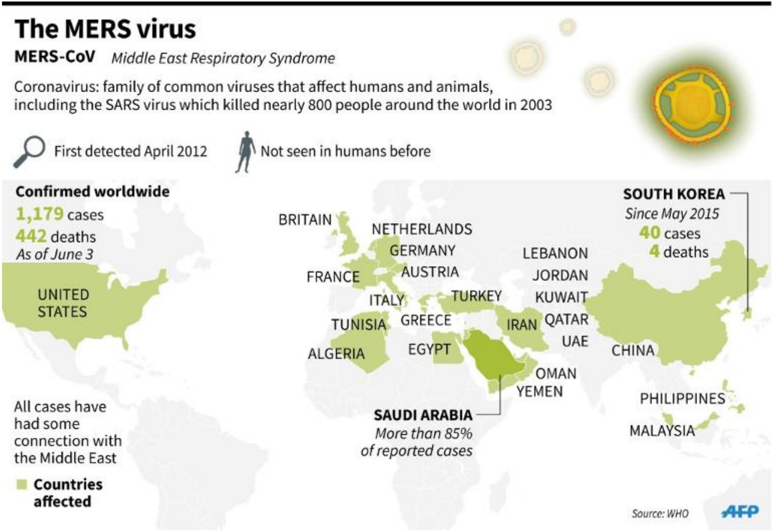
High-Dimensional Poisson DAG Model Learning Using ℓ_1-Regularized Regression

Gunwoong Park, Sion Park

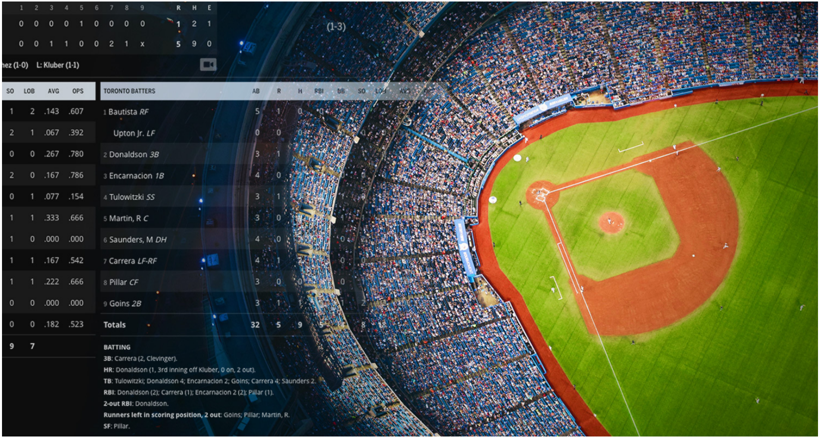
University of Seoul, Statistics

2019-01-15

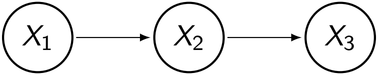
Infection Path - MERS



Sports Data - Player's ability

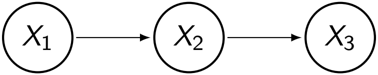


What is the Directed Graphical Model?



- X_1, X_2 and X_3 are random variables.
- X_1 affects X_2. Similarly X_2 also affects X_3.
- X_1 is a parent of X_2. Conversely, X_2 is a child of X_1.

Directed Acyclic Graphical (DAG) Model: Notations



- $X_{\mathrm{Pa}(j)}$ is a parents set of X_j. $e.g.X_{\mathrm{Pa}(3)} = \{X_2\}$.

- The set $De(k)$ denotes the set of all descendants of node k.
  Non-desendants set $Nd(k) := V \setminus (\{k\} \cup De(k))$.
  $e.g.De(1) = \{2, 3\}, Nd(3) = \{1, 2\}$

- Ordering : an indicator that indicates causal ordering.
  In this graph,

$$\pi = \{1, 2, 3\}$$

DAG Model: Factorization

$$X_1 \longrightarrow X_2 \longrightarrow X_3$$

Factorization(Lauritzen, 1996):

$$f(X_1, X_2, \ldots, X_p) = \prod_{j=1}^{p} f(X_j | X_{\mathrm{Pa}(j)}).$$

- In this graph,

$$\begin{aligned} f(X_1, X_2, X_3) &= f(X_1) f(X_2|X_1) f(X_3|X_2, X_1) \\ &= f(X_1) f(X_2|X_1) f(X_3|X_2). \end{aligned}$$

