Cramér Type Moderate Deviations for Self-normalized Processes

Qi-Man Shao

Hong Kong University of Science and Technology

Chinese University of Hong Kong (after September 1, 2012)

This talk is based on joint work with Wenxin Zhou and Weidong Liu.

Outline

1. Introduction and motivation
2. Classical moderate deviation vs the self-normalized moderate deviation
3. Studentized non-linear statistics
4. Studentized U-statistics
5. Hotelling's T^2 statistic

Let W_n be a random variable of interest.

- Aim: Estimate $P(W_n \geq x)$.

Let W_n be a random variable of interest.

- Aim: Estimate $P(W_n \geq x)$.
- Suppose that $W_n \xrightarrow{d.} Y$. A common practice is:
  - Use $P(Y \geq x)$ to estimate $P(W_n \geq x)$.

Let W_n be a random variable of interest.

- Aim: Estimate $P(W_n \geq x)$.
- Suppose that $W_n \xrightarrow{d.} Y$. A common practice is:
  - Use $P(Y \geq x)$ to estimate $P(W_n \geq x)$.
- Question:

What is the error of approximation?

- Two types of error:
  1. Absolute error: Berry-Esseen type bound

$$|P(W_n \geq x) - P(Y \geq x)| = \text{error}$$

- Two types of error:

1. Absolute error: Berry-Esseen type bound

$$|P(W_n \geq x) - P(Y \geq x)| = \text{error}$$

2. Relative error: Cramér type moderate deviation

$$\frac{P(W_n \geq x)}{P(Y \geq x)} = 1 + \text{error}$$

- Two types of error:

1. Absolute error: Berry-Esseen type bound

$$|P(W_n \geq x) - P(Y \geq x)| = \text{error}$$

2. Relative error: Cramér type moderate deviation

$$\frac{P(W_n \geq x)}{P(Y \geq x)} = 1 + \text{error}$$

- Our focus: Relative error, Cramér type moderate deviation, especially, what is the largest possible a_n such that

$$\frac{P(W_n \geq x)}{P(Y \geq x)} \to 1$$

holds uniformly in $x \in [0, a_n]$.

- Two types of error:
  1. Absolute error: Berry-Esseen type bound
  $$|P(W_n \geq x) - P(Y \geq x)| = \text{error}$$
  2. Relative error: Cramér type moderate deviation
  $$\frac{P(W_n \geq x)}{P(Y \geq x)} = 1 + \text{error}$$

- Our focus: Relative error, Cramér type moderate deviation, especially, what is the largest possible a_n such that
$$\frac{P(W_n \geq x)}{P(Y \geq x)} \to 1$$
holds uniformly in $x \in [0, a_n]$. Why?

- In many applications, $P(Y \geq x)$ itself is very small. Only when the relative error is small, can $P(W_n \geq x)$ be approximated by $P(Y \geq x)$;

- In many applications, $P(Y \geq x)$ itself is very small. Only when the relative error is small, can $P(W_n \geq x)$ be approximated by $P(Y \geq x)$;

BBC News - Higgs boson-like particle discovery claimed at LHC http://www.bbc.co.uk/news/worl

BBC NEWS

SCIENCE & ENVIRONMENT

4 July 2012 Last updated at 07:35 GMT

Higgs boson-like particle discovery claimed at LHC

COMMENTS (1665)

By Paul Rincon
Science editor, BBC News website, Geneva

Cern scientists reporting from the Large Hadron Collider (LHC) have claimed the discovery of a new particle consistent with the Higgs boson.

The particle has been the subject of a 45-year hunt to explain how matter attains its mass.

Both of the Higgs boson-hunting experiments at the LHC see a level of certainty in their data worthy of a "discovery".

More work will be needed to be certain that what they see is a Higgs, however.

The results announced at Cern (European Organization for Nuclear Research), home of the LHC in Geneva, were met with loud applause and cheering.

Prof Peter Higgs, after whom the particle is named, wiped a tear from his eye as the teams finished their presentations in the Ce auditorium.

"I would like to add my congratulations to everyone involved in this achievement," he added later.

"It's really an incredible thing that it's happened in my lifetime."

loud applause and cheering.

Prof Peter Higgs, after whom the particle is named, wiped a tear from his eye as the teams finished their presentations in the C auditorium.

"I would like to add my congratulations to everyone involved in this achievement," he added later.

"It's really an incredible thing that it's happened in my lifetime."

