Report of Variable Selection of β Model

Ergan Shang

USTC

October 17, 2022

Overview

Overview

1. Background and Introduction
   - Introduction
   - Algorithm
2. Convergence Rate
3. Parameter Tuning

Background

A networks with n nodes is presented by a graph $G_n = G_n(V, E)$, where V is the set of nodes or vertices and E is the set of edges or links. Let $A = (A_{i,j})_{i,j=1}^{n}$ be the adjacency matrix where $A_{i,j} \in \{0, 1\}$ is an indicator whether nodes i, j are connected:

$$A_{i,j} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are connected} \\ 0 & \text{if nodes } i \text{ and } j \text{ are not connected} \end{cases}.$$

Generally, a network is generated by the probability $P(A_{ij} = 1) := p_{ij}$, bring about those sufficient statistic sequence: $d_i = \sum_{j=1}^{n} A_{ij} = \sum_{j \neq i} A_{ij}$ and $d_+ = \sum_{1 \leq i < j \leq n} A_{ij}$ because we assume that there is no self-loop.

The β Model

Real-life networks are composed of several critical nodes with more connections with others, along with many "silent" nodes equipped with few links. In order to capture the feature of **heterogeneity** and **homophily**, we abandon the Erdős-Rényi model:

$$P(A_{ij} = 1) = \frac{e^{\mu}}{1 + e^{\mu}}.$$

Instead, we propose the β model as

$$P(A_{ij} = 1) = p_{ij} := \frac{e^{\beta_i + \beta_j}}{1 + e^{\beta_i + \beta_j}},$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_n)^\top \in \mathbb{R}^n$ is an unknown parameter. Correspondingly, the negative log-likelihood is

$$\ell_n(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \beta_i d_i + \sum_{1 \le i < j \le n} log(1 + e^{\beta_i + \beta_j}).$$

Optimization Problem

To recover the property of Erdős-Rényi model, we impose a concentration between several parameters by solving the following optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^n} \ell_n(\boldsymbol{\beta}) \quad \text{subject to } \|\boldsymbol{\beta} - \beta_0 \mathbf{1}\|_0 \leq s,$$

where $\mathbf{1}$ is an n-dimensional vector with its elements being 1, $s \in \{1, 2, \cdots, n-1\}$ is an integer-value parameter to control the sparsity of the network and also $\beta_0 \in \mathbb{R}$ is also unknown.
We denote the true parameter as $\boldsymbol{\beta}^*$ and the MLE as $\hat{\boldsymbol{\beta}}$.

Surrogate Function and MM Algorithm

Literatures nowadays often treat the problem as convex optimization. In order to propose algorithms tractable, we propose the surrogate function as

$$S(\beta|\beta^m) = \ell_n(\beta^m) + (G^m)^\top(\beta - \beta^m) + \frac{n-1}{4}\|\beta - \beta^m\|_2^2.$$

where β^m is the output in the m-th iteration of algorithms. Next lemma shows the reasonability of applying Majorization Minimization (MM) Algorithm.

Lemma 1

The function $S(\beta|\beta^m)$ majorizes the objective function $\ell_n(\beta)$ at β^m. That is,

$$\begin{aligned} S(\beta|\beta^m) &\geq \ell_n(\beta) \text{ for all } \beta, \\ S(\beta^m|\beta^m) &= \ell_n(\beta^m). \end{aligned}$$

Primal-Dual Algorthm

Algorithm 1 Parameter Estimation for Fixed s

Input: Adjacency matrix A, sparsity level s, the maximum number of iterations $m_{\max}$, and tolerance ϵ.

1: Initialize $\beta_0^0 = \frac{1}{2}\mathrm{logit}\left(\frac{\sum_{i=1}^n d_i}{n(n-1)}\right)$, $\boldsymbol{\beta}_i^0 = \beta_0^0$ and $G_i^0 = -d_i + \sum_{j\neq i} \frac{\exp(\beta_i^0+\beta_j^0)}{1+\exp(\beta_i^0+\beta_j^0)}$ for $i = 1, \ldots, n$.

2: **For** $m \in \{0, 1, 2, \cdots, m_{\max}\}$ do

3: Determine $\Delta_j^{m+1} = \frac{n-1}{4}\left(\beta_0 - \boldsymbol{\beta}_j^m + \frac{2}{n-1}G_j^m\right)^2$, $j = 1, \cdots, n$.