Poisson DAG Model

$$\begin{tikzcd} X_1 \arrow[r] & X_2 \arrow[r] & X_3 \end{tikzcd}$$

- Each node follows a conditional Poisson distribution given its parents.
- $X_1 \sim Poi(\lambda)$, $X_2|X_1 \sim Poi(g_2(X_1))$, $X_3|X_2 \sim Poi(g_3(X_2))$.

  where $g_j(.)$ is a function of $X_{\mathrm{Pa}(j)}$ to determine X_j's parameters.
- $f(X_1, X_2, X_3) = f(X_1)f(X_2|X_1)f(X_3|X_2)$.
  - $X_1, X_2, X_3 \sim Poi(\lambda) \times Poi(g_2(X_1)) \times Poi(g_3(X_2))$.

Poisson Structural Equation Models

Poisson SEM(Structural Equation Models)

In Poisson DAG,

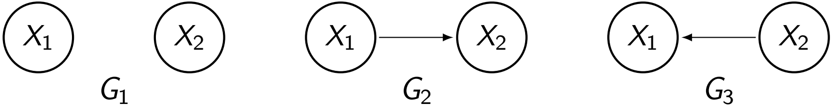
$$X_j \mid X_{\mathrm{Pa}(j)} \sim Poisson(g_j(X_{\mathrm{Pa}(j)})),$$

where rate parameter $g_j(X_{\mathrm{Pa}(j)}) = \exp(\theta_j + \sum_{k \in \mathrm{Pa}(j)} \theta_{jk} X_k)$

- Rate parameter function g_j is canonical form.
- Parents contribute to child's variability.

Question: Model Identifiability

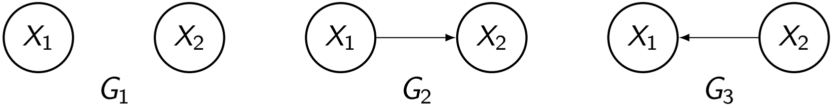
Is it possible to recover a graph from distribution? **Partially Yes.**



- We can distinguish G_2 and G_3 from G_1.
- We cannot identify a direction of an edge. Hence, we cannot distinguish G_2 and G_3.

Model Identifiablity

Is it possible to recover a graph from count data? **Yes**



Suppose that $X_1 \sim Poi(\lambda_1)$ and $X_2 \sim Poi(\lambda_2)$ where the parameter only depends on the parents in the graph.

- For G_1, $\text{Var}(X_1) = \mathbb{E}(X_1)$ and $\text{Var}(X_2) = \mathbb{E}(X_2)$.
- For G_2, $\text{Var}(X_1) = \mathbb{E}(X_1)$, while

$$\begin{aligned}\text{Var}(X_2) &= \mathbb{E}[\text{Var}(X_2 \mid X_1)] + \text{Var}[\mathbb{E}(X_2 \mid X_1)] \\ &= \mathbb{E}[\mathbb{E}(X_2 \mid X_1)] + \text{Var}[\mathbb{E}(X_2 \mid X_1)] > \mathbb{E}(X_2).\end{aligned}$$

Identifiability: Moment Ratio

Theorem: Identifiability

For any node $j \in V$, $non-empty$ $Pa_0(j) \subset Pa(j)$ and $S_j \subset Nd(j) \setminus Pa_0(j)$,

$$\mathbb{E}(X_j^2) > \mathbb{E}(\mathbb{E}(X_j \mid X_{S_j}) + \mathbb{E}(X_j \mid X_{S_j})^2),$$

the Poisson DAG model is identifiable.

- The identifiability condition in theorem is equivalent to

$$\mathbb{E}(\mathrm{Var}(\mathbb{E}(X_j \mid X_{\mathrm{Pa}(j)}) \mid X_{S_j})) > 0,$$

if all parents of node j contribute to its variability.