Prof Stephen Hawking joined in with an opinion on a topic often discussed in hushed tones.

"This is an important result and should earn Peter Higgs the Nobel Prize," he told BBC News.

"But it is a pity in a way because the great advances in physics have come from experiments that gave results we didn't expec

'Dramatic'

The CMS team claimed they had seen a "bump" in their data corresponding to a particle weighing in at 125.3 gigaelectronvolts (GeV) - about 133 times heavier than the protons that lie at the heart of every atom.

They claimed that by combining two data sets, they had attained a confidence level just at the "five-sigma" point - about a one-in-3.5 million chance that the signal they see would appear if there were no Higgs particle.

However, a full combination of the CMS data brings that number just back to 4.9 sigma - a one-in-two million chance.

Prof Joe Incandela, spokesman for the CMS, was unequivocal: "The results are preliminary but the five-sigma signal at aroun 125 GeV we're seeing is dramatic. This is indeed a new particle," he told the Geneva meeting.

Atlas results were even more promising, at a slightly higher mass: "We observe in our data clear signs of a new particle, at the level of five sigma, in the mass region around 126 GeV," said Dr Fabiola Gianotti, spokeswoman for the Atlas experiment at th LHC.

- Multiple hypothesis tests

  Consider the problem of testing simultaneously m (null) hypotheses, $H_1, H_2, \cdots, H_m$, of which m_0, are true. Let R be the number of hypotheses rejected. Table below summarizes the test results

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

- The proportion of errors committed by falsely rejecting null hypotheses: V/R

- False discovery rate (FDR): $E(V/R)$

- Benjamini-Hochberg FDR controlling procedure:
  Assume P-values are $p_1, p_2, \ldots, p_m$. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the ordered p-values, and denote by $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. Let

  $$k = \max\{i : \ p_{(i)} \leq \frac{i}{m}\alpha\}$$

  where $0 < \alpha < 1$. Then reject all $H_{(i)}$, for $1 \leq i \leq k$.

  If the test statistics are independent, then $E(V/R) \leq \alpha$.

- P-values are usually unknown, need to be estimated.

- Question: (Fan, Hall, Yao (2007))
  How large m can be before the accuracy of the estimated P-values becomes poor?

- P-values are usually unknown, need to be estimated.

- Question: (Fan, Hall, Yao (2007))
  How large m can be before the accuracy of the estimated P-values becomes poor?

- Korosok-Ma (2007), Fan, Hall, Yao (2007), Liu and Shao (2009), Shao (2010):

  - Let $T_{n,i}$ be the test statistic for H_i. Assume that the true P-value is $p_i = P(T_{n,i} \geq t_{n,i})$ and that there exist $a_{n,i}$ and functions f_i such that
  $$\max_{1 \leq i \leq m} \sup_{0 \leq x \leq a_{n,i}} \left| \frac{P(T_{n,i} \geq x)}{f_i(x)} - 1 \right| = o(1)$$
  as $n \to \infty$. If $m \leq \alpha / (2 \max_{1 \leq i \leq m} f_i(a_{n,i}))$, then the FDR is controlled at level α when it is based on the estimated P-values $\hat{p}_i = f_i(t_{n,i})$.

2. Classical moderate deviation vs self-normalized moderate deviation

Let $X, X_1, X_2, \cdots, X_n$ be independent identically distributed (i.i.d.) random variables and let

$$S_n = \sum_{i=1}^n X_i, \ \ V_n^2 = \sum_{i=1}^n X_i^2.$$

Assume $EX = 0$ and $\sigma^2 = EX^2 < \infty$.

Standardized sum: $S_n/(\sigma\sqrt{n})$

Self-normalized sum: S_n/V_n

2. Classical moderate deviation vs self-normalized moderate deviation

Let $X, X_1, X_2, \cdots, X_n$ be independent identically distributed (i.i.d.) random variables and let

$$S_n = \sum_{i=1}^{n} X_i, \quad V_n^2 = \sum_{i=1}^{n} X_i^2.$$

Assume $EX = 0$ and $\sigma^2 = EX^2 < \infty$.

