# Approximating the CLT using Stein's method

Ben Berckmoes

May 21, 2015
National University of Singapore

(joint work with B. Lowen, G. Molenberghs and J. Van Casteren)

# Estimation of $\mu$ with $X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

Independent observations of $N(\mu, 1)$

$$X_1, X_2, \ldots, X_k, \ldots$$

are contaminated according to the **inflated variance model**, i.e.

$$X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$$

with $p_k \in [0, 1]$ and $\sigma_k \in [1, \infty[$. Under which conditions is the sample mean

$$\widetilde{\mu}_n = \frac{1}{n} \sum_{k=1}^{n} X_k$$

(weakly) consistent for $\mu$ and asymptotically normal?

# Estimation of $\mu$ with $X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

**Proposition**

$$\mathbb{E}[\widetilde{\mu}_n] = \mu$$

*and*

$$\mathrm{Var}[\widetilde{\mu}_n] = \left(\frac{s_n}{n}\right)^2$$

*where*

$$s_n^2 = \sum_{k=1}^{n} [(1 - p_k) + p_k \sigma_k^2].$$

# Estimation of $\mu$ with $X_k \sim (1-p_k)N(\mu,1)+p_kN(\mu,\sigma_k^2)$

**Theorem**

*Suppose that*

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{k=1}^{n}p_k\sigma_k^2=0.$$

*Then*

$$\widetilde{\mu}_n\stackrel{\mathbb{P}}{\to}\mu.$$

# Estimation of $\mu$ with $X_k \sim (1 - p_k) N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

**Theorem**

*Suppose that*

$$\lim_{n \to \infty} \frac{1}{n} \max_{k=1}^{n} \sigma_k^2 = 0.$$

*Then*

$$\frac{n}{s_n} \left( \widetilde{\mu}_n - \mu \right) \xrightarrow{w} N(0, 1).$$

# Estimation of $\mu$ with $X_k \sim (1-p_k)N(\mu,1) + p_k N(\mu,\sigma_k^2)$

**Theorem**

*Suppose that*

$$\sigma_n \uparrow \infty \text{ and } \liminf_{n\to\infty} \frac{\sigma_n}{n} > 0$$

*and*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_k \sigma_k^2 = L.$$

*Then*

$$\widetilde{\mu}_n \stackrel{\mathbb{P}}{\to} \mu$$

*and*

$$\frac{n}{s_n}(\widetilde{\mu}_n - \mu) \stackrel{w}{\to} N(0,1) \Leftrightarrow L = 0.$$

What happens if $L \neq 0$?

# Central Limit Theorem

A **Feller standard triangular array (FSTA)** of rv's

$$\begin{array}{ccc} \xi_{1,1} & & \\ \xi_{2,1} & \xi_{2,2} & \\ \xi_{3,1} & \xi_{3,2} & \xi_{3,3} \\ & \vdots & \end{array}$$

has the following properties:

$(a)$ $\forall n : \xi_{n,1}, \ldots, \xi_{n,n}$ are independent,

$(b)$ $\forall n, k : \mathbb{E}\left[\xi_{n,k}\right] = 0$,

$(c)$ $\forall n : \sum_{k=1}^{n} \sigma_{n,k}^2 = 1$ with $\sigma_{n,k}^2 = \mathbb{E}\left[\xi_{n,k}^2\right]$,

$(d)$ $\max_{k=1}^{n} \sigma_{n,k}^2 \to 0$. (Feller negligible)

# Central Limit Theorem

**Theorem (CLT)**

*For an FSTA* $\{\xi_{n,k}\}$ *TFAE:*

$(a)$ $\sum_{k=1}^{n} \xi_{n,k} \overset{w}{\to} N(0,1)$.

$(b)$ $\forall \epsilon > 0 : \sum_{k=1}^{n} \mathbb{E}\left[\xi_{n,k}^2; \left|\xi_{n,k}\right| \geq \epsilon\right] \to 0$. *(Lindeberg's condition)*

# Approximate Central Limit Theorem

Let $\xi \sim N(0,1)$ and $K$ be the **Kolmogorov distance**. That is,

$$K(\eta,\zeta) = \sup_{x\in\mathbb{R}} \left|\mathbb{P}[\eta \leq x] - \mathbb{P}[\zeta \leq x]\right|.$$

Then

$$\limsup_{n\to\infty} K\left(\xi, \sum_{k=1}^{n} \xi_{n,k}\right) = 0 \Leftrightarrow \sum_{k=1}^{n} \xi_{n,k} \overset{w}{\to} \xi,$$

$$\limsup_{n\to\infty} K\left(\xi, \sum_{k=1}^{n} \xi_{n,k}\right) = \sup_{h\in\mathcal{H}} \limsup_{n\to\infty} \left|\mathbb{E}\left[h(\xi) - h\left(\sum_{k=1}^{n} \xi_{n,k}\right)\right]\right|,$$

$$\mathcal{H} = \left\{\mathbb{R} \overset{h}{\to} [0,1] \mid h \text{ strictly } \downarrow,\ C^{\infty},\ \lim_{x\to-\infty} h(x) = 1,\ \lim_{x\to\infty} h(x) = 0\right\}.$$