MRS Algorithm

Solve ℓ_1-regularized Poisson regression $\hat{\theta}(j)$

Ordering Estimation

Mean-variance Ratio Score.

Parents Estimation

$\mathrm{Pa}(j) := \{k \in V \mid \widehat{\theta}_{jk} \neq 0\}$.

ℓ_1-Regularized Poisson Regression

Assume that $X_j \mid X_S \sim Poisson(\lambda)$ then $\hat{\lambda} = g_j(X_S)$.

$$g_j(X_S) = \exp(\theta_j + \sum_{k \in S} \theta_{jk} X_k)$$

- The estimate of $\theta_S(j) = \left(\theta_j, (\theta_{jk})_{k \in S}\right)$:

$$\widehat{\theta}_S(j) := \arg\min \frac{1}{n} \sum_{i=1}^{n} \left[-X_j^{(i)} \left(\theta_j + \sum_{k \in S} \theta_{jk} X_k^{(i)} \right) \right.$$
$$\left. + \exp\left(\theta_j + \sum_{k \in S} \theta_{jk} X_k^{(i)} \right) \right] + \lambda_j \sum_{k \in S} |\theta_{jk}|.$$

- Use estimate that minimizes negative log likelihood function.

MRS Algorithm - Scoring

Mean-variance Ratio Score (MRS) :

$$\widehat{\mathcal{S}}_r(1,j) := \frac{\widehat{\mathbb{E}}(X_j^2)}{\widehat{\mathbb{E}}(X_j) + \widehat{\mathbb{E}}(X_j)^2} \quad \text{and} \quad \widehat{\mathcal{S}}_r(m,j) := \frac{\widehat{\mathbb{E}}(X_j^2)}{\widehat{\mathbb{E}}\big(\widehat{\mathbb{E}}(X_j \mid X_S) + \widehat{\mathbb{E}}(X_j \mid X_S)^2\big)}$$

- $\widehat{\mathbb{E}}(X_j) = \frac{1}{n}\sum_{i=1}^n X_j^{(i)}$ *and* $\widehat{\mathbb{E}}(X_j^2) = \frac{1}{n}\sum_{i=1}^n (X_j^{(i)})^2$
- $\widehat{\mathbb{E}}(X_j \mid X_S) = \exp\big(\widehat{\theta}_j^S + \sum_{k\in S}\widehat{\theta}_{jk}^S X_k\big)$ where $\widehat{\theta}_j^S$ and $\widehat{\theta}_{jk}^S$ are the solution of the ℓ_1-regularized GLM.
- Condition set $S = \hat{\pi}_{1:(m-1)}$

MRS Algorithm - Estimation

Ordering and parents estimation :

$$\hat{\pi}_m = \arg\min_j \hat{\mathcal{S}}(m, j) \quad and \quad \mathrm{Pa}(\hat{\pi}_m) := \{k \in S \mid \widehat{\theta}_{jk} \neq 0\}$$

In population,

- $\mathcal{S}_r(m, j) = 1$, when $\mathrm{Pa}(j) \subset S$.
- Otherwise, $\mathcal{S}_r(m, j) > 1$.

MRS Algorithm: Example

$$\begin{tikzcd} X_1 \arrow[r] & X_2 \arrow[r, dotted] & X_3 \arrow[r, dotted] & X_4 \end{tikzcd}$$

- When true ordering is $\pi = \{1, 2, 3, 4\}$ and $\hat{\pi}_{3:4}$ is unknown, $\hat{\pi}_{1:2} = \{1, 2\}$.
- In population, we choose X_3 as third element of estimated ordering, because $\mathcal{S}_r(3, 3) < \mathcal{S}_r(3, 4)$.

  $\hat{\pi}_3 = \{3\}$.
- $\hat{g}_3(X_{\mathrm{Pa}(3)}) = \exp(\hat{\theta_3} + \hat{\theta_{31}} X_1 + \hat{\theta_{32}} X_2)$, $\hat{\theta_{31}} = 0$ *and* $\hat{\theta_{32}} \neq 0$. Therefore, X_2 is estimated parent of X_3.