Standardized sum: $S_n/(\sigma\sqrt{n})$

Self-normalized sum: S_n/V_n

Remark: The student t-statistic T_n and the self-normalized sum have a close relationship

$$T_n = \frac{S_n}{V_n}\left(\frac{n-1}{n-(S_n/V_n)^2}\right)^{1/2}, \quad \{T_n \geq t\} = \left\{\frac{S_n}{V_n} \geq t\left(\frac{n}{n+t^2-1}\right)^{1/2}\right\}$$

▶ Classical Cramér moderate deviation

- If $Ee^{t_0\sqrt{|X|}} < \infty$ for $t_0 > 0$, then

$$\frac{P(S_n/(\sigma\sqrt{n}) \geq x)}{1-\Phi(x)} \to 1$$

uniformly in $x \in [0, o(n^{1/6}))$. Moreover,

$$\frac{P(S_n/(\sigma\sqrt{n}) \geq x)}{1-\Phi(x)} = 1 + O(1)\frac{(1+x^3)}{\sqrt{n}}$$

- Self-normalized moderate deviation
  - Shao (1999): If $E|X|^3 < \infty$, then

$$\frac{P(S_n/V_n \geq x)}{1 - \Phi(x)} \to 1$$

uniformly in $x \in [0, o(n^{1/6}))$.

- Self-normalized moderate deviation
  - Shao (1999): If $E|X|^3 < \infty$, then

$$\frac{P(S_n/V_n \geq x)}{1-\Phi(x)} \to 1$$

uniformly in $x \in [0, o(n^{1/6}))$.

  - Jing, Shao and Wang (2003): If $E|X|^3 < \infty$, then

$$\frac{P(S_n/V_n \geq x)}{1-\Phi(x)} = 1 + O(1)\frac{(1+x)^3 E|X|^3}{\sqrt{n}\sigma^3}$$

for $0 \leq x \leq n^{1/6}\sigma/(E|X|^3)^{1/3}$, where $|O(1)| \leq C$.

Let $\xi_1, ..., \xi_n$ be independent random variables with $E\xi_i = 0$ and $E\xi_i^2 < \infty$ satisfying

$$\sum_{i=1}^n E\xi_i^2 = 1.$$

Let

$$W_n = \sum_{i=1}^n \xi_i, \quad V_n^2 = \sum_{i=1}^n \xi_i^2$$

and D_1, D_2 be measurable functions of $\{\xi_i, 1 \le i \le n\}$.

Let $\xi_1, ..., \xi_n$ be independent random variables with $E\xi_i = 0$ and $E\xi_i^2 < \infty$ satisfying

$$\sum_{i=1}^{n} E\xi_i^2 = 1.$$

Let

$$W_n = \sum_{i=1}^{n} \xi_i, \quad V_n^2 = \sum_{i=1}^{n} \xi_i^2$$

and D_1, D_2 be measurable functions of $\{\xi_i, 1 \le i \le n\}$.

Assume

$$T_n = \frac{W_n + D_1}{V_n(1 + D_2)^{1/2}}.$$

Theorem (Shao and Wenxin Zhou (2011))

There is an absolute constant $A > 1$ such that

$$e^{O(1)\Delta_{n,x}}\big(1 - AR_{n,x}\big) \leq \frac{P(T_n \geq x)}{1 - \Phi(x)}$$

and

$$\begin{aligned} P(T_n \geq x) \quad &\leq \quad \big(1 - \Phi(x)\big)e^{O(1)\Delta n,x}\big(1 + AR_{n,x}\big) \\ &\quad +P\big(|D_1|/V_n > 1/(2x)\big) + P\big(|D_2| > 1/(2x^2)\big) \end{aligned}$$

for all $x > 1$ satisfying

$$\Delta_{n,x} \leq (1 + x)^2/A,\ x^2 \max_{1 \leq i \leq n} E\xi_i^2 \leq 1,$$

where

$$
\begin{aligned}
\Delta_{n,x} &= x^2 \sum_{i=1}^{n} E\xi_i^2 I(x|\xi_i| > 1) + x^3 \sum_{i=1}^{n} E|\xi_i|^3 I(x|\xi_i| \le 1), \\
R_{n,x} &= I_{n,0}^{-1} \Big\{ xE(|D_1| + x|D_2|) e^{\sum_{j=1}^{n}(x\xi_j - x^2\xi_j^2/2)} \\
&\quad + x \sum_{i=1}^{n} E(|\xi_i(D_1 - D_1^{(i)})| + x|\xi_i(D_2 - D_2^{(i)})|) e^{\sum_{j\neq i}^{n}(x\xi_j - x^2\xi_j^2/2)} \Big\}, \\
I_{n,0} &= \prod_{i=1}^{n} ee^{x\xi_i - x^2\xi_i^2/2},
\end{aligned}
$$

and $D_1^{(i)}$ and $D_2^{(i)}$ are any random variables that don't depend on ξ_i.

