Empirical and Quantile Process CLTs For Time Dependent Data

Kuelbs, Kurtz and Zinn

Part I: Empirical Process CLTs, KKZ

ABSTRACT. For stochastic processes $\{X_t : t \in E\}$, we establish sufficient conditions for the empirical process based on $\{I_{X_t \le y} - P(X_t \le y)\colon t \in E, y \in \mathbb{R}\}$ to satisfy the CLT uniformly in $t \in E, y \in \mathbb{R}$. Typically $E = [0, T]$, or a finite product of such intervals. Corollaries of our main result include examples of classical processes where the CLT holds, and we also show that it fails for Brownian motion tied down at zero and $E = [0, 1]$.

Part II: Empirical QuantileProcess CLTs, KZ

ABSTRACT. We establish empirical quantile process CLTs based on n independent copies of a stochastic process $\{X_t : t \in E\}$ that are uniform in $t \in E$ and quantile levels $\alpha \in I$, where I is a closed sub-interval of $(0, 1)$. Also included are additional empirical process CLTs. The process $\{X_t : t \in E\}$ may be chosen from a broad collection of Gaussian processes, compound Poisson processes, stationary independent increment stable processes, and martingales.

Motivation: Weak convergence of the scaled median of independent Brownian motions, Probab. Theory Relat. Fields (2007) by Jason Swanson.

Notation

E is a set, $D(E)$ is a collection of real valued functions on E.

$X = \{X_t : t \in E\}$ is a stochastic process with $P(X(\cdot) \in D(E)) = 1$.

P is the law of X on a sigma algebra of $D(E)$ containing $\mathcal{C}$.

$\mathcal{C} = \{C_{s,x} : s \in E, x \in \mathbb{R}\}$, where $C_{s,x} = \{z \in D(E) : z(s) \leq x\}$.

$\{X_j\}_{j=1}^{\infty}$ are i.i.d. copies of $\{X(t) : t \in E\}$ on a suitable probability space $(\Omega, \Sigma, \mathbb{P})$.

$F(t, y) = P(X_t \leq y) = P(X \in C_{t,y}), t \in E, y \in \mathbb{R}$.

The empirical distribution functions built on $\mathcal{C}$ (or built on the process X) are

$$F_n(t,y) = \frac{1}{n}\sum_{j=1}^{n} I_{X_j(t)\leq y} = \frac{1}{n}\sum_{j=1}^{n} I_{X_j \in C_{t,y}}, (t,y) \in E \times \mathbb{R}.$$

The empirical processes indexed by $\mathcal{C}$ (or just $E \times \mathbb{R}$) and built on X are

$$\{\nu_n(t,y) = \sqrt{n}(F_n(t,y) - F(t,y)) : (t,y) \in E \times \mathbb{R}\}, n \geq 1,$$

and X is the input process.

Part I: The Empirical Process CLT

In the language of empirical process theory, we are asking that the central limit theorem holds in $\ell_\infty(\mathcal{C})$ for the processes

$$\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[I_{X_j \in C} - P(X \in C) \right] : C \in \mathcal{C} \right\}, \tag{1}$$

or, equivalently, identifying $C_{t,y} \in \mathcal{C}$ with the point $(t, y) \in E \times \mathbb{R}$, that the CLT holds in $\ell_\infty(E \times \mathbb{R})$ for the empirical processes

$$\{\sqrt{n}(F_n(t, y) - F(t, y)) : (t, y) \in E \times \mathbb{R}\}. \tag{2}$$

Remark 1.

(i) To minimize notation we simply write or say that $\mathcal{C} \in CLT(P)$, or the CLT holds over $\mathcal{C}$ with respect to P. We also emphasize that we are looking at an empirical CLT for a class of sets, which has additional structure due to the "time" parameter of X.

