Gaussian Copula Graphic Model on Time Varying Count Data

Hao Pan 1401110056, Yimin Huang 1401110053 and Yu Zhang 1501110060
School of Mathematical Sciences, Peking University

Introduction

Recent methods for estimating sparse undirected graphs for real-valued data in high dimensional problems rely heavily on the assumption of normality. Liu et al.(2009) shows how to use a semiparametric Gaussian copula–or "nonparanormal"–for high dimensional inference. Just as additive models extend linear models by replacing linear functions with a set of one-dimensional smooth functions, the nonparanormal extends the normal by transforming the variables by smooth functions. Liu et al. derive a method for estimating the nonparanormal, study the methods theoretical properties, and show that it works well in many examples.

Furthermore, Liu et al.(2012) proposes a semiparametric approach called the nonparanormal skeptic for efficiently and robustly estimating high dimensional undirected graphical models. To achieve estimation robustness, they exploit nonparametric rank-based correlation coefficient estimators, including the Spearman's rho and Kendall's tau.

In our work, we want to use their methods in the real count data. Before that, we make a simulation to compare these results by ourselves. In next section, we will introduce the methods we used first. Then we will show our simulation results, and finally, we will use these methods in our real data.

Estimating Undirected Graphs

Let $X = (X_1, ..., X_p)$ denote a random vector with distribution $P = N(\mu, \Sigma)$. The undirected graph $G = (V, E)$ corresponding to P consists of a vertex set V and an edge set E. The set V has p elements, one for each component of X. The edge set E consists of ordered pairs (i, j) where $(i, j) \in E$ if there is a edge between X_i and X_j . The edge between (i, j) is excluded from E if and only if X i is independent of X_j given the other variables $O_{\setminus\{i,j\}} \equiv (X_s : 1 \le s \le p, s \ne i, j)$, written

$$X_i \amalg X_j | O_{\setminus\{i,j\}} \tag{1}$$

It is well-known that, for multivariate Gaussian distributions, (2.1) holds if and only if $\Omega_{ij} = 0$ where $\Omega = \Sigma^{-1}$. Let $X^{(1)}, X^{(2)}, ..., X^{(n)}$ be a random sample from P, where $X^{(i)} \in R^p$. If n is much larger than p, then we can estimate Σ using maximum likelihood, leading to the estimate $\hat{\Omega} = S^{-1}$, where

$$S = \frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \overline{X})(X^{(i)} - \overline{X})^T \tag{2}$$

is the sample covariance, with $\overline{X}$ the sample mean. The zeroes of Ω can then be estimated by applying hypothesis testing to $\hat{\Omega}$ (Drton and Perlman, 2007, 2008).

When $p > n$, maximum likelihood is no longer useful; in particular, the estimate $\hat{\Sigma}$ is not positive definite, having rank no greater than n.

Estimation of Σ Let $X^{(1)}, \cdots, X^{(n)}$ be a sample of size n where $X^{(i)} = (X_1^{(i)}, \cdots, X_p^{(i)})^T \in R^p$.

Instead of the marginal empirical distribution function $\hat{F}_j$, we use the following truncated or Winsorized estimator:

$$\tilde{F}_j(x) = \begin{cases} \delta_n, & \text{if } \hat{F}_j(x) < \delta_n; \\ \hat{F}_j(x), & \text{if } \delta_n \le \hat{F}_j(x) \le 1 - \delta_n; \\ (1 - \delta_n), & \text{if } \hat{F}_j(x) > 1 - \delta_n. \end{cases} \tag{3}$$

where δ_n is a truncation parameter, and $\hat{F}_j(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_j^{(i)} \le t\}}$.

Clearly, there is a bias-variance tradeoff in choosing δ_n. In what follows we use

$$\delta_n \equiv \frac{1}{4n^{1/4}\sqrt{\pi \log n}} \tag{4}$$

Given this estimate of the distribution of variable X_j, we then estimate the transformation function f_y by

$$\tilde{f}_j(x) \equiv \hat{\mu}_j + \hat{\sigma}_j \tilde{h}_j(x) \tag{5}$$

where

$$\tilde{h}_j(x) = \Phi^{-1}(\tilde{F}_j(x)) \tag{6}$$

and $\hat{\mu}_j$ and $\hat{\sigma}_j$ are the sample mean and the standard deviation:

$$\hat{\mu}_j \equiv \frac{1}{n}\sum_{i=1}^{n} X_j^{(i)} \text{ and } \hat{\sigma}_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_j^{(i)} - \hat{\mu}_j)^2} \tag{7}$$

Now, let $S_n(\tilde{f})$ be the sample covariance matrix of $\tilde{f}(X^{(1)}), \cdots, \tilde{f}(X^{(n)})$; that is,

$$\begin{aligned} S_n(\tilde{f}) &\equiv \frac{1}{n}\sum_{i=1}^{n}(\tilde{f}(X^{(i)}) - \mu_n(\tilde{f}))(\tilde{f}(X^{(i)}) - \mu_n(\tilde{f}))^T \\ \mu_n(\tilde{f}) &\equiv \frac{1}{n}\sum_{i=1}^{n}\tilde{f}(X^{(i)}) \end{aligned} \tag{8}$$

