Bootstrap Analysis in Network Vector Autoregression Model for Large-Scale Social Network

Manho Hong and Eunju Hwang
Department of Applied Statistics, Gachon University

INTRODUCTION

Large-scale social network services such as Facebook and Twitter have been dominating the online social media. In this work we consider a network vector autoregression model for the social network structure and study its statistical inference using two kinds of bootstrap methods. In the network vector autoregression model with large-size social media users, a continuous response of each user at a given time point is a linear combination of momentum effect, network effect, user-specific effect and independent noise. To analyze the model, a bootstrap version of the ordinary least square estimator is developed by means of the stationary bootstrap and the classical residual bootstrap. Its asymptotic properties and multivariate normal approximation are discussed on theory as well as visualized on a simulation study. For a finite sample validity, root mean square errors of the bootstrap estimates are computed and bootstrap confident intervals are constructed along with their empirical coverage probabilities and average lengths in a Monte-Carlo experiment.

MODEL and MAIN RESULT

❖ Network Vector Auto-Regression Model

To describe relational data in large-scale social network of N users, an adjacency matrix $A \in \mathbb{R}^{N\times N}$, whose entries represent link between pairs of users and Y_{it} which is response of user i at time t (e.g., log-transformed total tweet length) are used.

$$A = (a_{ij}) = \begin{cases} 1, & if\ user\ i\ follows\ user\ j \\ 0, & otherwise. \end{cases}$$

In Zhu et al. (2017), the network vector auto-regression model $\{Y_{it}\}$ is given by

$$Y_{it} = \beta_0 + Z_i^\top \gamma + \beta_1 n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-1)} + \beta_2 Y_{i(t-1)} + \varepsilon_{it}, 1 \le i \le N, 1 \le t \le T \quad (1)$$

$$where\ Z_i = (Z_{i1}, \cdots, Z_{iq})^\top \in \mathbb{R}^q : q\text{-}dimensional\ user\ specific\ random\ vector,$$

$n_i = \sum_{j\neq i} a_{ij} \in \mathbb{R}^1$: $total\ number\ of\ users\ that\ user\ i\ follows$, $\varepsilon_{it} \in \mathbb{R}^1$: $error\ with\ mean\ zero\ variance\ \sigma^2$

$$\gamma = (\gamma_1, \cdots, \gamma_q)^\top \in \mathbb{R}^q : user\ specific\ effect\ associated\ parameters, \beta = (\beta_0, \beta_1, \beta_2)^\top \in \mathbb{R}^3 : parameters.$$

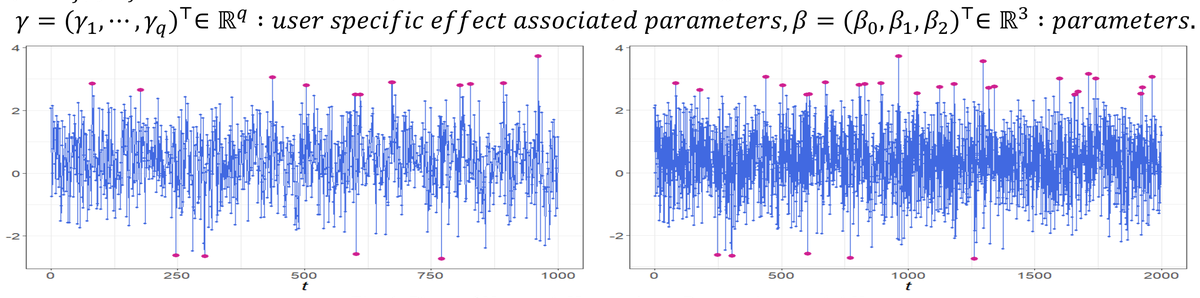


Fig. 1. Plots of Network Vector Auto-Regression model Y_{it}

❖ Ordinary Least Square Estimator(OLSE)