4: Update the active and inactive sets by
$$\mathcal{A}^{m+1} = \{j : \Delta_j^{m+1} \geq \Delta_{[s]}^{m+1}\}, \quad \mathcal{I}^{m+1} = (\mathcal{A}^{m+1})^c .$$

5: Determine $\boldsymbol{\beta}^m$ by $\boldsymbol{\beta}_{\mathcal{I}^{m+1}}^{m+1} = \beta_0^m$ and $\beta_{\mathcal{A}^{m+1}}^{m+1} = \left(\boldsymbol{\beta}^m - \frac{2}{n-1}G^m\right)_{\mathcal{A}^{m+1}}$.

6: Let $p = \dfrac{\sum_{i\in\mathcal{I}^{m+1}} d_i - \sum_{i\in\mathcal{A}^{m+1}} d_i + \sum_{i\neq j, i,j\in\mathcal{A}^{m+1}} \frac{\exp(\beta_i^{m+1}+\beta_j^{m+1})}{1+\exp(\beta_i^{m+1}+\beta_j^{m+1})}}{(n-s)(n-s-1)}$.

If $p < 1$, then $\beta_0^{m+1} = \frac{1}{2}\mathrm{logit}(p)$, else break.

7: **If** $\|\boldsymbol{\beta}^{m+1} - \boldsymbol{\beta}^m\|_2 < \epsilon$, break.

8: **end For**

9: **Return** $\hat{\mathcal{A}} = \mathcal{A}^{m+1}$ $\quad \hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{m+1}$.

Overview

Mild Conditions

We first give a condition on how well β_0^{m-1} approximates the truth β_0 and then we will give the theorem concerning the termination of the algorithm.

Condition 1

The ℓ_2 distance between the true β_0 and the estimator β_0^{m-1} satisfies that there exists constant $\tilde{C} \geq \sqrt{n-s}$ such that

$$(n-s)(\beta_0 - \beta_0^{m-1})^2 \leq \left(\frac{1}{1 - \frac{n-s}{\tilde{C}^2}} - 1 \right) \sum_{i=1}^{s} (\beta_i^m - \hat{\beta}_i)^2,$$

when the right support is identified.

From the condition, we can see that if β_0^{m-1} approximates β_0 well, i.e., the LHS will be smaller for the existence of $\tilde{C}$.

Convergence Rate

Theorem 1

Suppose β^{m+1} is the $(\mathrm{m}+1)$th iteration of 7, then there exists $\zeta \in (0,1)$, such that

$$\|\boldsymbol{\beta}^{m+1} - \hat{\boldsymbol{\beta}}\|_2 \leq \zeta^m \left(1 + \frac{\sqrt{n-s}}{\tilde{C} - \sqrt{n-s}}\right) \|\boldsymbol{\beta}_{\mathrm{ini}} - \hat{\boldsymbol{\beta}}\|_2,$$

if choosing $s = s^*$ and the right support is identified at the mth iteration, where $\boldsymbol{\beta}_{\mathrm{ini}}$ is the initial value in 7 and $\tilde{C}$ is the constant defined in 8.

Convergence Rate

Corollary 1

Let $L := \max_{1\leq i\leq n} |\beta_i^*| = o(loglogs)$. If the right sparsity level is chosen, i.e. $s = s^*$ and the right support is identified at the mth iteration, then the solution of the maximum liklelihood exists and

$$\|\boldsymbol{\beta}^{m+1} - \boldsymbol{\beta}^*\|_2 \leq \zeta^m \left(1 + \frac{\sqrt{n-s}}{\tilde{C} - \sqrt{n-s}}\right) \|\boldsymbol{\beta}_{\mathrm{ini}} - \hat{\boldsymbol{\beta}}\|_2 + s^{\frac{3}{2}} O_p\left(\frac{o((logs)^6)}{\sqrt{n}}\right),$$

where ζ and $\boldsymbol{\beta}^{m+1}$ is defined in 9.

Overview

1. Background and Introduction
   - Introduction
   - Algorithm
2. Convergence Rate
3. Parameter Tuning

GIC

In order to select the true sparsity level s, we propose the following Generalized Information Criterion (GIC):

$$GIC^* = 2\ell_n\left(\hat{\boldsymbol{\beta}}^{ora}(\mathcal{A})\right) + s \cdot \log\left(\frac{n(n-1)}{2}\right) \cdot \log(\log n),$$

where $\hat{\boldsymbol{\beta}}^{ora}(\mathcal{A}) := arg\min_{\{\boldsymbol{\beta}:\mathrm{supp}(\boldsymbol{\beta}-\beta_0\mathbf{1})=\mathcal{A}\}} \ell_n(\boldsymbol{\beta})$. For the convenience of the statement of theorems, we define

$$\delta_n = \inf_{\mathcal{A}\not\supset\mathcal{A}_0, |\mathcal{A}|\le K} \frac{2}{n(n-1)} I\left(\boldsymbol{\beta}^{KL}(\mathcal{A})\right),$$

