Self-normalized Cramér moderate deviations for a supercritical Galton-Watson process

Xiequan FAN

fanxiequan@hotmail.com

Joint work with Qi-Man SHAO (Southern University of Science and Technology)

Center for Applied Mathematics at Tianjin University

ECODEP Conference, 15 Sept. 2021

Outline

1. A supercritical Galton-Watson process
2. Time type self-normalized Cramér moderate deviations
3. Space type self-normalized Cramér moderate deviations
4. Applications to statistics
5. Future work

Outline

A Galton-Watson process (G-W process) can be used to describe the growth of population. It can be described as follows

$$Z_0 = 1, \quad Z_1 = X_{1,1}, \ \ldots, \ Z_{n+1} = \sum_{i=1}^{Z_n} X_{n,i}, \quad \text{for } n \geq 1. \tag{1.1}$$

where $X_{n,i}$ is the offspring number of the i-th individual of the generation n. Moreover, the random variables $(X_{n,i})_{i \geq 1}$ are independent of each other with common distribution law: for all n and i,

$$\mathbb{P}(X_{n,i} = k) = p_k, \quad k \geq 0,$$

and $(X_{n,i})_{i \geq 1}$ are also independent of Z_n.

Estimate the average offspring number of an individual:

$$m = \mathbb{E}Z_1 = \mathbb{E}X_{n,i} = \sum_{k=0}^{\infty} kp_k, \qquad n, i \geq 1.$$

Denote v the standard variance of Z_1, that is

$$v^2 = \mathbb{E}(Z_1 - m)^2. \tag{1.2}$$

To avoid triviality, assume that $v > 0$.

We assume that the set of extinction of the process $(Z_k)_{k\geq 0}$ is negligible. A typical case is $m > 1$.

An example for $m > 1$:

Example: An infectious disease model $(Z_n)_{n\geq 0}$ may be described as follows:

$$Z_0 = 1, \quad Z_{n+1} = Z_n + \sum_{i=1}^{Z_n} Y_{n,i}, \quad \text{for } n \geq 0, \tag{1.3}$$

where Z_n stands for the total population of patients with infectious disease at time n, and $Y_{n,i}$ is the number of patients infected by the i-th individual of Z_n in a unit time (for instance, one day). Taking

$$X_{n,i} = 1 + Y_{n,i},$$

then we have $m > 1$. [Attention: not fit for Covid-19.]

Time type data:
$(Z_k)_{n_0 \leq k \leq n_0+n}$ can be observed.

Space type data:
$(X_{n,i})_{1 \leq i \leq Z_n}$ can be observed for some given n.

Outline

Time type data: $(Z_k)_{n_0 \leq k \leq n_0+n}$ can be observed.

Define the following time type self-normalized process for the Lotka-Nagaev estimator Z_{k+1}/Z_k:

$$M_{n_0,n} = \frac{\sum_{k=n_0}^{n_0+n-1} \sqrt{Z_k}(\frac{Z_{k+1}}{Z_k} - m)}{\sqrt{\sum_{k=n_0}^{n_0+n-1} Z_k(\frac{Z_{k+1}}{Z_k} - m)^2}}. \tag{2.1}$$

We have the following time type self-normalized Cramér moderate deviations.

Theorem 1

Assume that $\mathbb{E}Z_1^{2+\rho} < \infty$ *for some* $\rho \in (0, 1]$.

[i] If $\rho \in (0, 1)$, *then for all* $x \in [0,\ o(\sqrt{n}))$,

$$\left| \ln \frac{\mathbb{P}(M_{n_0,n} \geq x)}{1 - \Phi(x)} \right| \leq C_\rho \left(\frac{x^{2+\rho}}{n^{\rho/2}} + \frac{(1+x)^{1-\rho(2+\rho)/4}}{n^{\rho(2-\rho)/8}} \right). \qquad (2.2)$$

[ii] If $\rho = 1$, *then for all* $x \in [0,\ o(\sqrt{n}))$,

$$\left| \ln \frac{\mathbb{P}(M_{n_0,n} \geq x)}{1 - \Phi(x)} \right| \leq C \left(\frac{x^3}{\sqrt{n}} + \frac{\ln n}{\sqrt{n}} + \frac{(1+x)^{1/4}}{n^{1/8}} \right). \qquad (2.3)$$

Moreover, the same inequalities remain valid when $M_{n_0,n}$ *is replaced by* $-M_{n_0,n}$.