Consistency of Algorithm

Theorem: Consistency

For any $\epsilon > 0$, there exists positive $C_\epsilon > 0$ such that if $n \geq C_\epsilon \left(d^2 \log^9 p\right)$,

$$P(G \neq \hat{G}) \leq \epsilon,$$

under the regularity conditions.

Notations

- $\hat{G}$: Estimated graph via the MRS algorithm.
- d: Maximum indegree of the graph.
- If the graph is sparse, the algorithm works in the high-dimensional setting.

Simulation Settings for Poisson SEM

200 **realizations of** p**-node Poisson SEM.**

Log-link:

$$g_j(X_{\text{Pa}(j)}) = \exp\left(\theta_j + \sum \theta_{jk} X_{\text{Pa}_k(j)}\right)$$

- For $d = 1$,

  $n \in \{25, 50, 75, \ldots, 250\}$,

  $\theta_j \in [1, 3]$,

  $\theta_{jk} \in [-1.5, -0.5] \cup [0.5, 1.5]$,

  $p \in \{20, 200\}$.

- For $d = 10$,

  $n \in \{100, 200, 300, \ldots, 1000\}$,

  $\theta_j \in [1, 3]$,

  $\theta_{jk} \in [-1, -0.1] \cup [0.1, 1]$,

  $p \in \{20, 200\}$.

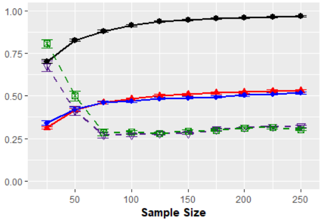
Simulation: Measurement

Average measurement of 200 realization sets.

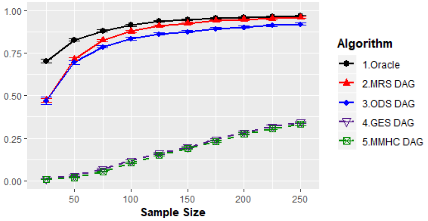
Measurement of result:

- Average precision : $\frac{\#\ \ of\ \ correctly\ \ estimated\ \ edges}{\#\ \ of\ \ estimated\ \ edges}$
- Average recall : $\frac{\#\ \ of\ \ correctly\ \ estimated\ \ edges}{\#\ \ of\ \ true\ \ edges}$