We rewrite (1) as follows:

$$Y_{it} = \beta_0 + Z_i^\top \gamma + \beta_1 w_i^\top \mathbb{Y}_{t-1} + \beta_2 Y_{i(t-1)} + \varepsilon_{it} = X_{i(t-1)}^\top \theta + \varepsilon_{it} \quad (2)$$

where

$$w_i = \left(\frac{a_{i1}}{n_i}, \frac{a_{i2}}{n_i}, \cdots, \frac{a_{iN}}{n_i}\right)^\top \in \mathbb{R}^N,\ W = (w_1, w_1, \cdots, w_N),\ \mathbb{Y}_{t-1} = (Y_{1(t-1)}, Y_{2(t-1)}, \cdots, Y_{N(t-1)})^\top \in \mathbb{R}^N,$$

$$X_{i(t-1)} = (1,\ w_i^\top \mathbb{Y}_{t-1}, Y_{i(t-1)}, Z_i^\top)^\top \in \mathbb{R}^{q+3},\ \theta = (\beta_0, \beta_1, \beta_2, \gamma^\top)^\top.$$

We can finally rewrite (2) in matrix form with $i = 1, 2, \cdots, N$ in rows as

$$\mathbb{Y}_t = \mathbb{X}_{t-1}\theta + \mathcal{E}_t,\ \ where\ \mathbb{X}_t = (X_{1t}, X_{2t}, \cdots, X_{Nt})^\top \in \mathbb{R}^{N\times(q+3)},\ \mathcal{E}_t = (\varepsilon_{1t}, \varepsilon_{2t},\ \cdots, \varepsilon_{Nt})^\top \in \mathbb{R}^N.$$

The OLSE is given by

$$\hat{\theta} = \left(\sum_{t=1}^{T} \mathbb{X}_{t-1}^\top \mathbb{X}_{t-1}\right)^{-1} \left(\sum_{t=1}^{T} \mathbb{X}_{t-1}^\top \mathbb{Y}_t\right) = \left(\sum_{t=1}^{T}\sum_{i=1}^{N} X_{i(t-1)} X_{i(t-1)}^\top\right)^{-1} \left(\sum_{t=1}^{T}\sum_{i=1}^{N} X_{i(t-1)} Y_{it}\right).$$

❖ Monte-Carlo Simulation

To demonstrate the performance of the proposed model, Monte-Carlo simulations are used. The asymptotic properties and normal approximation of OLSE are verified through simulation. Specifically, we follow Nowicki and Snijders (2001), and randomly assign for each user label with equal probability.

$$User\ label\ K\ is\ simulated\ from\ discrete\ uniform\ \{1, 2, \ldots, K_{max}\}$$

$$Adjacency\ matrix\ A\ is\ simulated\ as\ a_{ij} = \begin{cases} 0.3 \times N^{-0.3}, & if\ label\ of\ user\ i\ is\ equal\ with\ label\ of\ user\ j \\ 0.3 \times N^{-1}, & otherwise. \end{cases}$$