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d random variables, and let $h(x, y)$ be a symmetric kernel, i.e., $h(x, y) = h(y, x)$. $\theta = Eh(X_1, X_2)$.

4. Studentized U-Statistics

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d random variables, and let $h(x, y)$ be a symmetric kernel, i.e., $h(x, y) = h(y, x)$. $\theta = Eh(X_1, X_2)$.

U-statistic (Hoeffding (1948)):

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_i, X_j)$$

The standardized U-statistic:

$$\frac{\sqrt{n}}{2\,\sigma_1}(U_n - \theta).$$

where $\sigma_1^2 := \mathrm{Var}(g(X)) > 0$ and $g(x) = E(h(x, X))$.

Studentized U-statistic:

$$T_n = \frac{\sqrt{n}}{2\, s_1}(U_n - \theta)$$

where

$$s_1^2 = \frac{(n-1)}{(n-2)^2} \sum_{i=1}^{n} \left(\frac{1}{n-1} \sum_{j \neq i} h(X_i, X_j) - U_n \right)^2$$

Studentized U-statistic:

$$T_n = \frac{\sqrt{n}}{2\,s_1}(U_n - \theta)$$

where

$$s_1^2 = \frac{(n-1)}{(n-2)^2}\sum_{i=1}^{n}\left(\frac{1}{n-1}\sum_{j\neq i} h(X_i, X_j) - U_n\right)^2$$

Hoeffding's decomposition: (assume $\theta = 0$)

$$T_n = \frac{W_n + D_1}{V_n(1 + D_2)^{1/2}}$$

where

$$W_n = \sum_{i=1}^{n} \xi_i,\ \xi_i = g(X_i)/(\sigma_1\sqrt{n}),\ V_n^2 = \sum_{i=1}^{n} \xi_i^2$$

D_1 and D_2 are small.

- Berry-Esseen bounds: Callaert and Veraverbeke (1981), Zhao (1983), Wang, Jing and Zhao (2000), ...

- Cramér type moderate deviations: Vandemaele and Veraverbeke (1985), Wang (1998), Lai, Shao and Wang (2009)

- Lai, Shao and Wang (2009):
  Assume that $\sigma_1 > 0$ and $E|h(X_1, X_2)|^3 < \infty$. If

$$h^2(x_1, x_2) \le c_0(\sigma_1^2 + g^2(x_1) + g^2(x_2))$$

for some $c_0 > 0$, then

$$\frac{P(T_n \ge x)}{1 - \Phi(x)} \to 1$$

holds uniformly in $x \in [0, o(n^{1/6}))$.

- Lai, Shao and Wang (2009):
  Assume that $\sigma_1 > 0$ and $E|h(X_1, X_2)|^3 < \infty$. If

$$h^2(x_1, x_2) \leq c_0(\sigma_1^2 + g^2(x_1) + g^2(x_2))$$

  for some $c_0 > 0$, then

$$\frac{P(T_n \geq x)}{1 - \Phi(x)} \to 1$$

  holds uniformly in $x \in [0, o(n^{1/6}))$.

- Shao and Wenxin Zhou (2011):

$$\frac{P(T_n \geq x)}{1 - \Phi(x)} = 1 + O(1)\frac{(1+x)^3}{\sqrt{n}}$$

  for $x \in [0, o(n^{1/6}))$.

5. Hotelling's T^2 statistics

Let $d \geq 2$ and $\boldsymbol{X}$ be a $d \times 1$ random vector with mean vector $\boldsymbol{\mu}$ and non-degenerate covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample of $n(n > d)$ independent observations of $\boldsymbol{X}$.

Hotelling's T^2 statistic:

$$T_n^2 = (\boldsymbol{S}_n - n\boldsymbol{\mu})' \bar{\boldsymbol{V}}_n^{-1} (\boldsymbol{S}_n - n\boldsymbol{\mu}),$$

where

$$\boldsymbol{S}_n = \sum_{i=1}^{n} \boldsymbol{X}_i,\ \bar{\boldsymbol{X}} = \boldsymbol{S}_n/n,\ \bar{\boldsymbol{V}}_n = \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'.$$

Let $d \geq 2$ and $\boldsymbol{X}$ be a $d \times 1$ random vector with mean vector $\boldsymbol{\mu}$ and non-degenerate covariance matrix $\boldsymbol{\Sigma}$. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be a random sample of $n(n > d)$ independent observations of $\boldsymbol{X}$.