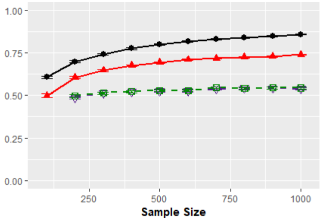
Simulation: PoissonSEM



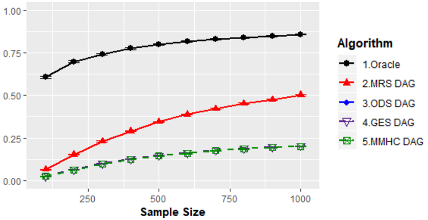
(a) Precision: $p=20, d=1$



(b) Recall: $p=20, d=1$

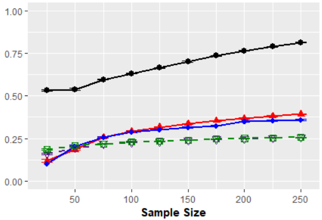


(c) Precision: $p=20, d=10$

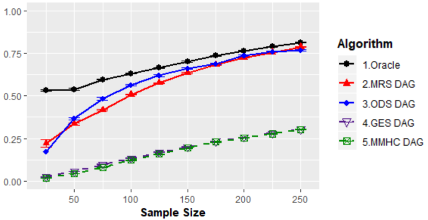


(d) Recall: $p=20, d=10$

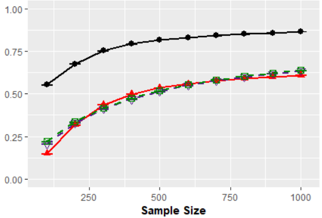
Simulation: Poisson SEM 2



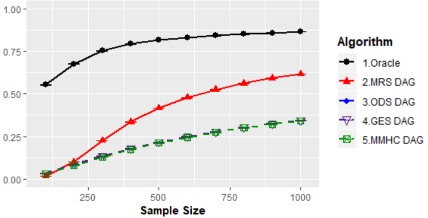
(a) Precision: p=200, d=1



(b) Recall: p=200, d=1

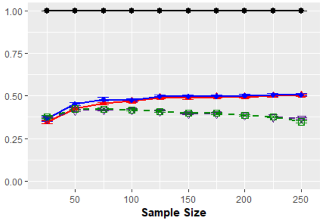


(c) Precision: p=200, d=10

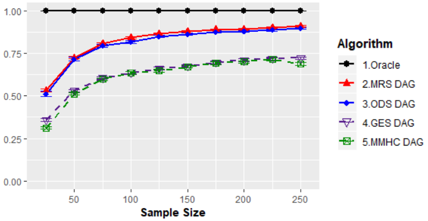


(d) Recall: p=200, d=10

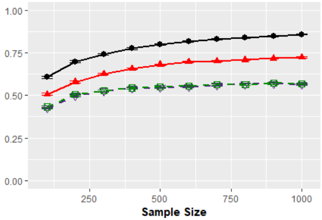
Simulation: Poisson SEM - MEC



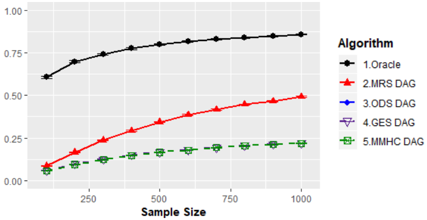
(a) Precision: $p=20, d=1$



(b) Recall: $p=20, d=1$

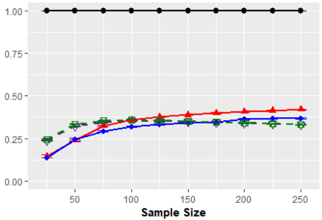


(c) Precision: $p=20, d=10$

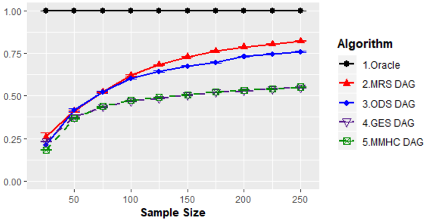


(d) Recall: $p=20, d=10$

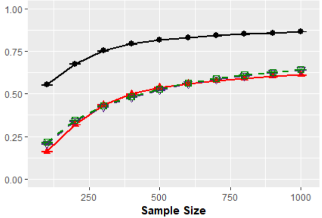
Simulation: Poisson SEM - MEC 2



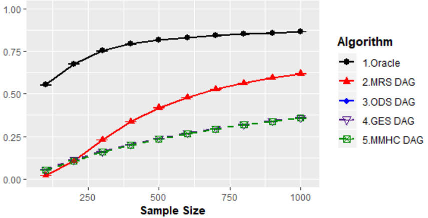
(a) Precision: $p{=}200, d{=}1$



(b) Recall: $p{=}200, d{=}1$



(c) Precision: $p{=}200, d{=}10$



(d) Recall: $p{=}200, d{=}10$

Simulation Setting for Poisson DAG

200 **realizations of p-node Poisson DAG.**

Identity-link:

$$g_j(X_{\mathrm{Pa}(j)}) = \theta_j + \sum \theta_{jk} X_{\mathrm{Pa}_k(j)}$$

- For $d = 2$,

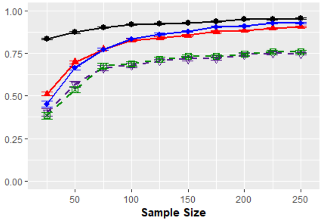
  $n \in \{25, 50, 75, \ldots, 250\}$,

  $\theta_j \in [1, 10]$,

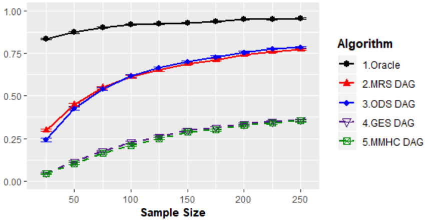
  $\theta_{jk} \in [-1.5, -0.5] \cup [0.5, 1.5]$,

  $p \in \{20, 100\}$.