$$Random\ error\ \varepsilon_{it}\ is\ simulated\ from\ N(0,1).$$

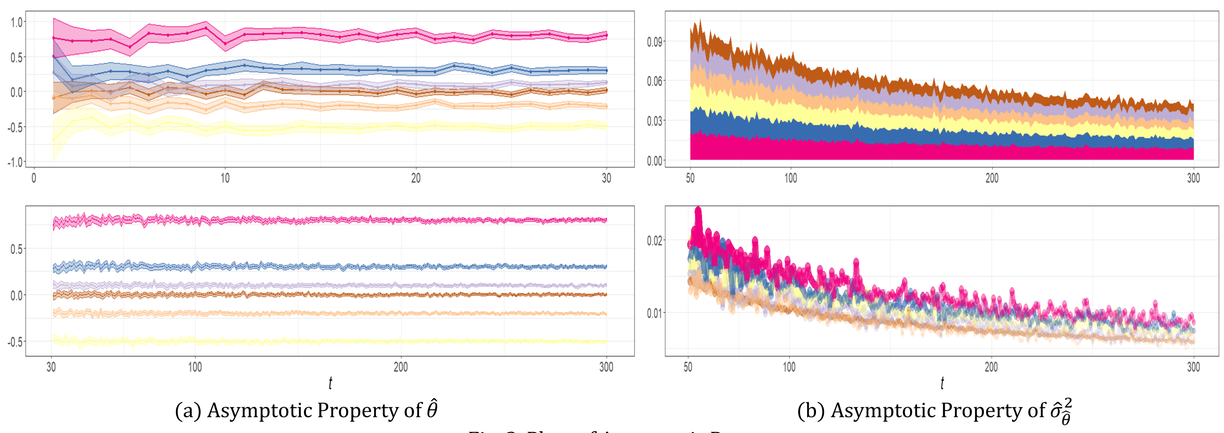
$$User\ specific\ vector\ Z_i = (Z_{i1}, \cdots, Z_{i3})^\top \in \mathbb{R}^3\ is\ simulated\ from\ N_3(\mathbf{0},\ \Sigma_z),$$

$$where\ \Sigma_z = (\sigma_{j_1 j_2}),\ \sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}.$$

$$Initial\ value\ \mathbb{Y}_0\ is\ simulated\ from\ N_N(\mu, \Gamma(0)),$$

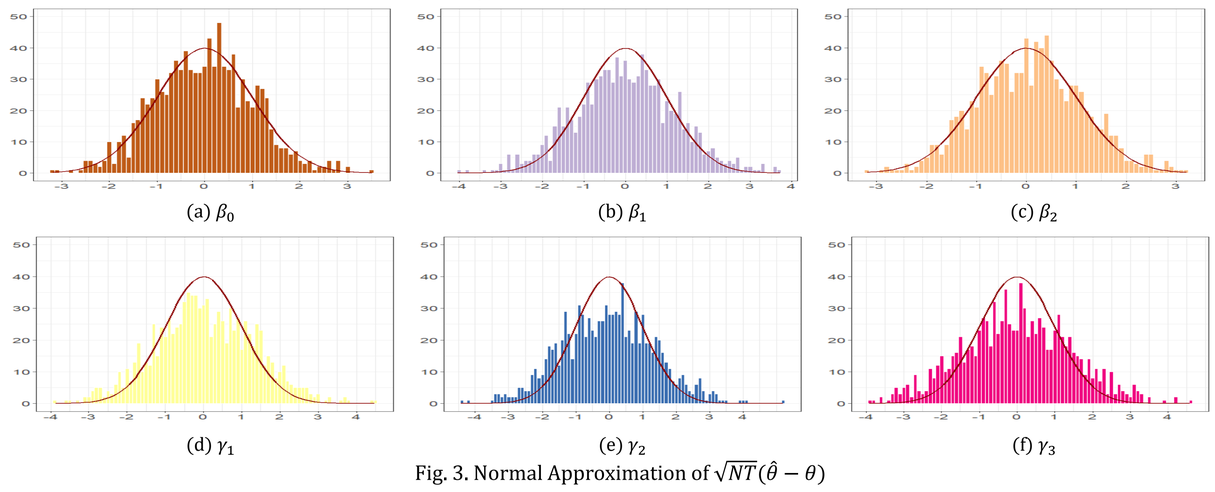
$$where\ \mu = (I - G)^{-1},\quad vec\{\Gamma(0)\} = \sigma^2 (I - G \otimes G)^{-1} vec(I),\quad G = \beta_1 W + \beta_2 I.$$

$$Parameter\ \theta\ is\ fixed\ to\ be\ \theta = (\beta_0, \beta_1, \beta_2, \gamma^\top)^\top = (0, 0.1, -0.2, -0.5, 0.3, 0.8).$$



(a) Asymptotic Property of $\hat{\theta}$ (b) Asymptotic Property of $\hat{\sigma}_{\hat{\theta}}^2$

Fig. 2. Plots of Asymptotic Property



(a) β_0 (b) β_1 (c) β_2

(d) γ_1 (e) γ_2 (f) γ_3

Fig. 3. Normal Approximation of $\sqrt{NT}(\hat{\theta} - \theta)$

Reference

[1] Xuening Z, Rui P, Guodong L, Yuewen L and Hansheng W. (2017), Network Vector AutoRegression, The Annals of Statistics Volume 45, Number 3, 1096-1123.