Hotelling's T^2 statistic:

$$T_n^2 = (\boldsymbol{S}_n - n\boldsymbol{\mu})' \bar{\boldsymbol{V}}_n^{-1} (\boldsymbol{S}_n - n\boldsymbol{\mu}),$$

where

$$\boldsymbol{S}_n = \sum_{i=1}^{n} \boldsymbol{X}_i, \ \bar{\boldsymbol{X}} = \boldsymbol{S}_n/n, \ \bar{\boldsymbol{V}}_n = \sum_{i=1}^{n} (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})'.$$

- T_n^2 has a limiting χ^2-distribution with d degrees of freedom.
- Fujikoshi (1997) and Kano (1995): Egeworth expansion with an error of $o(1/n)$ if $E|\boldsymbol{X}|^8 < \infty$

- Dembo and Shao (2006): Assume $\boldsymbol{\mu} = 0$
    - For $x > 0$

$$\lim_{n\to\infty} P\Big(T_n^2 \geq x\,n\Big)^{1/n} = K\Big(\sqrt{x/(1+x)}\Big),$$

where

$$K(\alpha) = \sup_{b\geq 0} \sup_{||\boldsymbol{\theta}||=1} \inf_{t\geq 0} E\exp\Big(t\big(b\boldsymbol{\theta}'\boldsymbol{X} - \alpha((\boldsymbol{\theta}'X)^2 + b^2)/2\big)\Big).$$

- Dembo and Shao (2006): Assume $\boldsymbol{\mu} = 0$
  - For $x > 0$

$$\lim_{n\to\infty} P\Big(T_n^2 \geq x\,n\Big)^{1/n} = K\Big(\sqrt{x/(1+x)}\Big),$$

where

$$K(\alpha) = \sup_{b\geq 0}\ \sup_{||\boldsymbol{\theta}||=1}\ \inf_{t\geq 0} E\exp\Big(t\big(b\boldsymbol{\theta}'\boldsymbol{X} - \alpha((\boldsymbol{\theta}'X)^2 + b^2)/2\big)\Big).$$

  - For any $x_n \to \infty$ and $x_n = o(n)$,

$$\ln P\Big(T_n^2 \geq x_n\Big) \sim -\frac{1}{2}x_n$$

Theorem (Weidong Liu and Shao (2012))

Suppose that $E\|\boldsymbol{X}\|^{3+\delta} < \infty$ *for some* $\delta > 0$*. Then*

$$\frac{P\left(T_n^2 \geq x\right)}{P\left(\chi^2(d) \geq x\right)} \to 1$$

uniformly for $x \in [0, o(n^{1/3}))$*.*

Theorem (Weidong Liu and Shao (2012))

Suppose that $E\|\boldsymbol{X}\|^{3+\delta} < \infty$ *for some* $\delta > 0$. *Then*

$$\frac{P\left(T_n^2 \geq x\right)}{P\left(\chi^2(d) \geq x\right)} \to 1$$

uniformly for $x \in [0, o(n^{1/3}))$.

- Similar result holds for two-sample Hotelling T^2 statistic
- The results have been successfully applied to global tests of means and control FDR in multiple tests for means

Conjecture: If $E|\boldsymbol{X}|^3 < \infty$, then

$$\lim_{n\to\infty} P(T_n^2 \geq x)/P(\chi_d^2 \geq x) = 1$$

holds uniformly in $0 \leq x \leq o(n^{1/3})$; Moreover

$$P(T_n^2 \geq x)/P(\chi_d^2 \geq x) = 1 + O(1)\frac{(1+x)^{3/2}E|\boldsymbol{X}|^3}{n^{1/2}|\boldsymbol{\Sigma}|^{3/2}}.$$

- Concluding remark:
  - Craḿer type moderate deviations for the self-normalized process require very little moment conditions and hence the results are more appealing to applications.
  - More studies in this direction and their applications in probability, statistics and other fields are worthy to be further explored.

Thank you!

V. de la Pena, T.L. Lai and Q.M. Shao (2009). Self-normalized Processes: Theory and Statistical Applications. Springer.