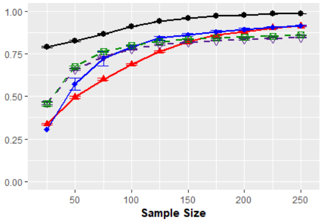
Simulation: Identity Link Function



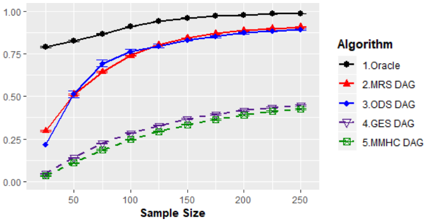
(a) Precision: $p=20, d=2$



(b) Recall: $p=20, d=2$

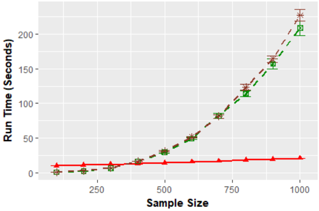


(c) Precision: $p=100, d=2$

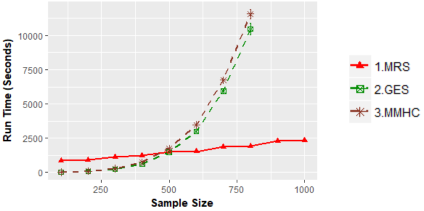


(d) Recall: $p=100, d=2$

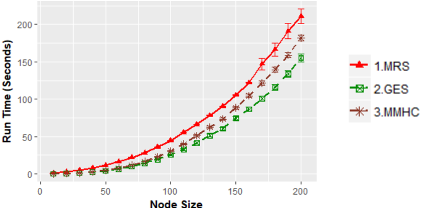
Simulation: Computational Complexity



(a) Varying n for p =100, d =5



(b) Varying n for p =500, d =5



(c) Varying p for n =500, d =5

Real Multivariate Count Data: MLB Statistics

- 800 MLB players salary and batting statistics from the 2003 season from R package Lahman.
- We considered players in the top 25% in terms of the number of games played.

feature	name	feature	name
G	Games played	SB	Stolen Bases
AB	At Bats	BB	Bases on Balls
R	Runs	SO	Strikeouts
H	Hits	IBB	Intentional Walks
X2B	Doubles	HBP	times Hit by Pitch
X3B	Triples	SH	Sacrifice Hits
HR	Home Runs	SF	Sacrifice Flies
RBI	Run Batted In	GIDP	times Ground into Double Plays
CS	times Caught Stealing	salary	Annual salary

Real Multivariate Count Data: MLB Statistics

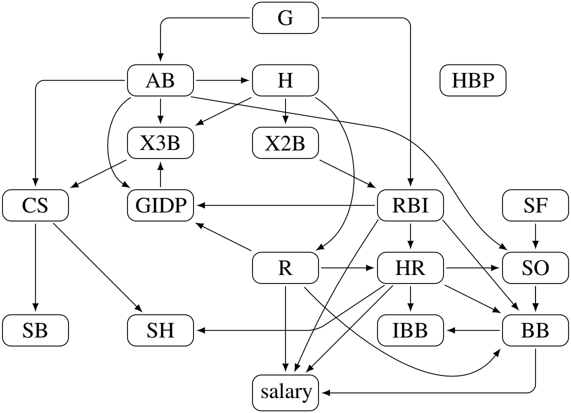
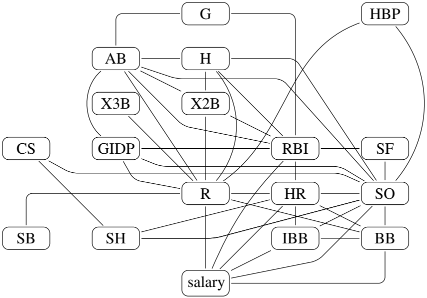
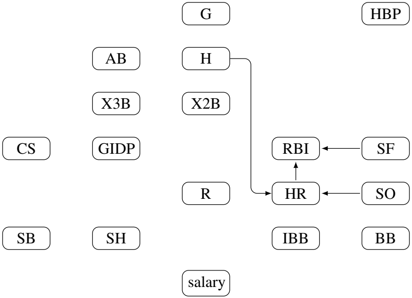