[2] Nowicki K. and Snijders T. A. B. (2001), Estimation and Prediction for Stochastic Block structures. Journal of the American Statistical Association. Volume 96, 1077–1087.

❖ Bootstrap Estimator and Theorem

The OLSE residuals are computed as

$$\hat{\varepsilon}_{it} = Y_{it} - X_{i(t-1)}^\top \hat{\theta}, \qquad 1 \le i \le N, 1 \le t \le T.$$

Bootstrap version of residual array $\{\hat{\varepsilon}_{it}^*\}$ is generated through its $empirical\ distribution\ function\ \hat{F}_{\mathcal{E}}(\cdot)$:

$$\hat{F}_{\mathcal{E}}(x) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{i=1}^{N} \mathbb{I}_{(-\infty, \hat{\varepsilon}_{it}]}(x).$$

Using the stationary bootstrap sample $\{\mathbb{Y}_t^* = (Y_{1t}^*, \ldots, Y_{Nt}^*)^\top : 0 \le t \le T-1\}$ obtained by

$$Y_{U_1}, Y_{U_1+1}, \cdots, Y_{U_1+L_1-1}, \cdots, Y_{U_k}, Y_{U_k+1}, \cdots, Y_{U_k+L_k-1}, \cdots$$

where $\{U_i : i = 1, 2, \cdots\}$ are $i.i.d.$ with $discrete\ uniform\{1, 2, \ldots, T\}$ with probability $\frac{1}{T}$ and $\{L_i : i = 1, 2, \cdots\}$ are $i.i.d.$ with $geometric\left(p = 0.05 \times \frac{T}{100}^{-\frac{1}{3}}\right)$ the stationary bootstrap version of array $\mathbb{X}_t^*$ is defined as

$$\mathbb{X}_t^* = (X_{1t}^*, \ldots, X_{Nt}^*)^\top\ with\ X_{it}^* = (1,\ w_i^\top \mathbb{Y}_t^*, Y_{it}^*, Z_i^\top)^\top, 1 \le i \le N.$$

We define bootstrap estimator $\hat{\theta}^*$ of θ as

$$\hat{\theta}^* = \left(\sum_{t=1}^{T} \mathbb{X}_{t-1}^{*\top} \mathbb{X}_{t-1}^*\right)^{-1} \left(\sum_{t=1}^{T} \mathbb{X}_{t-1}^{*\top} \tilde{\mathbb{Y}}_t^*\right) = \left(\sum_{t=1}^{T}\sum_{i=1}^{N} X_{i(t-1)}^* X_{i(t-1)}^{*\top}\right)^{-1} \left(\sum_{t=1}^{T}\sum_{i=1}^{N} X_{i(t-1)}^* \tilde{Y}_{it}^*\right)$$

$$where\ \tilde{\mathbb{Y}}_{it}^* = (\tilde{Y}_{1t}^*, \ldots, \tilde{Y}_{Nt}^*)^\top,\ \tilde{Y}_{it}^* = \mathbb{X}_{t-1}^{*\top} \hat{\theta} + \hat{\varepsilon}_{it}^*.$$

