understand this time course. At each time point, we want to estimate a graph, which changes with time. Thus we want to look at the evolution of this graph, which is very high dimensional, we want to cluster this graph to partition the brain into different regions, and study the network in these regions. So this is the application for neuroscience. Graphical models have been applied to lots of application domains in finance and genomics. This is also a fundamental problem in statistics - how do we do high-dimensional dense estimation. From the computational side, we also have a lot of issues.

From the statistical perspective, we have two goals:

1. We want to estimate graph $\hat{G}$ with certain properties, namely that $\lim_{n\to\infty} \mathbf{P}\{\hat{G} = G\} = 1$ (point estimation).

2. Testing problem. After we estimate the graph, we want to test the H_0: How can we be sure an edge exists in the graph, so we need to deliver p-value in high dimensions. We also want to know how many connected components, and test hypotheses, etc. (How does this relate to planed bisection?)

3 What has been done?

Gaussian graphical models (parametric models) - you just assume a joint multivariate Gaussian distribution. Under the Gaussian assumption, We have the interesting result that $X_j \perp X_k |$rest is completely encoded by whether $\Sigma^{-1}_{jk} = 0$. This implication is if and only if for the precision matrix.

So Glasso (Yuan and Lin 2006) propose a simple statistic (note that the precision matrix is a sufficient statistic) that uses the sample covariance. Nodewise regression, graphical Dantzig, etc.

We are particularly interesting in $SCIO$(Liu and Luo 2012). Instea dof estimating the whole precision matrix at once, it only estimates the j^{th} column and solves $\min_\beta \frac{1}{2}\beta^T\hat{\Sigma}\beta - e_j^T\beta + \lambda\|\beta\|_1$. So we have a quadratic form in sample covariance and with the j^{th} column; thus β serves as a surrogate for the j^{th} column. Regularization is just to ensure sparsity on the j^{th} column. Each β encodes the neighborhood, so this is sparsity over neighbors (neighborhood selection). It scales with nodes, not edges. Right now it scales to around 10000 nodes. The theory for Gaussian models for this estimator is very well-established. But if you see the Q-Q plot, you often see that the Gaussian assumption is wrong. The data is not Gaussian! So what should we do?

4 Our Goals

We want to relax the Gaussian assumption. IF we want non-Gaussian, we need nonparametric parameters to make the model more flexible. We also don't want to lose our statistical and computational effiencies. We also want to be economical on computation time.

We want to introduce nonparanormal graphical model (non-parametric Gaussian), semiparametric exponeential family graphical model, forest structured graphical model. All of these are legitimate graphical models, and are richer than Gaussian. They are flexible, but don't lose statistical and computational efficiency.

5 Nonparanormal Model

Non-paranormal model. We want to extend Gaussian to Gaussian copula family. A random vector X is nonparanormal if there exists a sum of marginal strictly increasing transformations, such that after the transformation, the data is Gaussian ($f_j(t) = \frac{1}{\sigma_j}(t - \mu_j)$). Instead of assuming directly normal, we believe there is a translation such that afterwards, it is normal. So this lets you recover an arbitrary Gaussian distribution

- these are infinite dimensional parameters; we only know they must be strictly increasing. Thus this is nonparametric. This is just bigger than the Gaussian family; Gaussian is a special case. Marginal transformations can give you a lot of interesting looking distributions with multiple modes; here, you can have arbitrary number of nodes. Even though this is a marginal transformation, it is much richer. But why do we want to define this family in this way? It is due to an interesting property. We can use a Jacobian transformation to direclty write down the density of the nonparanormal distribution. There is a theorem implying if your density has a Gaussian form, you get the result that $X_j \perp X_k|\text{rest}$ iff $\Sigma^{-1}_{jk} = 0$. However, this is not jointly convex, so how do we estimate the parameters? Assuming the data actually follows a nonparanormal distribution, then how should we fit this model? We use the same estimator as before to estimate the j^{th} column. There is one difference: Here instead of $\hat{\Sigma}$, we use the Kendell-Tau correlation matrix? We can use $\hat{\Sigma}_{jk} = \sin(\frac{\pi}{2}\hat{\tau}_{jk})$. You augment Kendell-Tau by calculating the differences between the priors: We only care about the relative order of the differences. Thus we have $\hat{\tau}_{jk} = (2/(n(n-1))) \sum_{i<i'} \text{sgn}(x_{ij} - x_{i'j})\text{sgn}(x_{ik} - x_{i'k})$. This is exactly rank correlation. We are not conditioning on choice of f yet. So here, we use the same Gaussian graphical model solver, but here, we need to just change the estimator. By Kruskal 1948, he shows that the Pearson correlation coefficient $\Sigma_{jk} = \sin(\frac{\pi}{2}\mathbf{E}[\hat{\tau}_{jk}])$, where the $\hat{\tau}$ is the estimated Kendall's tau. This results from integrating over polar integration. So you go from a rank estimate to a Pearson's correlation coefficient with a sin transform. The relative order matters though; however, relative order is invariant to monotone increasing distributions; thus, this holds for the nonparanormal family. Is there an efficiency loss? One is in terms of graphical recovery: Here you lose almost nothing. How about the limiting variance for $\hat{\Sigma}_{jk}$? We only get 65%, but this ends up not mattering, which I will show later. So the main point: We advocate use **rank covariance** matrix instead of plain correlation matrix, but do some transformations first (to get you the theory). This is also highly robust (very simple monotone condition holds when there's noise with high probability).