# Approximate Central Limit Theorem

The **classical method** (e.g. Fourier analysis, Gaussian transforms) performs an **analysis of** $h$ which leads to

$$\left|\mathbb{E}\left[h(\xi) - h\left(\sum_{k=1}^{n} \xi_{n,k}\right)\right]\right|$$
$$\leq \frac{1}{6}\left\|h'''\right\|_\infty \left(\mathbb{E}\left[|\xi|^3\right] \max_{k=1}^{n} \sigma_{n,k} + \epsilon\right) + \left\|h''\right\|_\infty \sum_{k=1}^{n} \mathbb{E}\left[\xi_{n,k}^2 ; \left|\xi_{n,k}\right| \geq \epsilon\right]$$

which, after taking the lim sup, recalling Feller's negligibility condition and letting $\epsilon \downarrow 0$, reduces to

$$\limsup_{n\to\infty} \left|\mathbb{E}\left[h(\xi) - h\left(\sum_{k=1}^{n} \xi_{n,k}\right)\right]\right|$$
$$\leq \left\|h''\right\|_\infty \left(\sup_{\epsilon>0} \limsup_{n\to\infty} \sum_{k=1}^{n} \mathbb{E}\left[\xi_{n,k}^2 ; \left|\xi_{n,k}\right| \geq \epsilon\right]\right).$$

# Approximate Central Limit Theorem

We call

$$\mathrm{Lin}\left(\left\{\xi_{n,k}\right\}\right) = \sup_{\epsilon>0} \limsup_{n\to\infty} \sum_{k=1}^{n} \mathbb{E}\left[\xi_{n,k}^2 ; \left|\xi_{n,k}\right| \geq \epsilon\right]$$

the **Lindeberg index**. It has the following properties:

(*a*) $\mathrm{Lin}\left(\left\{\xi_{n,k}\right\}\right) = 0 \Leftrightarrow \left\{\xi_{n,k}\right\}$ satisfies Lindeberg's condition,

(*b*) $0 \leq \mathrm{Lin}\left(\left\{\xi_{n,k}\right\}\right) \leq 1$.

# Approximate Central Limit Theorem

The classical method has thus produced the inequality

$$\limsup_{n\to\infty}\left|\mathbb{E}\left[h(\xi)-h\left(\sum_{k=1}^{n}\xi_{n,k}\right)\right]\right|\leq\left\|h''\right\|_{\infty}\mathrm{Lin}\left(\left\{\xi_{n,k}\right\}\right)\qquad(1)$$

which holds for every test function $h$. This proves that Lindeberg's condition is sufficient for normal convergence.
However, since $\|h''\|_{\infty}$ **blows up** if we let $h$ run through $\mathcal{H}$, (1) is useless to derive an upper bound for the number $\limsup_{n\to\infty}K\left(\xi,\sum_{k=1}^{n}\xi_{n,k}\right)$.

# Approximate Central Limit Theorem

The **Stein-Chen method** starts with the observation

$$\left|\mathbb{E}\left[h(\xi) - h\left(\sum_{k=1}^{n} \xi_{n,k}\right)\right]\right|$$

$$= \left|\mathbb{E}\left[\left(\sum_{k=1}^{n} \xi_{n,k}\right) f_h\left(\sum_{k=1}^{n} \xi_{n,k}\right) - f_h'\left(\sum_{k=1}^{n} \xi_{n,k}\right)\right]\right|$$

where

$$f_h(x) = e^{x^2/2} \int_{-\infty}^{x} \left(h(t) - \mathbb{E}[h(\xi)]\right) e^{-t^2/2} dt,$$

# Approximate Central Limit Theorem

and then performs **an analysis of** $f_h$ which leads to

$$\left|\mathbb{E}\left[h(\xi) - h\left(\sum_{k=1}^{n} \xi_{n,k}\right)\right]\right|$$

$$\leq \frac{1}{2}\left\|f_h''\right\|_\infty \epsilon + \left(\sup_{x_1,x_2\in\mathbb{R}}\left|f_h'(x_1) - f_h'(x_2)\right|\right)\sum_{k=1}^{n}\mathbb{E}\left[\left|\xi_{n,k}\right|^2 ; \left|\xi_{n,k}\right| \geq \epsilon\right]$$

$$+ \left(\sup_{x_1,x_2\in\mathbb{R}}\left|f_h''(x_1) - f_h''(x_2)\right|\right)\max_{k=1}^{n}\sigma_{n,k}$$

# Approximate Central Limit Theorem

which, after taking the lim sup, recalling Feller's negligibility condition and letting $\epsilon \downarrow 0$, reduces to

$$\limsup_{n\to\infty} \left| \mathbb{E}\left[ h(\xi) - h\left( \sum_{k=1}^{n} \xi_{n,k} \right) \right] \right| \leq \left( \sup_{x_1, x_2 \in \mathbb{R}} \left| f_h'(x_1) - f_h'(x_2) \right| \right) \mathrm{Lin}\left( \left\{ \xi_{n,k} \right\} \right).$$

Now $\sup_{x_1, x_2 \in \mathbb{R}} \left| f_h'(x_1) - f_h'(x_2) \right|$ **does not blow up** if we let $h$ run through $\mathcal{H}$ as it is always **bounded by 1**.