Corollary 2

Assume that $\mathbb{E}Z_1^{2+\rho} < \infty$ *for some* $\rho \in (0, 1]$. *It holds*

$$\frac{\mathbb{P}(M_{n_0,n} \geq x)}{1 - \Phi(x)} = 1 + o(1) \qquad (2.4)$$

uniformly for $n_0 \in \mathbb{N}$ *and for* $x \in [0,\ o(n^{\rho/(4+2\rho)}))$ *as* $n \to \infty$.

Advantages: $\mathbb{E}Z_1^{2+\rho} < \infty; \Phi(x); v^2$.

Maximum likelihood method:

$$M_{n_0,n} = \frac{\sum_{k=n_0}^{n_0+n-1} \sqrt{Z_k}(\frac{Z_{k+1}}{Z_k} - m)}{\sqrt{\sum_{k=n_0}^{n_0+n-1} Z_k(\frac{Z_{k+1}}{Z_k} - m)^2}} = 0 \Longleftrightarrow \sum_{k=n_0}^{n_0+n-1} \sqrt{Z_k}(\frac{Z_{k+1}}{Z_k} - m) = 0,$$

which implies that

$$\overline{m}_n := \frac{1}{\sum_{k=n_0}^{n_0+n-1} \sqrt{Z_k}} \sum_{k=n_0}^{n_0+n-1} \sqrt{Z_k}(\frac{Z_{k+1}}{Z_k})$$

can be regarded as a random weighted Lotka-Nagaev estimator of m.

Proof for Time type self-normalized CMD:

Denote

$$\xi_{k+1} = \sqrt{Z_k}(Z_{k+1}/Z_k - m), \quad (m > 1)$$

$\mathfrak{F}_{n_0} = \{\emptyset, \Omega\}$ and $\mathfrak{F}_{k+1} = \sigma\{Z_i : n_0 \le i \le k+1\}$ for all $k \ge n_0$. Notice that $X_{k,i}$ is independent of Z_k. Then it is easy to verify that if $\mathbb{E}Z_1^{2+\rho} < \infty$, then

$$\begin{aligned} \mathbb{E}[\xi_{k+1}|\mathfrak{F}_k] &= 0, \\ \mathbb{E}[\xi_{k+1}^2|\mathfrak{F}_k] &= \mathrm{Var}(Z_1), \\ \mathbb{E}[|\xi_{k+1}|^{2+\rho}|\mathfrak{F}_k] &\le C_{\rho,\mathrm{Var}(Z_1),\mathbb{E}Z_1^{2+\rho}}\mathbb{E}[\xi_{k+1}^2|\mathfrak{F}_k]. \end{aligned}$$

CMD for self-normalized martingales

Assume that $(\xi_i, \mathcal{F}_i)_{i\geq 1}$ is a sequence of martingale differences satisfying the following conditions:

(A1) There exist two positive constants ρ and H such that

$$\mathbb{E}[|\xi_i|^{2+\rho}|\mathcal{F}_{i-1}] \leq H^\rho\, \mathbb{E}[\xi_i^2|\mathcal{F}_{i-1}] \ \text{ for all } \ 1 \leq i \leq n;$$

(A2) There exist two constants $\sigma > 0$ and H such that

$$\| \sum_{i=1}^{n} \mathbb{E}[\xi_i^2|\mathcal{F}_{i-1}] - n\sigma^2 \|_\infty \leq H.$$

Denote by $S_n = \sum_{i=1}^{n} \xi_i, \ \ V^2 = \sum_{i=1}^{n} \xi_i^2$. And self-normalized martingale is defined by S_n/V_n.

CMD for Self-normalized martingales in F., Grama, Liu and Shao (Bernoulli, 2019)

Lemma 3

Assume conditions (A1) and (A2).

- *If* $\rho \in (0,1)$, *then for all* $0 \leq x = o(\sqrt{n})$,

$$\left| \ln \frac{\mathbb{P}(S_n/V_n > x)}{1-\Phi(x)} \right| \leq C_\rho \left(\frac{x^{2+\rho}}{n^{\rho/2}} + \frac{(1+x)^{1-\rho(2+\rho)/4}}{n^{\rho(2-\rho)/8}} \right). \tag{2.5}$$

- *If* $\rho = 1$, *then for all* $0 \leq x = o(\sqrt{n})$,

$$\left| \ln \frac{\mathbb{P}(S_n/V_n > x)}{1-\Phi(x)} \right| \leq C \left(\frac{x^3}{\sqrt{n}} + \frac{\ln n}{\sqrt{n}} + \frac{(1+x)^{1/4}}{n^{1/8}} \right).$$

As

$$M_{n_0,n} = \frac{\sum_{k=n_0}^{n_0+n-1} \xi_{k+1}}{\sqrt{\sum_{k=n_0}^{n_0+n-1} \xi_{k+1}^2}} \tag{2.6}$$

is a self-normalized martingale. Applying CMD for self-normalized martingales to $(\xi_k, \mathfrak{F}_k)_{n_0 \leq k \leq n_0+n-1}$, we obtain the time type self-normalized CMD.