However note that our optimization is non-convex. We still do work to find local solution that has good statistical significance. We propose PICASA (80-page long paper), it is very long because the analysis is very long. This algorithm is pretty simple, it is basically the same as pathwise coordinate optimization. This algorithm works $\lambda_0 > \cdots > \lambda_n$, this is similar to interior point methods. Then you do a bunch of coordinate updates, do a self loop. Then do we need to update the active set? We check approximate KKT conditions. If we satisfy, we output a $\hat{\beta}_\lambda$, and use it to initialize the next round. PICASA uses proximal gradient pilot, which is a simple algorithm, but has a lot of impliciations in theory.

5.1 PICASA

Theorem 5.1. *PICASA.*
Let X be nonparanormal($\Sigma, \{f_j\}_{j=1}^d$). Then we have geometric iteration complexity is $c \log 1/\epsilon$, statistical rate of convergence is $\|\hat{\Theta} - \Theta\|_2 \leq \sqrt{(\log d)/n}$, thus it is minimax optimal, finally we have $\hat{G} = G$, so we have consistency.

Proof. To prove this theorem, you have to cleverly use local geometry and check approximate KKT conditions. □

As for empirical results, then if the data is nonGaussian, nonparanormal does not do worse than Gaussian.

Now if the distribution is truly Gaussian, what do we use by using nonparanormal. The answer is almost nothing. If we look at the ROC for graph recovery, we see that the Gaussian and nonparanormal line up almost exactly. As the problem becomes more challenging, both methods become much worse, and get closer together. If it's too easy problem, then we

5.2 The Main Point

Don't use sample covariance, use the **rank covariance** matrix, but use it carefully: transform it, and then use nonconvex optimization (our algorithm, not convex optimization) since this problem is non-convex.

6 Semiparametric exponential family model

Each conditional distribution follows a generalized linear model (GLM). If the j^{th} column follows the generalized distribution, we are saying these are how the nodes are distributed. We add on a f_j, as a base measure. If we can show that $f_j(t) = -t^2/2$, then we have that the $p(x_j|x_{-j})$ is a Gaussian model. Thus Gaussian model is a subcase. If $f_j(t) = 0$, then it is Bernoulli. Then $f_j(t) = -log(t!)$ results in a Poisson. Thus we can model mixed data and all kinds of data, thus this is also a powerful model family. Why is this interesting in graphical model setting? If $\beta_{jk} = 0$, this suggests that $X_j \perp X_k|\text{rest}$ iff $\beta_{jk} = 0$, where β_{jk} are the coefficients of the combinations of $x_j x_k$ in the exponent. So we just need to fit this model. We can do conditional log-likelihood maximiziation, with a sparse penalty. The properties of this estimator have been studied a lot in Chen et. al(2014), Yang et al (2014). Challenge: WE NEED TO KNOW THE BASE MEASURE for the model. So this amounts to model selection. Can we just treat the base measure as a nuisance parameters without losing statistical and computational efficiencies?

Now I will give you the SPARC procedure. We have n samples and d variables. We take the augmented data by created $Z_{12} = X_1 - X_2$ for all $X_1, \cdots, X_n$ where $X_i \in \mathbb{R}^d$. Thus we get $\binom{n}{2}$ extra data. Then we do a logistic loss using the Zs as variables and an additional sparsity constraint. Thus we have a penalized logistic loss ($\hat{\beta}_j = \text{argmin}\binom{n}{2}^{-1}\sum_{i<i'}\log(1+\exp(-Z_{ii',j}\beta_j^T Z_{ii',-j})$. This is the whole procedure. Thus what we want to truly focus on is implementing a really good logistic operation.

