Structure learning of undirected graphical models for count data

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

Università degli Studi di Padova

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 focussed on graphical models over binary, multinomial and Gaussian random variables only.

Litle work makes use of Poisson assumption: SPGM, QPGM, TPGM [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 $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. Let $V = \{1, 2, \ldots, p\}$ and assume each conditional distribution $X_s|\mathbf{x}_{\{V\setminus\{s\}\}}$ follows a Poisson distribution,

$$X_s|\mathbf{x}_{\{V\setminus\{s\}\}} \sim \text{Pois}(f_s(\mathbf{x}_{\{V\setminus\{s\}\}})),$$

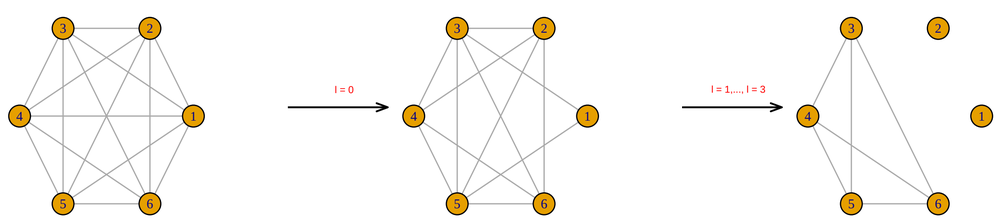
where, $f_s(\mathbf{x}_{\{V\setminus\{s\}\}}) = \exp\{\sum_{t\neq s} \theta^*_{st} x_t\}$. We note that there is one edge between s and t, if and only if $\theta^*_{st} \neq 0$ or $\theta^*_{ts} \neq 0$.
Problem: Learning an undirected (possibly sparse) graphical structure from given data, i.e. identifying the set of non-zero parameters θ^*_{st}.
Solution: Conditional independence tests, i.e. Wald type tests on the parameters θ_{st}. In detail, assume $X_s|\mathbf{x}_K \sim \text{Pois}(\exp\{\sum_{t\in K} \theta^*_{st|K} x_t\})$, $\forall s \in V, K \subset \{1, \ldots, p\}\setminus\{s\}$. The test statistic for the hypothesis $H_0 : \theta_{st|K} = 0$ is given by $Z_{st|K} = \dfrac{\sqrt{n}\hat{\theta}_{st|K}}{\sqrt{\left[J(\hat{\theta}_{s|K})^{-1}\right]_{tt}}}$, where $[A]_{jj}$ denotes the element in position (j, j) of matrix A.

Algorithm

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

Algorithm 1 The PC-LPGM algorithm

Input: n independent realizations of the p-random vector $\mathbf{X}$; i.e., $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}$; an ordering $order(V)$ on the variables, (and a stopping level m).
Output: An estimated undirected graph $\hat{G}$.
Form the complete undirected graph $\tilde{G}$ on the vertex set V.
$l = -1$; $\quad \hat{G} = \tilde{G}$
repeat
  $l = l + 1$
  for all vertices $s \in V$, **do**
    let $K = \text{adj}(G, s)$
  end for
  repeat
    Select a (new) ordered pair of nodes s, t that are adjacent in $\hat{G}$ s.t.
    $|K\setminus\{t\}| \geq l$, using $order(V)$.
    repeat
      choose a (new) set $S \subset K\setminus\{t\}$ with $|S| = l$, using $order(V)$.
      if $H_0 : \theta_{st|S} = 0$ not rejected
        delete edge (s, t) from $\hat{G}$
      end if
    until edge (s, t) is deleted or all $S \subset K\setminus\{t\}$ with $|S| = l$ have been considered.
  until all ordered pair of adjacent variables s and t such that $|K\setminus\{t\}| \geq l$ and
    $S \subset K\setminus\{t\}$ with $|S| = l$ have been tested for conditional independence.
until $l = m$ or for each ordered pair of adjacent nodes s, t: $|\text{adj}(G, s)\setminus\{t\}| < l$.



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., $n = 200, 1000, 2000$.

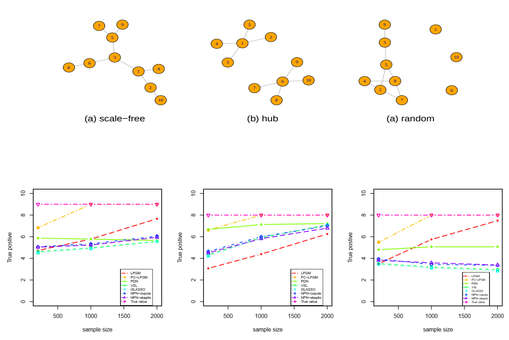
We adopt two measures: PPV that stands for Positive Predictive Value and is defined as $TP/(TP + FP)$; and Sensitivity (Se), defined as $TP/(TP + FN)$, where TP (true positive), FP (false positive), and FN (false negative) refer to the inferred edges.

Other approaches to compare

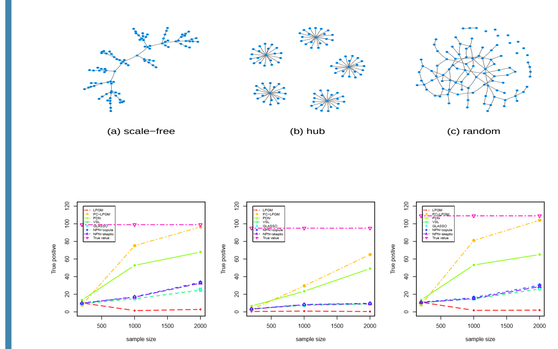
PC-LPGM	The PC-LPGM algorithm, i.e. the proposed algorithm.
LPGM	The local Poisson graphical model algorithm [2].
PDNs	The Poisson Dependency Network algorithm [3].
VSL	The variable selection with lasso algorithm [5] on log-transformed data $log(1 + X)$.
GLASSO	The graphical lasso algorithm [6] on log-transformed data $log(1 + X)$.
NPN-Copula	The nonparanormal-Copula algorithm [7].
NPN-SKEPTIC	The nonparanormal-SKEPTIC algorithm [8].

Results

Results for p = 10



Results for p = 100



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 leading threesome.
- 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 more evident.

References

[1] E. Yang, P. Ravikumar, G. Allen, and Z. Liu. On poisson graphical models. In *Advances in Neural Information Processing Systems*, pages 1718–1726, 2013.

[2] G. Allen and Z. Liu. A local poisson graphical model for inferring networks from sequencing data. *NanoBioscience, IEEE Transactions on*, 12(3):189–198, 2013.

[3] F. Hadiji, A. Molina, S. Natarajan, and K. Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning*, 100(2-3):477–507, 2015.

[4] H. Liu, K. Roeder, and L. Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440, 2010.

[5] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, pages 1436–1462, 2006.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[7] H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

[8] H. Liu, F. Han, M. Yuan, J. Lafferty, and L. Wasserman. The nonparanormal skeptic. *arXiv preprint arXiv:1206.6488*, 2012.