and

$$\boldsymbol{\beta}^{KL}(\mathcal{A}) := arg \min_{\{\mathrm{supp}(\boldsymbol{\beta}-\beta_0\mathbf{1})=\mathcal{A}\}} I(\boldsymbol{\beta}(\mathcal{A})),$$

where $I(\boldsymbol{\beta}(\mathcal{A})) = \mathbb{E}[log(f^*/g_{\mathcal{A}})]$, with $\boldsymbol{\beta}(\mathcal{A})$ is an n-dimensional parameter vector with support of $\boldsymbol{\beta} - \beta_0\mathbf{1}$ being $\mathcal{A}$, f^* is the density of underlying true model, and $g_{\mathcal{A}}$ is the density of the model with population parameter $\boldsymbol{\beta}(\mathcal{A})$.

Mild Conditions

Now we ask for the constrain for the sparsity level and also the boundedness of $\hat{\beta}^{ora}$ and β^{KL}.

Condition 2

To ensure identifiability, we assume the unique minimizer $\beta^{KL}(\mathcal{A})$ for every $\mathcal{A}$ satisfying $|\mathcal{A}| \le K$, with $K = O(\sqrt{n})$.

Condition 3

The norm of $\hat{\beta}^{ora}(\mathcal{A})$ and $\beta^{KL}(\mathcal{A})$ should be bounded by its support, namely, $\begin{cases} \|\hat{\beta}^{ora}(\mathcal{A})\|_2^2 \le C|\mathcal{A}| \\ \|\beta^{KL}(\mathcal{A})\|_2^2 \le C|\mathcal{A}| \end{cases}$ for some constant C.

Final Algorithm

Algorithm 2 Parameter Estimation

Input: Adjacency matrix A and the maximum support size K.

1: Set degree sequence $\mathbf{d} = (d_1, \cdots, d_n)^T$.

2: Let $d_{(1)} > d_{(2)} > \cdots > d_{(m)}$ denote the distinctive values of d_i's.

3: Denote $S_k = \left\{ i \in \{1, \cdots, n\} : d_i = d_{(k)} \right\}$ and $s_k = |S_k|$, where $k = 1, 2, \cdots, m$.

4: **For** $s = s_1, s_1 + s_2, \cdots, \sum_{k=1}^{m} s_k$ **do**

5: $(\hat{\boldsymbol{\beta}}_s, \hat{\mathcal{A}}_s) =$**Algorithm 1**(s).

6: **end For**

7: Compute the minimum of GIC*:

$$s_{min} = arg \min_s GIC^*(\hat{\boldsymbol{\beta}}_s).$$

8: **Return** $(\hat{\boldsymbol{\beta}}_{s_{min}}, \hat{\mathcal{A}}_{s_{min}})$.

Theorems and Conclusions

Theorem 2

Supposing $(\hat{\beta}, \hat{\mathcal{A}})$ is the solution of 13 and under Conditions 1 - 3, if $\delta_n K^{-1} R_n^{-1} \to \infty$ and $\frac{n(n-1)}{2}\delta_n s^{-1}\left(\log\left(\frac{n(n-1)}{2}\right)\log\log n\right)^{-1} \to \infty$, then as $n \to \infty$,

$$P\left(\hat{\mathcal{A}} = \mathcal{A}^*\right) \to 1,$$

where $R_n = \frac{\sqrt{\log n}}{n(n-1)}$.

Theorems and Conclusions

Corollary 2

Suppose $\beta^m_{s_{min}}$ is the mth iteration after choosing s_{min} via GIC*, $\hat{\mathcal{A}}$ is the solution of 13 and under the same conditions of Theorem 2, then for some constant $\zeta \in (0,1)$, with probability at least $1 - C(L)n^{-2}$, the solution of the MLE $\hat{\beta}$ exists satisfying

$$\|\beta^{m+1}_{s_{min}} - \beta^*\|_2 \le \zeta^m \left(1 + \frac{\sqrt{n - s^*}}{\tilde{C} - \sqrt{n - s^*}}\right) \|\beta^{\text{ini}}_{s_{min}} - \hat{\beta}\|_2 + s^{*\frac{3}{2}} O_p\left(\frac{o((logs^*)^6)}{\sqrt{n}}\right),$$

where $\beta^{\text{ini}}_{s_{min}}$ is the initial value chosen when the sparsity level s_{min} is selected in 7. Moreover, we have

$$P(\hat{\mathcal{A}} = \mathcal{A}^*) \to 1,$$

as $n \to \infty$.

Thank you !