(ii) Recall that the limiting Gaussian measure γ_P in the CLT of (1), or (2), is required to be Radon on $\ell_\infty(\mathcal{C})$, or $\ell_\infty(E \times \mathbb{R})$. In addition, the centered Gaussian process $\{G(t, y) : (t, y) \in E \times \mathbb{R}\}$ having covariance function

$$(3) \quad E(G(s,x)G(t,y)) = P(X_s \leq x, X_t \leq y) - P(X_s \leq x)P(X_t \leq y),$$

and L^2-distance

$$(4) \qquad d_G((s,x),(t,y)) = E((G(s,x) - G(t,y)))^2)^{\frac{1}{2}},$$

admits a version all of whose trajectories are uniformly bounded and uniformly d_G continuous on $E \times \mathbb{R}$, and induces law γ_P on $\ell_\infty(E \times \mathbb{R})$. Of course, it also is required that for every bounded, continuous $F : \ell_\infty(E \times \mathbb{R}) \to \mathbb{R}$,

$$\lim_{n \to \infty} \mathbb{E}^* F(\nu_n) = \mathbb{E} F(G).$$

First Guesses on the Influence of the Time Parameter

(i) Since the class of half spaces $\{(-\infty, x] : x \in \mathbb{R}\}$ is universal for the classical empirical CLT, a first guess was that perhaps $\{X(t) : t \in E\}$ satisfying the CLT would suffice for our empirical CLT.

(ii) Maybe a Lipshitz condition on $\{X(t) : t \in E\}$ would imply $\mathcal{C} \in CLT(P)$.

A distributional transform:

Given the distribution function F of any real valued random variable, Y, and V a uniform random variable independent of Y, we define the **distributional transform of** F to be

$$\tilde{F}(x, V) = F(x^-) + V(F(x) - F(x^-)).$$

Then,

$$\tilde{F}(Y, V) \text{ is uniform on } [0, 1] \ ,$$

$\tilde{F}(x, V)$ is non-decreasing in x, and $\tilde{F} = F$ if F is continuous.

Theorem 1.

Let ρ be given by $\rho^2(s,t) = E(H(s) - H(t))^2$, for some centered Gaussian process H that is sample bounded and uniformly continuous on (E, ρ) with probability one. Let $\{X(t) : t \in E\}$ be such that for some $L < \infty$, and all $\epsilon > 0$,

$$(5) \qquad \sup_{t \in E} P^*\Big(\sup_{\{s : \rho(s,t) \le \epsilon\}} |\tilde{F}_t(X_s) - \tilde{F}_t(X_t)| > \epsilon^2 \Big) \le L\epsilon^2,$$

where, to simplify notation, we let

$$\tilde{F}_t(x) \equiv (\tilde{F}_t)(x, V), x \in \mathbb{R}$$

be the distributional transform of F_t, the distribution function of X_t. Then, $\mathcal{C} \in CLT(P)$, i.e. the empirical CLT holds in $\ell_\infty(\mathcal{C})$.

Remark 2.

(i) A single uniform V, independent of the process $\{X(t) : t \in E\}$, is used for all the distributional transforms in (5).

(ii) A situation where the distributional transforms are useful occurs when one has a point $t_0 \in E$ such that $P(X(t) = X(t_0) \text{ for all t in E}) = 1$, and F_{t_0} is possibly discontinuous. In this situation, (5) holds for the Gaussian process $H(t) = g$ for all $t \in E$, g a standard Gaussian random variable, and $X(t_0)$ having any distribution function F_{t_0}. Thus Theorem 1 implies the classical empirical CLT when the class of sets consists of half-lines for all laws F_{t_0} on $\mathbb{R}$.

(iii) The L-condition of (5) may fail when there is only one $t \in E$ with discontinuous distribution function F_t, and also even when $\mathcal{C}$ is P-pregaussian. Furthermore, there are examples where $\mathcal{C}$ is P-pregaussian and X satisfies a modified L-condition, but the CLT of Theorem 1 over $\mathcal{C}$ fails. The modified L-condition used is given as in (5) with the distributions $\tilde{F}_t$ replaced by the F_t. Hence, one needs to assume something more, and our results show the L-condition for the process $\{\tilde{F}_s(X_s) : s \in E\}$ is sufficient.