# Approximate Central Limit Theorem

Therefore we get

**Theorem (Approximate CLT)**

*For an FSTA* $\{\xi_{n,k}\}$

$$\limsup_{n\to\infty} K\left(N(0,1), \sum_{k=1}^{n} \xi_{n,k}\right) \leq \mathrm{Lin}\left(\{\xi_{n,k}\}\right).$$

# Estimation of $\mu$ with $X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

**Theorem**

*Suppose that*

$$\sigma_n \uparrow \infty \text{ and } \liminf_{n\to\infty} \frac{\sigma_n}{n} > 0$$

*and*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_k \sigma_k^2 = L.$$

*Then*

$$\widetilde{\mu}_n \xrightarrow{\mathbb{P}} \mu$$

*and*

$$\frac{n}{s_n}(\widetilde{\mu}_n - \mu) \xrightarrow{w} N(0,1) \Leftrightarrow L = 0.$$

What happens if $L \neq 0$?

# Estimation of $\mu$ with $X_k \sim (1-p_k)N(\mu,1) + p_k N(\mu,\sigma_k^2)$

**Proposition**

$$\left\{\frac{1}{s_n}(X_k-\mu)\right\} \textit{ is an } \text{FSTA}$$

*and*

$$\sum_{k=1}^{n}\frac{1}{s_n}(X_k-\mu) = \frac{n}{s_n}(\widetilde{\mu}_n-\mu)$$

*and*

$$\mathrm{Lin}\left(\left\{\frac{1}{s_n}(X_k-\mu)\right\}\right) = \frac{L}{1+L}.$$

# Estimation of $\mu$ with $X_k \sim (1-p_k)N(\mu,1) + p_k N(\mu,\sigma_k^2)$

**Theorem**

*Suppose that*

$$\sigma_n \uparrow \infty \text{ and } \liminf_{n\to\infty} \frac{\sigma_n}{n} > 0$$

*and*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} p_k \sigma_k^2 = L.$$

*Then*

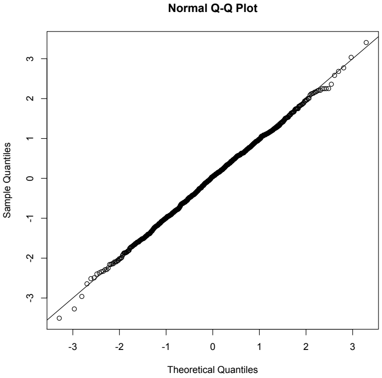
$$\widetilde{\mu}_n \overset{\mathbb{P}}{\to} \mu$$

*and*

$$\limsup_{n\to\infty} K\left(N(0,1), \frac{n}{s_n}(\widetilde{\mu}_n - \mu)\right) \leq \frac{L}{1+L}.$$

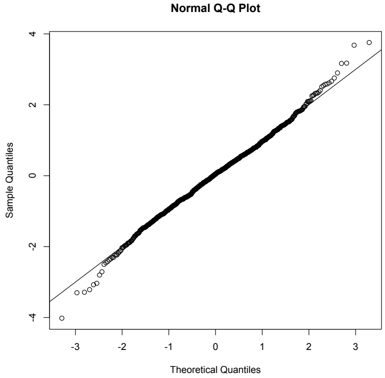
# Estimation of $\mu$ with $X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

Figure: Lin = 0.02



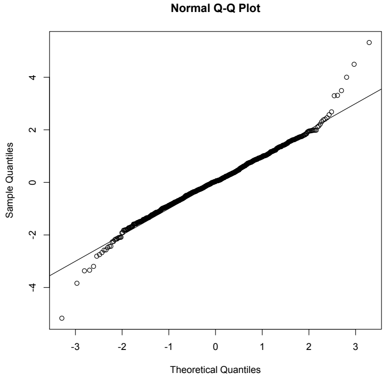
# Estimation of $\mu$ with $X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

Figure: Lin = 0.18



# Estimation of $\mu$ with $X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

Figure: Lin = 0.44



# Estimation of $\mu$ with $X_k \sim (1 - p_k)N(\mu, 1) + p_k N(\mu, \sigma_k^2)$

Figure: Lin = 0.82