Theorem 1. Under the stationary condition $|\beta_1| + |\beta_2| < 1$, as $\min\{N, T\} \to \infty$,

$$\sqrt{NT}(\hat{\theta}^* - \theta) \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1}), \qquad where\ \Sigma = \begin{pmatrix} 1 & c_\beta & c_\beta & \mathbf{0}^\top \\ c_\beta & \Sigma_1 & \Sigma_2 & \kappa_8 \gamma^\top \Sigma_z \\ c_\beta & \Sigma_2 & \Sigma_3 & \kappa_3 \gamma^\top \Sigma_z \\ \mathbf{0} & \kappa_8 \Sigma_z \gamma & \kappa_3 \Sigma_z \gamma & \Sigma_z \end{pmatrix}$$

$c_\beta = \beta_0(1 - \beta_1 - \beta_2)^{-1}, \Sigma_1 = c_\beta^2 + \kappa_5 \gamma^\top \Sigma_z \gamma + \kappa_6, \Sigma_2 = c_\beta^2 + \kappa_7 \gamma^\top \Sigma_z \gamma + \kappa_2, \Sigma_3 = c_\beta^2 + \kappa_4 \gamma^\top \Sigma_z \gamma + \kappa_1,$
$\kappa_1 = N^{-1} tr\{\Gamma(0)\},\ \kappa_2 = N^{-1} tr\{W\Gamma(0)\}, \kappa_3 = N^{-1} tr\{(I - G)^{-1}\}, \kappa_4 = N^{-1} tr\{Q\},\ \kappa_5 = N^{-1} tr\{WQW^\top\},$
$\kappa_6 = N^{-1} tr\{W\Gamma(0)W^\top\}, \kappa_7 = N^{-1} tr\{WQ\}, \kappa_8 = N^{-1} tr\{W(I - G)^{-1}\}.$

Theorem 2. As $\min\{N, T\} \to \infty\ and\ Tp \to \infty$

$$\sup_x \left|P^*\left(\sqrt{NT}\left[\hat{\theta}^* - \hat{\theta}\right] \le x\right) - P\left(\sqrt{NT}\left[\hat{\theta} - \theta\right] \le x\right)\right| \xrightarrow{p} 0.$$

❖ Bootstrap Confidence Interval

According to Theorem 2, we construct a confidence interval for each bootstrap estimator $\hat{\theta}^*$. Let $q_{0.025}^*$ and $q_{0.975}^*$ be the 0.025 and 0.975 quantiles of the bootstrap estimates. The 95% confidence interval based on the bootstrap is constructed by $0.95 = P\left(q_{0.025}^* - \hat{\theta} \le \hat{\theta}^* - \hat{\theta} \le q_{0.975}^* - \hat{\theta}\right)$

$$\approx P\left(q_{0.025}^* - \hat{\theta} \le \hat{\theta} - \theta \le q_{0.975}^* - \hat{\theta}\right) = P\left(2\hat{\theta} - q_{0.975}^* \le \theta \le 2\hat{\theta} - q_{0.025}^*\right).$$

Thus, 95% $bootstrap\ confidence\ interval = [2\hat{\theta} - q_{0.975}^*,\ 2\hat{\theta} - q_{0.025}^*]$. Evaluation of bootstrap estimates via root mean square error, coverage probability and average length are presented.