(iv) Our proof depends on Talagrand's generic chaining version of necessary and sufficient conditions for the Gaussian process $\{H(t) : t \in E\}$ to be uniformly continuous and sample bounded on (E, ρ). Combined with the L-condition, we can then show our process is P-pregaussian and also verify the conditions to apply the local conditions sufficient for $\mathcal{C} \in CLT(P)$ given in Theorem 4.4 of AGOZ.

The Brownian Motion Example

Let $X_1, \cdots, X_n$ be i.i.d. copies of sample continuous Brownian motion on $E = [0, T]$, and assume they start at zero when $t = 0$ with probability one. In this case we take $D(E) = C([0, T])$, P is Wiener measure on the Borel subsets of $D(E)$, and we show

$$(6) \qquad P(\Delta^{\mathcal{C}_Q}(\{X_1, \cdots, X_n\}) \\ \equiv \mathrm{card}\{C \cap \{X_1, \cdots, X_n\} : \ C \in \mathcal{C}_Q\} = 2^n) = 1, n \geq 1,$$

where Q is the rational numbers and $\mathcal{C}_Q = \{C_{t,y} : t \in [0, T] \cap Q, y \in \mathbb{R}\}$. Thus $\mathcal{C}_Q \notin CLT(P)$, as a necessary condition for that requires

$$\frac{\ln \Delta^{\mathcal{C}_Q}(\{X_1, \ldots, X_n\})}{\sqrt{n}} \to 0 \text{ in (outer) probability.}$$

Hence **the CLT fails when the input process is standard Brownian motion tied down at zero.** Nevertheless, if $Y = \{Y_t : t \in [0, T]\}$ is standard tied down Brownian motion on $[0, T]$, and for $t \in [0, T]$

$$X_t = Y_t + Z,$$

where Z is independent of the process Y and has bounded density function, then the **CLT holds**.

The proof of (6) depends on the LIL for Brownian motion at $t = 0$, and **a similar result also holds for fractional Brownian motions and the Brownian sheet, i.e the CLT fails for tied down inputs, but it holds when we add a Z as above**. For Brownian motion, as well as other processes with stationary independent increments, one can use the Hewitt-Savage zero-one law instead of the LIL to verify (6). However, for fractional Brownian motions that approach is not applicable. The proof that the empirical CLT holds when we add Z in these situations, as well as for some other examples, was done in KKZ, and depends on verifying the L-condition of Theorem 1.

Part II: The Empirical Quantile Process CLT

The **quantiles and empirical quantiles** are defined as the left-continuous inverses of $F(t,x)$ and $F_n(t,x)$ in the variable x, respectively:

$$\tau_\alpha(t) = \inf\{x\colon F(t,x) \geq \alpha\}$$

and

$$\tau_\alpha^n(t) = \inf\{x\colon F_n(t,x) \geq \alpha\}.$$

The **empirical quantile processes** are defined as

$$\sqrt{n}\big(\tau_\alpha^n(t) - \tau_\alpha(t)\big),$$

for these processes.

Remark 4.

Since we are seeking limit theorems with non-degenerate Gaussian limits, it is appropriate to mention that for $\alpha \in (0, 1)$ and t fixed, that is, for a one-dimensional situation, a necessary condition for the weak convergence of

$$\sqrt{n}\big(\tau_\alpha^n(t) - \tau_\alpha(t)\big) \Longrightarrow \xi, \tag{7}$$

where ξ has a strictly increasing, continuous distribution, is that the distribution function $F(t, \cdot)$ be differentiable at $\tau_\alpha(t)$ and $F'(t, \tau_\alpha(t)) \in (0, \infty)$. Hence $F(t, \cdot)$ is strictly increasing near $\tau_\alpha(t)$ as a function of x, and if we keep t fixed, but ask that (7) holds for all $\alpha \in (0, 1)$, then $F(t, x)$ will be differentiable, with strictly positive derivative $F'(t, x)$ on the the set $J_t = \{x : 0 < F(t, x) < 1\}$.