Outline

Space type data: $(X_{n,i})_{1\leq i\leq Z_n}$ can be observed for some given n.

[Athreya, AAP, 1994] LDP:
$p_1 > 0$ and $\mathbb{E}e^{cZ_1} < \infty, c > 0$, $\mathbb{P}(|\frac{Z_{n+1}}{Z_n} - m| \geq \varepsilon)$

[He, Adv. AP, 2016] LDP:
$p_1 > 0$ and $\mathbb{P}(Z_1 > x) \sim x^{-\beta}L(x), 1 < \beta < 2$, $\mathbb{P}(\frac{Z_{n+1}}{Z_n} - m \geq \varepsilon)$

[Chu, JAP, 2018] Self-normalized LDP:
$p_0 = 0, p_1 > 0$, and $\mathbb{E}Z_1 < \infty$, $\mathbb{P}\left(\frac{\sqrt{Z_n}}{\sqrt{\sum_{i=1}^{Z_n}(X_{n,i}-\frac{Z_{n+1}}{Z_n})^2}}|\frac{Z_{n+1}}{Z_n} - m| \geq \varepsilon\right)$

Denote

$$T_n = \frac{Z_n \left(\frac{Z_{n+1}}{Z_n} - m \right)}{\sqrt{\sum_{i=1}^{Z_n} (X_{n,i} - \frac{Z_{n+1}}{Z_n})^2}} = \frac{\frac{\sum_{i=1}^{Z_n} X_{n,i} - m}{Z_n}}{\sqrt{\frac{1}{Z_n} \sum_{i=1}^{Z_n} (X_{n,i} - \frac{Z_{n+1}}{Z_n})^2}} \sqrt{Z_n}$$

the space type self-normalized process for the Lotka-Nagaev estimator.

For T_n, we have the following space type self-normalized CMD.

Theorem 4

Assume either $p_0 = 0$ *or* $p_1 > 0$, *and* $\mathbb{E}Z_1^{2+\rho} < \infty$ *for some* $\rho \in (0, 1]$. *Then*

$$\left| \ln \frac{\mathbb{P}(T_n \geq x)}{1 - \Phi(x)} \right| = O\Big(\frac{1 + x^{2+\rho}}{n^{\rho/2}} \Big) \qquad (3.1)$$

uniformly for $x \in [0, o(\sqrt{n}))$ *as* $n \to \infty$. *Moreover, the same equality remains valid when* T_n *is replaced by* $-T_n$.

Corollary 5

Assume the conditions of Theorem 4. Then

$$\frac{\mathbb{P}\big(T_n \geq x\big)}{1-\Phi(x)} = 1 + O\Big(\frac{1+x^{2+\rho}}{n^{\rho/2}}\Big) \tag{3.2}$$

uniformly for $x \in [0,\, O(n^{\rho/(4+2\rho)}))$ *as* $n \to \infty$*. In particular, it implies that*

$$\frac{\mathbb{P}\big(T_n \geq x\big)}{1-\Phi(x)} = 1 + o(1) \tag{3.3}$$

uniformly for $x \in [0,\, o(n^{\rho/(4+2\rho)}))$ *as* $n \to \infty$*. Moreover, the same equalities remain valid when* T_n *is replaced by* $-T_n$*.*

From Theorem 4, we get the following Berry-Esseen bound for T_n.

Corollary 6

Assume the conditions of Theorem 4. Then

$$\sup_{x\in\mathbb{R}} \Big|\mathbb{P}(T_n \leq x) - \Phi(x)\Big| \leq \frac{C_\rho}{n^{\rho/2}}, \tag{3.4}$$

where C_ρ does not depend on n.