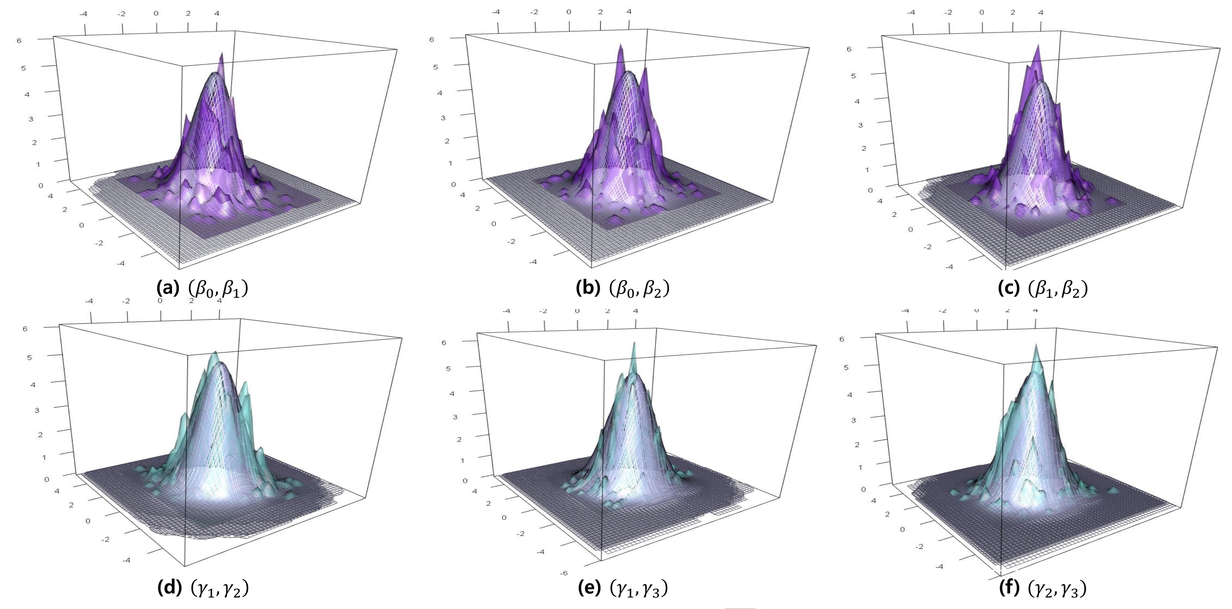
K	T	β_0 = 0	β_1 = 0.1	β_2 = -0.2	γ_1 = -0.5	γ_2 = 0.3	γ_3 = 0.8
5	50	(1.55)	(1.63)	(1.42)	(1.74)	(1.94)	(2.11)
		93.9%	95.9%	96.0%	94.8%	95.3%	95.5%
		[5.77]	[6.49]	[5.64]	[6.86]	[7.44]	[8.43]
	100	(1.00)	(1.23)	(1.01)	(1.29)	(1.41)	(1.48)
		95.8%	94.2%	94.8%	94.4%	95.0%	95.7%
		[4.08]	[4.75]	[3.99]	[4.85]	[5.62]	[5.97]
	200	(0.72)	(0.85)	(0.74)	(0.88)	(0.93)	(1.21)
		95.5%	94.8%	95.5%	95.8%	96.1%	94.7%
		[2.92]	[3.40]	[2.83]	[3.58]	[3.74]	[4.68]
10	50	(1.45)	(1.60)	(1.45)	(1.74)	(1.91)	(2.27)
		95.0%	95.5%	95.5%	94.6%	93.9%	94.1%
		[5.75]	[6.35]	[5.66]	[6.83]	[7.36]	[8.85]
	100	(1.05)	(1.13)	(1.02)	(1.16)	(1.44)	(1.52)
		94.7%	96.3%	95.4%	96.6%	95.6%	94.9%
		[4.05]	[4.51]	[4.00]	[4.74]	[5.81]	[5.98]
	200	(0.74)	(0.93)	(0.73)	(0.91)	(0.96)	(1.01)
		95.5%	95.3%	95.0%	94.6%	93.8%	96.4%
		[2.92]	[3.64]	[2.82]	[3.43]	[3.60]	[4.07]
20	50	(1.44)	(1.81)	(1.42)	(1.60)	(1.82)	(2.08)
		95.4%	94.5%	95.7%	95.3%	96.4%	94.1%
		[5.78]	[6.78]	[5.67]	[6.56]	[7.38]	[7.89]
	100	(1.01)	(1.31)	(1.03)	(1.43)	(1.41)	(1.53)
		95.7%	94.7%	95.2%	94.3%	95.3%	94.1%
		[4.03]	[4.99]	[3.99]	[5.44]	[5.41]	[5.70]
	200	(0.71)	(0.95)	(0.73)	(0.87)	(0.89)	(0.97)
		95.4%	95.3%	94.5%	95.7%	94.5%	94.6%
		[2.88]	[3.86]	[2.83]	[3.40]	[3.52]	[3.81]