Moreover, if $F'(t,x)$ is locally in L_1 with respect to Lebesgue measure on J_t, then a standard real analysis fact implies $F'(t,x)$ is the density of $F(t,\cdot)$ and it is strictly positive on J_t. For many of the base processes we study, $J_t = \mathbb{R}$ for all $t \in E$, but should that not be the case, it can always be arranged by adding an independent random variable Z with strictly positive density to our base process in order to have a suitable input process.

Theorem 2.

Assume for all $t \in E$ that the distribution functions $F(t,x)$ are strictly increasing, their densities $f(t,\cdot)$ satisfy

$$\lim_{\delta\to 0} \sup_{t\in E} \sup_{|u-v|\le\delta} |f(t,u)-f(t,v)| = 0, \tag{8}$$

and for every closed interval I in $(0,1)$ there is an $\theta(I) > 0$ such that

$$\inf_{t\in E,\alpha\in I,|x-\tau_\alpha(t)|\le\theta(I)} f(t,x) \equiv c_{I,\theta(I)} > 0. \tag{9}$$

In addition, assume the CLT holds on $\mathcal{C}$ with respect to P and has centered Gaussian limit process $\{G(t,x)\colon\ (t,x) \in E\times\mathbb{R}\}$. Then, for I a closed subinterval of $(0,1)$, the quantile processes $\{\sqrt{n}(\tau_\alpha^n(t)-\tau_\alpha(t))\colon\ n\ge 1\}$ satisfy the CLT in $\ell_\infty(E\times I)$ with respect to P and have Gaussian limit process

$$\{\frac{G(t,\tau_\alpha(t))}{f(t,\tau_\alpha(t))}\colon\ (t,\alpha)\in E\times I\}. \tag{10}$$

Remark 5.

(i) The proof of Theorem 2 involves the almost sure version of the CLT for empirical processes as presented, for example, in D. This is applicable since we are assuming $\mathcal{C} \in CLT(P)$, and the perfect mappings obtained are important for our proof. Then, using an idea of Vervaat, we are able pass from the CLT for empirical processes, to a CLT for their **"inverses"**(actually difference of inverses), the empirical quantile processes.

(ii) Since the empirical CLT appears as an assumption for the empirical quantile CLT, we were motivated to broaden the class of processes where the empirical CLT holds. This has been done, and it now includes a more general collection of Gaussian processes than we have already mentioned, compound Poisson processes, stationary independent increment stable processes(and others), and a broad collection of martingales.

(iii) The results in the paper by Jason Swanson mentioned earlier deal only with the case $\alpha = 1/2$(medians), and allow the Brownian motion to start at zero at time zero, whereas the assumptions in (8) and (9) do not allow us to apply Theorem 2 in that situation. However, the CLT results obtained here are uniform for $(t, \alpha) \in E \times I$ for a whole range of processes. Of course, for stationary independent increment stable processes on $[0, T]$, to apply Theorem 2 we must also add an independent Z if the process started at zero at time zero, i.e. the Hewitt Savage zero-one law shows the empirical CLT over $\mathcal{C}$ will fail as for Brownian motion. Another difference is that since Swanson deals only with sample continuous Brownian motions, the median process is continuous, and his weak convergence is proven in $C([0, T])$, whereas our results are in an ℓ_∞-type space.

Can one prove an empirical quantile CLT that includes stationary independent increment stable processes on $[0, T]$ which start at zero at time zero? Can we obtain Swanson's result in $C([0, T])$ for Brownian motion and in a cadlag space of functions for the other stable processes?

