Statistics and Probability Letters 119 (2016) 163–169

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Asymptotics of score test in the generalized β-model for networks

Ting Yan [a], Yunpeng Zhao [b,*]

[a] Department of Statistics and Hubei Key Laboratory of Mathematical Sciences, Central China Normal University, Wuhan, 430079, China

[b] Department of Statistics, George Mason University, VA 22030-4444, USA

ARTICLE INFO

Article history:
Received 17 March 2016
Received in revised form 23 July 2016
Accepted 24 July 2016
Available online 9 August 2016

MSC:
62E20
62F05

Keywords:
Central limit theorem
Network models
Score test

ABSTRACT

We show that the scaled score test statistic under a simple null in the generalized β-model for undirected networks asymptotically follows standard normal distribution when the number of network vertices goes to infinity.

1. Introduction

Pairwise relationships among a set of objects (e.g., friendships between individuals, citations between papers) can be conveniently represented in a network. Many statistical models have been developed to describe the generation mechanism of networks (see Bhattacharyya and Bickel, 2015; Goldenberg et al., 2010; Matias and Robin, 2014 for recent review), with applications in biology, computer science, social sciences and many other areas. Meanwhile, the non-standard structure of network data poses new challenge for statistical inference since typically, only one realized network is available (Fienberg, 2012).

The degrees of network vertices are one of the most important network statistics and provide important insights to understand more complex network structures such as the "small-world phenomenon" in Chung and Lu (2002) and some refined network statistics (e.g., "alternating k-star statistic") developed by Snijders et al. (2006). One natural approach for modeling the degrees is to put them as sufficient statistics for exponential family distributions on graphs according to the Koopman–Pitman–Darmois theorem (Koopman, 1936; Pitman, 1936; Darmois, 1935) or the principle of maximum entropy. Chatterjee et al. (2011) coined this model as β-model and proved the uniform consistency of the maximum likelihood estimator (MLE); Yan and Xu (2013) established its asymptotic normality. Hillar and Wibisono (2013) generalized the β-model to weighted edges and also proved that the MLE is uniformly consistent.

In this note we further establish the asymptotic normality of one of the score test statistics under the generalized β-model for the finite discrete weighted edges, i.e., $(2n)^{-1/2}(T-n) \xrightarrow{D} N(0,1)$ as n goes to infinity, where T is the score test

* Corresponding author.
E-mail addresses: tingyanty@mail.ccnu.edu.cn (T. Yan), yzhao15@gmu.edu (Y. Zhao).

http://dx.doi.org/10.1016/j.spl.2016.07.022
0167-7152/

statistic under a simple null and n is the number of vertices of the network. In this model, each vertex is assigned with one parameter to measure its strength to participate in network connection. As n increases, the number of parameter diverges. It makes the asymptotic inference nonstandard. To the best of our knowledge, this is the first asymptotic result for the score test with a diverging number of parameters in network models as n the number of vertices of the network goes to infinity. A key step in the proof is that we use the approximate inverse of the Fisher information matrix in Yan and Xu (2013) to obtain the approximately explicit expression for T. Another technical step is to construct a dependency graph to obtain the asymptotic distribution for the weighted quadratic sum of the centered degrees.

The rest of the paper is organized as follows. In Section 2, we lay out the main result, i.e., the asymptotic distribution of our score test statistic in the generalized β-model. Simulation studies are provided in Section 3. Section 4 concludes with some discussion and future work. All proofs are given in Appendix A.

2. Main results

Consider an undirected graph $\mathcal{G}_n$ on n vertices labeled by $1, \ldots, n$ with no self-loops. Let a_{ij} be the weight of edge (i, j) taking q discrete values from the set $\Omega = \{0, \ldots, q-1\}$, where q is assumed to be fixed. The β-model is only involved with the binary edges (i.e., $q = 2$). Here, we consider a generalized β-model for finite discrete weighted edges proposed by Hillar and Wibisono (2013). Rinaldo et al. (2013) defined a different version of generalized β-model by assuming that a_{ij} comes from n Bernoulli trials, which we did not consider here. Define the degree of vertex i by $d_i = \sum_{j\neq i} a_{ij}$ and the degree sequence of $\mathcal{G}_n$ by $\mathbf{d} = (d_1, \ldots, d_n)^T$. The probability density function of $\mathcal{G}_n$ under the generalized β-model is

$$p_\beta(\mathcal{G}_n) = \exp\big(\boldsymbol{\beta}^\top \mathbf{d} - z(\boldsymbol{\beta})\big), \tag{1}$$

where $z(\boldsymbol{\beta})$ is the normalizing constant. The parameters $\beta_1, \ldots, \beta_n$ measure the strength of each vertex participating in network connection. It can be obtained from (1) that the edges a_{ij}s for all $1 \le i < j \le n$ are mutually independent with the probability:

$$P(a_{ij} = a) = \frac{e^{a(\beta_i+\beta_j)}}{\sum\limits_{k=0}^{q-1} e^{k(\beta_i+\beta_j)}}, \quad a = 0, 1, \ldots, q-1.$$