This is the first method of estimating Σ, In Liu et al.(2012), there are other two methods could be used. We consider the following statistics:

(Spearman's rho)

$$\hat{\rho}_{jk} = \frac{\sum_{i=1}^{n}(r_j^i - \overline{r}_j)(r_k^i - \overline{r}_k)}{\sqrt{\sum_{i=1}^{n}(r_j^i - \overline{r}_j)^2 \cdot \sum_{i=1}^{n}(r_k^i - \overline{r}_k)^2}} \tag{9}$$

(Kendall's tau)

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)}\sum_{1 \le i < i' \le n} sign((x_j^i - x_j^{i'})(x_k^i - x_k^{i'})) \tag{10}$$

We define the following estimators $\hat{S}^\rho = [\hat{S}_{jk}^\rho]$ and $\hat{S}^\tau = [\hat{S}_{jk}^\tau]$ for the unknown correlation matrix Σ

$$\hat{S}_{jk}^\rho = \begin{cases} 2\sin(\frac{\pi}{6}\hat{\rho}_{jk}), & j \ne k; \\ 1, & j = k. \end{cases} \tag{11}$$

and

$$\hat{S}_{jk}^\tau = \begin{cases} 2\sin(\frac{\pi}{2}\hat{\tau}_{jk}), & j \ne k; \\ 1, & j = k. \end{cases} \tag{12}$$

Estimation of Ω Introduced by Liu et al.(2012), there are four methods could be applied, and according to their results, these methods can be classified into two types. One is glasso and CLIME, the other one is MeinshausenCBhlmann procedure and graphical Dantzig selector. Here, we just use two methods because of the similarity.

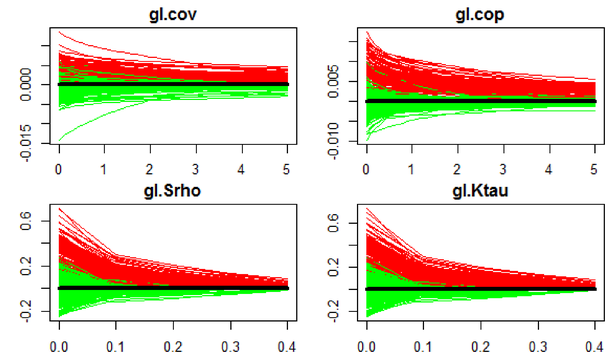
We can plug the estimated correlation matrix $\hat{S}$ into the graphical lasso:

$$\hat{\Omega}^{glasso} = \arg\min_{\Omega \succ 0}\{tr(\hat{S}\Omega) - \log|\Omega| + \lambda\sum_{j,k}|\Omega_{jk}|\} \tag{13}$$

We can also use the MeinshausenCBhlmann procedure to estimate the graph. As has been discussed in Friedman, Hastie and Tibshirani (2008), the correlation matrix is a sufficient statistic for the MeinshausenCBhlmann procedure.

Simulation

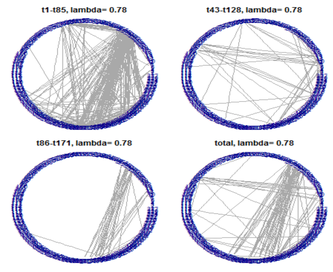
We generate graphs as in Liu et al.(2009) and Meinshausen and Bhlmann (2006), and use 3 methods to estimate Σ and 2 methods to estimate Ω. Since that there is not much space, we do not describe the process of generating, and just show the comparison among the different estimations of Σ with glasso. The paths for the relevant variables (nonzero inverse covariance entries) are plotted as red lines; the paths for the irrelevant variables are plotted as green lines. We can find gl.Ktau and gl.Srho are better that others, in real data, we will use gl.Srho to do the work.



Real Data

The real data is a type of time varying lattice data, with entries donate a count of disease cases at specific location on a specific date. The dates ($n = 171$) are collected by day, while the locations (p=227) are at subdistrict level. Note that during the time varying collection, the propagation of disease will be changed. We suppose that the latent relationship of cases count on corresponding locations behaviors similarly.

Due to the result of simulation, we choose the Spearmans rho estimates to infer correlation by glasso, at each of total times $[0, t_n]$ and 3 time intervals, $[0, t_n/2]$, $[t_n/4, 3t_n/4]$, $[t_n/2, t_n]$. The first half interval shows a intensity communication of disease case than the second and third intervals. While the first graph is also denser than the total time scale correlation estimation at the same sparse penalty level ($\lambda = 0.78$). Our exploration shows that the time varying affection should not be ignored, and the time varying graphic model would have more attention for further research.



Contribution and References

Pan Hao: Real data analysis; Huang Yimin: Simulation; Zhang Yu: Paper review.

[1] Liu H, Lafferty J, Wasserman L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs[J]. The Journal of Machine Learning Research, 2009, 10: 2295-2328.
[2] Liu H, Han F, Yuan M, et al. High-dimensional semiparametric Gaussian copula graphical models[J]. The Annals of Statistics, 2012, 40(4): 2293-2326.
[3] Zhou S, Lafferty J, Wasserman L. Time varying undirected graphs[J]. Machine Learning, 2010, 80(2-3): 295-319.
[4] Kolar M, Xing E P. On time varying undirected graphs[J]. 2011.