The answer to both questions is **yes**, and the scaling property of the input process emerges in an important way. Hence we only show these results for the stable processes mentioned, fractional Brownian motions, and the Brownian sheet. The result for stable processes is as follows.

Theorem 3.

Let $\{X(t) \colon t \geq 0\}$ be a symmetric r-stable process with stationary, independent increments, cadlag sample paths, and such that $P(X(0) = 0) = 1$. Also, assume the empirical quantile processes $\tau_\alpha^n(t)$ are built from i.i.d. copies of $\{X(t) : t \geq 0\}$ with cadlag paths, and I is a closed subinterval of $(0, 1)$. Then, the quantile processes

$$\{\sqrt{n}(\tau_\alpha^n(t) - \tau_\alpha(t)) \colon \ n \geq 1\}$$

satisfy the CLT in $\ell_\infty([0, T] \times I)$ with centered Gaussian limit process

$$\{W(t, \alpha) \colon \ (t, \alpha) \in [0, T] \times I\},$$

where $W(0, \alpha) = 0, \alpha \in I$,

$$W(t, \alpha) = \frac{G(t, \tau_\alpha(t))}{f(t, \tau_\alpha(t))}, (t, \alpha) \in (0, T] \times I,$$

and for $(s, \beta), (t, \alpha) \in (0, T] \times I$ the covariance function is given by

$$E(W(s, \beta)W(t, \alpha)) = \frac{P(X(s) \leq \tau_\beta(s), X(t) \leq \tau_\alpha(t)) - \alpha\beta}{f(s, \tau_\beta(s))f(t, \tau_\alpha(t))}.$$

Remarks – (The CLT and sample path properties of the input process X)

(1) If $\alpha \in (0,1)$ is fixed in Theorem 3, then the empirical quantile CLT holds in the Banach subspace $D_\alpha([0,T])$ of $\ell_\infty([0,T] \times \{\alpha\})$, where

$$D_\alpha([0,T]) = \{f(\cdot,\alpha) : f(\cdot,\alpha) \text{ is cadlag on } [0,\mathrm{T}]\}.$$

Moreover, if $r = 2$ and the empirical quantile processes $\tau_\alpha^n(t)$ are built from i.i.d. copies of $\{X(t) : t \geq 0\}$ with sample continuous paths, then for $\alpha \in (0,1)$ fixed the empirical quantile CLT holds in the Banach subspace $C_\alpha([0,T])$ of $\ell_\infty([0,T] \times \{\alpha\})$, where

$$C_\alpha([0,T]) = \{f(\cdot,\alpha) : f(\cdot,\alpha) \text{ is continuous on } [0,\mathrm{T}]\}.$$

Furthermore, if $r = 2, \alpha = \frac{1}{2}$, the covariance of the limiting Gaussian process is

$$E(W(s,\frac{1}{2})W(t,\frac{1}{2})) = \frac{P(X(s) \leq 0), X(t) \leq 0) - \frac{1}{4}}{f(s,0)f(t,0)} = \sqrt{st}\sin^{-1}(\frac{\mathrm{s} \wedge \mathrm{t}}{\sqrt{\mathrm{st}}}).$$

(2) The fact that the empirical quantile CLT holds in these smaller Banach subspaces for $\alpha \in (0,1)$ fixed, can be extended to hold in $[0, T] \times I$. For example, if $I = [a,b] \subset (0,1)$, in the Brownian motion situation the subspace can taken to be all the functions $f(t,\alpha)$ that are continuous in $t \in [0, T]$ for each $\alpha \in I$, and, for each $t \in [0, T]$, left continuous in $\alpha \in (a,b]$, with right limits on $[a,b)$. For stable processes with $0 < r < 2$ and cadlag input processes, the Banach subspace consists of functions $f(t,\alpha)$ which are in cadlag in $t \in [0, T]$ for each $\alpha \in I$, and for each $t \in [0, T]$, left continuous in $\alpha \in (a,b]$, with right limits on $[a,b)$.