When $q = 2$, it reduces to the β-model.

In general, only one realization of a random network is observed. Based on a single observed network, the log-likelihood function is $\ell(\boldsymbol{\beta}) = \boldsymbol{\beta}^\top \mathbf{d} - z(\boldsymbol{\beta})$. The solution to $\nabla z(\boldsymbol{\beta}) = \mathbf{d}$ is the MLE of $\boldsymbol{\beta}$ and $E_\beta(\mathbf{d}) = \nabla z(\boldsymbol{\beta})$ by the property of exponential family (Brown, 1986)]. For convenience, we will suppress the subscript β hereafter. Let $V_n = (v_{ij})_{i,j=1,\ldots,n}$ be the Fisher information matrix of the parameters $\beta_1, \ldots, \beta_n$, which is also the covariance matrix of $\mathbf{d}$. By Yan et al. (2015), we have

$$v_{ij} = \frac{\sum\limits_{0\le k<l\le q-1} (k-l)^2 e^{(k+l)(\beta_i+\beta_j)}}{\left(\sum\limits_{a=0}^{q-1} e^{a(\beta_i+\beta_j)}\right)^2}, \quad j \neq i, \; v_{ii} = \sum_{j\neq i} v_{ij}.$$

V_n is the diagonally dominant matrix with nonnegative entries. Let $U(\boldsymbol{\beta})$ be the score function of the log-likelihood $\ell(\boldsymbol{\beta})$:

$$U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{d} - E(\mathbf{d}),$$

and define

$$T_n(\boldsymbol{\beta}) = U^\top(\boldsymbol{\beta})[V_n(\boldsymbol{\beta})]^{-1} U(\boldsymbol{\beta}).$$

Then $T_n(\boldsymbol{\beta}_0)$ is the score test statistic under the simple null: $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$. We will investigate the asymptotic distribution of $T_n(\boldsymbol{\beta})$ for a general $\boldsymbol{\beta}$ as n goes to infinity. We use the notations T_n, V_n, U instead of $T_n(\boldsymbol{\beta}), V_n(\boldsymbol{\beta}), U(\boldsymbol{\beta})$ hereafter for convenience.

The dimensions of U and V_n will increase with n, which makes the study of the asymptotic distribution of T_n difficult. In order to obtain the asymptotic distribution of T_n, we need to handle the inverse of V_n with a large dimension. Since V_n^{-1} does not have a closed form, we use its approximation $S_n = \mathrm{diag}(1/v_{11}, \ldots, 1/v_{nn})$ proposed by Yan et al. (2015), whose approximate error is given in Proposition 1 of their paper. As a result, T_n is divided into the sum of two parts: the quadratic sum $\sum_i (d_i - E(d_i))^2/v_{ii}$ and a remainder. Therefore, the object is to derive the asymptotic distribution of $\sum_i (d_i - E(d_i))^2/v_{ii}$ and bound the remainder. For any fixed k, the vector $((d_{i_1} - E(d_{i_1}))/v_{i_1,i_1}^{1/2}, \ldots, (d_{i_k} - E(d_{i_k}))/v_{i_k,i_k}^{1/2})$ is asymptotically independent standard normal as shown in Proposition 2 of Yan et al. (2015). In order to obtain the asymptotic distribution of $\sum_i (d_i - E(d_i))^2/v_{ii}$, we consider a more general weighted quadratic sum $\sum_i c_i (d_i - E(d_i))^2$. Let $\tilde{a}_{ij} = a_{ij} - E(a_{ij})$ be the centered random variable of a_{ij} and define $\tilde{a}_{ii} = 0$ for all $i = 1, \ldots, n$. By noting that $Var(\mathbf{d}) = V_n$, the expectation and variance of this sum are given below, the calculation of which is given in the supplementary material (see Appendix B).