Proof for Space type self-normalized CMD:

$$
\begin{aligned}
\mathbb{P}\Big(T_n \geq x\Big) &= \mathbb{P}\left(\sum_{i=1}^{Z_n}(X_{n,i}-m) \geq x\sqrt{(X_{n,i}-\frac{Z_{n+1}}{Z_n})^2}\right) \\
&= \sum_{k=1}^{\infty}\mathbb{P}(Z_n=k)\mathbb{P}\left(\sum_{i=1}^{k}(X_{n,i}-m) \geq x\sqrt{(X_{n,i}-\frac{Z_{n+1}}{k})^2}\right) \\
&= \sum_{k=1}^{\infty}\mathbb{P}(Z_n=k)\mathbb{P}\left(\sum_{i=1}^{k}(X_{n,i}-m) \geq x\sqrt{(X_{n,i}-\overline{X}_{n,.})^2}\right) \\
&=: \sum_{k=1}^{\infty}\mathbb{P}(Z_n=k)I_k(x) \\
&= \sum_{k=1}^{n}\mathbb{P}(Z_n=k)I_k(x) + \sum_{k=n+1}^{\infty}\mathbb{P}(Z_n=k)I_k(x).
\end{aligned}
$$

Lemma 7 (Athreya, AAP, 1994)

If $p_1 > 0$, *then*

$$\lim_{n\to\infty} \frac{\mathbb{E}[s^{Z_n}]}{p_1^n} = Q(s), \quad 0 \le s < 1. \tag{3.5}$$

Set $s_0 = \frac{1+p_1}{2} \in (0,1)$.

$$\begin{aligned}
\sum_{k=1}^{n} \mathbb{P}(Z_n = k) I_k(x) &\le \sum_{k=1}^{n} \mathbb{P}(Z_n = k) = \mathbb{P}(Z_n \le n) \\
&= \mathbb{P}(s_0^{Z_n} \ge s_0^n) \\
&\le s_0^{-n} \mathbb{E}[s_0^{Z_n}] \\
&\le CQ(s_0)(\frac{p_1}{s_0})^n \\
&= C_1 \exp\{-nc_0\}.
\end{aligned} \tag{3.6}$$

Lemma 8 (Jing, Shao, Wang, AoP, 2003)

When $k \geq n$, it holds

$$\left| \ln \frac{I_k(x)}{1 - \Phi(x)} \right| \leq C_\rho \frac{1 + x^{2+\rho}}{n^{\rho/2}} \tag{3.7}$$

uniformly for $0 \leq x = o(\sqrt{n})$ as $n \to \infty$.

For $0 \leq x = o(\sqrt{n})$, we have

$$\begin{aligned} \mathbb{P}\Big(T_n \geq x\Big) &= \sum_{k=1}^{n-1} \mathbb{P}(Z_n = k) I_k(x) + \sum_{k=n}^{\infty} \mathbb{P}(Z_n = k) I_k(x) \\ &\leq C_1 \exp\{-nc_0\} + \Big(1 - \Phi(x)\Big) \exp\left\{C_\rho \frac{1 + x^{2+\rho}}{n^{\rho/2}}\right\} \\ &\leq \Big(1 - \Phi(x)\Big) \exp\left\{C'_\rho \frac{1 + x^{2+\rho}}{n^{\rho/2}}\right\}. \end{aligned}$$

For $0 \leq x = o(\sqrt{n})$, we have

$$
\begin{aligned}
\mathbb{P}\Big(T_n \geq x\Big) &\geq \sum_{k=n}^{\infty} \mathbb{P}(Z_n = k) I_k(x) \\
&\geq \mathbb{P}(Z_n \geq n)\Big(1 - \Phi(x)\Big) \exp\Big\{ - C_\rho \frac{1 + x^{2+\rho}}{n^{\rho/2}} \Big\} \\
&= \Big(1 - \mathbb{P}(Z_n < n)\Big)\Big(1 - \Phi(x)\Big) \exp\Big\{ - C_\rho \frac{1 + x^{2+\rho}}{n^{\rho/2}} \Big\} \\
&\geq \Big(1 - C_1 \exp\{-nc_0\}\Big)\Big(1 - \Phi(x)\Big) \exp\Big\{ - C_\rho \frac{1 + x^{2+\rho}}{n^{\rho/2}} \Big\} \\
&\geq \Big(1 - \Phi(x)\Big) \exp\Big\{ - C'_\rho \frac{1 + x^{2+\rho}}{n^{\rho/2}} \Big\}.
\end{aligned}
$$

Outline

Application to p-value for hypothesis testing:

The data $(Z_k)_{n_0 \leq k \leq n_0+n}$ can be observed. Let $(z_k)_{k=n_0,\ldots,n_0+n}$ be an observation of $(Z_k)_{n_0 \leq k \leq n_0+n}$. In order to estimate the offspring mean m, the observations for Harris' estimator and weighted estimator are respectively

$$\widehat{m}_n = \frac{\sum_{k=n_0}^{n_0+n-1} z_{k+1}}{\sum_{k=n_0}^{n_0+n-1} z_k} \quad \text{and} \quad \widehat{m}_n = \frac{1}{\sum_{k=n_0}^{n_0+n-1} \sqrt{z_k}} \sum_{k=n_0}^{n_0+n-1} \sqrt{z_k}\left(\frac{z_{k+1}}{z_k}\right).$$

The p-value $\mathbb{P}(M_{n_0,n} > |\widetilde{m}_n|)$, by Theorem 1, is almost equal to $2(1 - \Phi(|\widetilde{m}_n|))$, where

$$\widetilde{m}_n = \frac{\sum_{k=n_0}^{n_0+n-1} \sqrt{z_k}(z_{k+1}/z_k - \widehat{m}_n)}{\sqrt{\sum_{k=n_0}^{n_0+n-1} z_k(z_{k+1}/z_k - \widehat{m}_n)^2}}.$$

Construction of confidence interval for m:

The time type data $(Z_k)_{n_0 \leq k \leq n_0+n}$ can be observed. We prove that

$$\frac{\mathbb{P}(M_{n_0,n} \geq x)}{1-\Phi(x)} = 1+o(1) \quad \text{and} \quad \frac{\mathbb{P}(M_{n_0,n} \leq -x)}{1-\Phi(x)} = 1+o(1)$$

uniformly for $0 \leq x = o(n^{\rho/(4+2\rho)})$.

When $\Phi^{-1}(1-\kappa_n/2) = o(n^{\rho/(4+2\rho)})$, $(1-\kappa_n)$-confidence intervals for m is given by the following equations:

$$-\Phi^{-1}(1-\kappa_n/2) \leq M_{n_0,n} = \frac{\sum_{k=n_0}^{n_0+n-1} \sqrt{Z_k}(Z_{k+1}/Z_k - m)}{\sqrt{\sum_{k=n_0}^{n_0+n-1} Z_k(Z_{k+1}/Z_k - m)^2}} \leq \Phi^{-1}(1-\kappa_n/2).$$

$$\Phi^{-1}(1-\kappa_n/2) = o(n^{\rho/(4+2\rho)}) \iff \left|\ln \kappa_n\right| = o\big(n^{\rho/(2+\rho)}\big).$$

The space type data $(X_{n,i})_{1\leq i\leq Z_n}$ can be observed. Corollary 5 implies that

$$\frac{\mathbb{P}(T_n \geq x)}{1-\Phi\left(x\right)} = 1+o(1) \quad \text{and} \quad \frac{\mathbb{P}(T_n \leq -x)}{1-\Phi\left(x\right)} = 1+o(1)$$

uniformly for $0 \leq x = o(n^{\rho/(4+2\rho)})$.

When $\Phi^{-1}(1-\kappa_n/2) = o(n^{\rho/(4+2\rho)})$, $(1-\kappa_n)$-confidence intervals for m is given by the following equations:

$$-\Phi^{-1}(1-\kappa_n/2) \leq T_n =: \frac{Z_n\left(\frac{Z_{n+1}}{Z_n}-m\right)}{\sqrt{\sum_{i=1}^{Z_n}(X_{n,i}-\frac{Z_{n+1}}{Z_n})^2}} \leq \Phi^{-1}(1-\kappa_n/2).$$

$$\Phi^{-1}(1-\kappa_n/2) = o\big(n^{\rho/(4+2\rho)}\big) \iff \big|\ln\kappa_n\big| = o\big(n^{\rho/(2+\rho)}\big).$$

Outline

Martingale differences $\left(\sqrt{Z_k}(\frac{Z_{k+1}}{Z_k} - m), \mathcal{F}_{k+1}\right)_{k \geq n_0}$:

Normalized CMD; Concentration inequalities;
Erdös-Kac's theorem; Donsker's theorem ...

Three open questions:

Time and Space type data: $(X_{k,i})_{0 \leq k \leq n,\ 1 \leq i \leq Z_k}$ can be observed;

Some data of $(Z_k)_{k \geq 0}$ are missing. Hint: $\frac{Z_{k+l}}{Z_k} \approx m^l$.

Can we establish similar result for BPRE.

Thank you for your attention!