What is the intuition? This operation is like statistical chromatography. We are modeling the data at a more refined granularity. The main intuition is that if we got any sequence of data $\{X_{1j}, \cdots, X_{nj}\}$ is a rank statistic (R) + an order statistics (O) (small to large). Traditionally, people model the conditional distribution of the j^{th} distribution. Then we are more doing something like a discriminative model. What happens is we calculate the first part, we basically get a softmax (and does not depend on the base transformation). Here we use a property of the exponential family distribution, to get f_js to cancel. So this is called partial likelihood. From the statistical view, we will throw away information and thus get a larger variance, but that's ok: It's not too bad. There are a lot of interesting computational problems. We can take a lower order approximation of the conditional rank likelihood.

We have a theorem:

Theorem 6.1. *Let β_j^* be the true parameter. Then*

$$\boldsymbol{E}[\nabla l(\beta_j^*)] = 0$$

As a practical implementation, we can use a sparse logistic regression solver.

We get the same no-free-lunch theorem as for the nonparanormal setting. Again, we use the same ideas to prove this (local geometry, eigenvalues, etc). It says that in terms of convergence we don't lose anything.

If a parametric model is incorrect, then SPARK is better than the parametric model. If the parametric model is correct, then SPARC is worse. So here, we do have an efficiency loss (it is larger than for nonparanormal). We gain a lot in computation, but lose. What is reflected in this picture is that this model family is maybe bigger than the nonparanormal family, and we pay for this in our increased loss.

We are trying to avoid doing things that are not necessary (like estimating the base measure, not computationally efficient, since our goal is the graph. But in theory we could estimate the base measure).

7 Forest-structured Graphical Model

What is the point of the third one: Well, what if the true distribution is not nonparanormal or sparse exponential family graphical models? Here, we want to trade off structural flexibility for greater nonparametricness. We want to extend the Gaussian copula to fully nonparametric. We keep relaxing the distributional family. Here we need to regularize the underlying graph. Here we restrict to $F = (V, E)$ is an acyclic graph. Then a distribution supported on a forest allows us to factorize $F = (V, E)$. Thus $p_F(x) = \prod_{j,k \in E} p(x_j, x_k)/(p(x_j)p(x_k)) \prod_{l \in V} p(x_l)$.

To estimate these forest graphical models, we take the KL divergence between the true distribution and the forest distribution (we want to project the true density). These are subject to some constraints, namely that the number of edges in the forest is no larger than k (these are regularizations on our estimator). This problem is equivalent to maximum weight forest problem (Kruskal 1956). Here the graph edges are pairwise mutual information. So how shall we implement this, we only have data? To calculate mutual information, we can use our estimators of $\hat{p}(x_j, x_k)$.

7.1 Forest Density Estimation

We basically just use Kruskal's greedy algorithm to do this: We have a graph with empirical mutual information edges. We sort edges by empirical mutual information. Then greedily add edges such that no cycles are formed. Then we output the graph after k edges are in.

Thus, after we calculate the pairwise mutual information matrix (despite the high dimension), we only need to run this greedily algorithm, and then we are good. Everything is minimax optimal.

What we want to show is that this method is simple, but the theory is deep. Using concentration and stuff, you can bound things in terms of theory and show things are optimal.

8 Conclusion

Let us have a unifed view of these three methods. All of them belong to an family called a second order nonparametric Markov network.

$$p(x) = \frac{1}{Z} \exp(\sum_{j=1}^{d} \psi_j(x_j) + \sum_{k<l} \psi_{kl}(x_k, x_l))$$

This is not highly specified, so our three models provide the specifications. Nonparanormal says $\psi_{kl}(x_k, x_l) = \Omega_{kl} \cdot f_k(x_k) f_l(x_l)$. Forest density says only involved $d - 1$ directions (no cycles) over interaction terms $\psi_{kl}(x_k, x_l)$. Then exponential family graph gives us $\psi_{kl} = \beta_{kl} x_k x_l$. Thus this is the tradeoff between structual complexity and flexibility.

Look at Eloyan et. al 2012, Qiu et al. 2015 for the applications of graphical models to neuroscience.

This is in fact just a starting group. Our own group has done a ton of extensions: extend nonparanormal to more complex, time varying, multiple nonparanormal graphs, matrix-variate nonparanormal graphs, discrete nonparanormal graphs, differential nonparanormal nework analysis.

Motivate new empirical process and random matrix theory.

This is a family of methodologies, and it is pretty active in statistics for NIPS, ICML, JMLR, etc. So it is the right time to do something.