Figure: Estimated Graph by MRS Algorithm

Real Multivariate Count Data: MLB Statistics



(a) Estimated Graph by Poisson MRF



(b) Estimated Graph by MMHC

Summary

- By calculationg moment ratio via ℓ_1-regularized regression, we can learn Poisson SEM.
- MRS algorithm via is consistent, and sample bound $n = \Omega(d^2 \log^9 p)$.
- MRS algorithm works well for identity link function.
  - MRS algorithm learns Poisson DAG for MLB statistics successfully.

Reference

- Gunwoong Park and Sion Park. High-Dimensional Poisson DAG Model Learning via ℓ_1-Regularized Regression.
- Gunwoong Park and Garvesh Raskutti, Learning Large-Scale Poisson DAG Models based on OverDispersion Scoring.
- E. Yang, G. Allen, Z. Liu, and P. K. Ravikumar, Graphical models via generalized linear models.

Thank You

Additional Explanation: Assumption 1

Dependence Assumption:

For any $j \in V$ with $S_j \in \{\{\pi_1\}, \{\pi_1, \pi_2\}, ..., \{\pi_1, ..., \pi_{j-1}\}\}$ and non-empty $T_j = S_j \cap \mathrm{Pa}(j)$, there exists positive constants $\rho_{\min}$ and $\rho_{\max}$ such that

$$\min_{j \in V} \lambda_{\min}\left(Q^{j,S_j}_{T_j,T_j}\right) \geq \rho_{\min}, \quad \text{and} \quad \max_{j \in V} \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} X^{(i)}_{\mathrm{Pa}(j)}(X^{(i)}_{\mathrm{Pa}(j)})^T\right) \leq \rho_{\max},$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and largest eigenvalues of the matrix A, respectively.

Additional Explanation: Assumption 2

Incoherence Assumption:

For any $j \in V$ with $S_j \in \{\{\pi_1\}, \{\pi_1, \pi_2\}, ..., \{\pi_1, ..., \pi_{j-1}\}\}$ and non-empty $T_j = S_j \cap \mathrm{Pa}(j)$, there exists a constant $\alpha \in (0, 1]$ such that

$$\max_{j,S_j} \max_{t \in T_j^c} \| Q^{j,S_j}_{tT_j} (Q^{j,S_j}_{T_j T_j})^{-1} \|_1 \leq 1 - \alpha.$$

Concentration Bound Assumption:

For any $j \in V$ and $\theta^*_{S_j}(j) = (\theta^{S_j}_j, \theta^{S_j}_{jk})$, there exists a constant $M_{\max} > 0$ such that

$$\max_{j,S_j} \mathbb{E}\left(e^{\exp(\theta^{S_j}_j + \sum_{k \in S_j} \theta^{S_j}_{jk} X_k)}\right) < M_{max}.$$

Additional Explanation: Assumption 3

For all $j \in V$ and any non-empty $\mathrm{Pa}_0(j) \subset \mathrm{Pa}(j)$ where $S_j \subset \mathrm{Nd}(j) \setminus \mathrm{Pa}_0(j)$, there exists an $M_{\mathrm{min}} > 0$ such that

$$\mathbb{E}(X_j^2) > (1 + M_{\mathrm{min}})\mathbb{E}(\mathbb{E}(X_j \mid X_{S_j}) + \mathbb{E}(X_j \mid X_{S_j})^2).$$

ODS algorithm implementations

n	100	200	300	400	500	600	700	800	900	1000
p = 20	199	175	107	64	1	0	0	0	0	0
p = 50	200	200	200	199	192	179	151	140	99	86

Number of failures in ODS algorithm implementations from among 200 sets of samples for different node sizes $p \in \{20, 50\}$, and sample sizes $n \in \{100, 200, ..., 1000\}$, when the indegree is $d = 5$.