Structure learning of undirected graphical models for count data

Kim Hue Nguyen and Monica Chiozza
Department of Statistical Sciences, University of Padova, Padova, Italy
nguyen@stat.unipd.it

Purpose

Count data are increasingly ubiquitous in big-data settings such as genomic sequencing data, user-rating data, spatial incidence data, site visits...

Research has so far mainly focused on graphical models over binary, multinomial and Gaussian random variables only.

Little work makes use of Poisson assumptions: SPCAM, QPCAM, TPM [1], LPGM [2], PDNs [3]. Yet, some problems remain to be solved: existing of a consistent joint distribution, possible inaccurate inferences when dealing with models of high dimension...

Here, we will concentrate on investigating a new algorithm for structure learning of undirected Poisson graphical models, called PC-LPGM:

- able to reconstruct the underlying structure from a set of given data;
- feasible up to high dimensional data;
- out performing on average state-of-the-art algorithms.

Problem

Consider a p-random vector $X = (X_1, X_2, \ldots, X_p)$. Let $V = \{1, 2, \ldots, p\}$ and assume each conditional distribution $X_{i|\{X_{\neq i}\}}$ follows a Poisson distribution,

$$X_{i|\{X_{\neq i}\}} \sim \text{Poiss}(\theta_{i|\{X_{\neq i}\}}),$$

where, $\lambda_{i|\{X_{\neq i}\}} = \exp(\sum_{j \neq i} \theta_{ij} x_{ij})$. We note that there is one edge between $s$ and $t$, if and only if $\theta_{st} \neq 0$ or $\theta_{ts} \neq 0$.

Problem: Learning an undirected (possibly sparse) graphical structure from given data, i.e. identifying the set of non-zero parameters $\theta_{ij}$.

Solution: Conditional independence tests, i.e. Wald type test on the parameters $\theta_{ij}$. In detail, assume $X_{ij|\{X_{\neq i \neq j}\}} \sim \text{Poiss}(\sum_{k \neq i \neq j} \theta_{ijk} x_{ik} x_{k})$. For all $s \neq t \in V \setminus \{1, \ldots, p\}$, the test statistic for the hypothesis $H_0: \theta_{st} = 0$ is given by $Z_{st} = \sqrt{n} \hat{\theta}_{st} / \hat{\theta}_{st}^2$. The test statistic for the hypothesis $H_0: \theta_{st} = 0$ is given by $Z_{st} = \sqrt{n} \hat{\theta}_{st} / \hat{\theta}_{st}^2$, where $[a]_{ij}$ denotes the element in position $(i, j)$ of matrix $A$.

Algorithm

Let $\text{adj}(G, s) = \{t \in G: (s, t) \in E\}$ denotes the set of all nodes that are adjacent to $s$ on the graph $G$.

Algorithm 1 The PC-LPGM algorithm

1. Input: $n$ independent realizations of the $p$-random vector $X$, i.e. $X_1^n, X_2^n, \ldots, X_p^n$.
2. Output: An estimated undirected graph $\hat{G}$.
3. Form the complete undirected graph $\hat{G}$ on the vertex set $V$.
4. $l = 1, \ldots, 3$
5. repeat
6. $l = l + 1$
7. for all vertices $s \in V$, do
8. let $K = \text{adj}(G, s)$
9. end for
10. repeat
11. Select a (new) ordered pair of nodes $s, t$ that are adjacent in $\hat{G}$.
12. if $|K| \geq 1$, using order($|V|$).
13. repeat
14. Choose a (new) set $S \subseteq K$ with $|S| = l$ using order($|V|$).
15. if $H_0: \theta_{st} = 0$ not rejected
16. delete edge $(s, t)$ from $\hat{G}$
17. end if
18. until edge $(s, t)$ is deleted or all $S \subseteq K$ with $|S| = l$ have been considered;
19. end if
20. until $l = m$ or for each ordered pair at adjacent nodes $s, t$: $|\text{adj}(G, s)| < l$.

Simulation setup

For two different cardinalities, i.e., $p = 10$ and $p = 100$, we consider three graphs of different structure: (i) a scale-free graph, (ii) a hub graph, (iii) a random graph. For each graph, 500 datasets were sampled as in [2] for three sample sizes, i.e., $n = 200, 1000, 2000$.

We adopt two measures: PPV that stands for Positive Predictive Value and is defined as $TP/(TP + FP)$, and Sensitivity (Se), defined as $TP/(TP + FN)$, where $TP$ (true positive), $FP$ (false positive), and $FN$ (false negative) refer to the inferred edges.

Results

For $p = 10$, PC-LPGM outperforms the other approaches on average in term of reconstructing the structure from given data.

When $p = 10$, PC-LPGM reaches the highest TP value, followed by PDN and LPGM. Among the algorithms with highest PPV, PC-LPGM shows a sensitivity approaching 1 already at the sample size $n = 1000$.

PC-LPGM is far better than that of the competing algorithms employing the Poisson assumption, i.e., PDN and LPGM. This might be explained in terms of difference between penalization and restriction of the conditional sets.

Gaussian based methods (VSL, GLASSO) perform reasonably well, with an inferior score with respect to the learning thesesome.

Sophisticated techniques that replace the Gaussian distribution with a more flexible continuous distribution such as the nonparanormal distribution, e.g., NPN-Copula, NPN-SKEPTIC can show slight gains in accuracy over the naive analysis.

Results for the high dimensional setting ($p = 100$) are somehow comparable. The PC-LPGM outperforms all competing methods, and differences among algorithms are even more evident.

Conclusion

PC-LPGM outperforms the other approaches on average in term of reconstructing the structure from given data.

When $p = 10$, PC-LPGM reaches the highest TP value, followed by PDN and LPGM. Among the algorithms with highest PPV, PC-LPGM shows a sensitivity approaching 1 already at the sample size $n = 1000$.

PC-LPGM is far better than that of the competing algorithms employing the Poisson assumption, i.e., PDN and LPGM. This might be explained in terms of difference between penalization and restriction of the conditional sets.

Gaussian based methods (VSL, GLASSO) perform reasonably well, with an inferior score with respect to the learning thesesome.

Sophisticated techniques that replace the Gaussian distribution with a more flexible continuous distribution such as the nonparanormal distribution, e.g., NPN-Copula, NPN-SKEPTIC can show slight gains in accuracy over the naive analysis.

Results for the high dimensional setting ($p = 100$) are somehow comparable. The PC-LPGM outperforms all competing methods, and differences among algorithms are even more evident.

References


Other approaches to compare

- **PC-LPGM**
- **LPDM**
- **PDNs**
- **VSL**
- **GLASSO**
- **NPN-Copula**
- **NPN-SKEPTIC**

The local Poisson graphical model algorithm [2].

The Poisson Dependency Network algorithm [3].

The variable selection with lasso algorithm [5] on log-transformed data $\log(1 + X)$.

The graphical lasso algorithm [6] on log-transformed data $\log(1 + X)$.

The nonparanormal-Copula algorithm [7].

The nonparanormal-SKEPTIC algorithm [8].