K	T	β_0 = 0.1	β_1 = -0.3	β_2 = 0.2	γ_1 = 0.7	γ_2 = 0.1	γ_3 = -0.8
5	50	(1.44)	(1.34)	(1.40)	(2.10)	(1.78)	(2.33)
		95.4%	95.4%	94.8%	95.6%	94.8%	96.8%
		[5.66]	[5.25]	[5.49]	[8.34]	[6.82]	[9.59]
	100	(1.13)	(1.11)	(1.02)	(1.62)	(1.32)	(1.93)
		94.0%	94.9%	94.4%	94.6%	95.8%	94.8%
		[4.15]	[4.31]	[3.88]	[6.16]	[5.42]	[7.35]
	200	(0.72)	(0.68)	(0.69)	(1.10)	(0.96)	(1.25)
		96.3%	95.7%	95.7%	95.1%	94.4%	95.5%
		[2.85]	[2.62]	[2.73]	[4.29]	[3.72]	[4.80]
10	50	(1.52)	(1.41)	(1.38)	(2.08)	(2.01)	(2.35)
		94.1%	94.4%	95.3%	94.1%	93.4%	95.4%
		[5.85]	[5.36]	[5.48]	[8.18]	[7.64]	[9.22]
	100	(1.05)	(0.88)	(0.99)	(1.54)	(1.42)	(1.59)
		94.6%	94.2%	94.9%	95.2%	94.4%	94.2%
		[4.04]	[3.46]	[3.86]	[6.19]	[5.41]	[6.14]
	200	(0.75)	(0.71)	(0.69)	(0.99)	(0.95)	(1.14)
		94.9%	94.5%	95.1%	94.6%	94.9%	94.5%
		[2.95]	[2.78]	[2.74]	[3.93]	[3.76]	[4.47]
20	50	(1.59)	(1.58)	(1.38)	(2.05)	(1.80)	(2.37)
		93.7%	95.2%	95.8%	96.8%	94.3%	95.5%
		[5.92]	[6.36]	[5.53]	[8.38]	[6.94]	[9.19]
	100	(1.01)	(0.93)	(1.01)	(1.51)	(1.28)	(1.62)
		95.4%	94.2%	95.0%	94.0%	95.8%	94.4%
		[4.06]	[3.71]	[3.88]	[5.84]	[5.07]	[6.24]
	200	(0.76)	(0.81)	(0.69)	(1.04)	(1.03)	(1.16)
		94.4%	94.4%	96.6%	95.1%	94.8%	96.2%
		[2.85]	[3.05]	[2.77]	[4.19]	[3.97]	[4.65]

Table 1. (RMSE$\times 10^2$) of Bootstrap Estimates, Coverage Probability(%) and [Average Length$\times 10^2$].
# of Replications = 1000, # of Bootstrap Samples = 1000.

❖ Asymptotic Normality of Bootstrap Estimates

According to Theorem 1, $\sqrt{NT}(\hat{\theta}^* - \theta)$ approximates to the multivariate normal distribution. To verify this, three pairs of bivariate normalized approximations of beta and gamma, respectively, via Monte-Carlo simulations are presented.



(a) (β_0, β_1) (b) (β_0, β_2) (c) (β_1, β_2)

(d) (γ_1, γ_2) (e) (γ_1, γ_3) (f) (γ_2, γ_3)

Fig. 4. Normal Approximation of $\sqrt{NT}(\hat{\theta}^* - \theta)$

저자 : 홍만호, 가천대학교 응용통계학과 학부생, ghdaksgh123@naver.com
황은주, 가천대학교 응용통계학과 조교수, ehwang@gachon.ac.kr
사사 : 본 연구는 가천대학교 캡스톤 디자인 지원을 받아 수행되었음.

아름다운 인재의 샘 가천대학교