Strong Invariance Principle for the Generalized Quantile Process under Dependence

Martin Wendler

RUHR-UNIVERSITÄT BOCHUM

Dependence in Probability and Statistics
CIRM Workshop, Luminy, April 4-8, 2011

Outline

Kiefer-Müller Process

Let $(X_n)_{n\in\mathbb{N}}$ be iid uniformly on $[0,1]$ and $F_n(t) := \frac{1}{n}\sum_{i=}^{n} \mathbb{1}_{\{X_n \le t\}}$. Then $\left(\frac{1}{\sqrt{n}}(ns(F_{ns}(t)-t))\right)_{t,s\in[0,1]}$ converges weakly to a Gaussian process $(G(t,s))_{t,s\in[0,1]}$ with

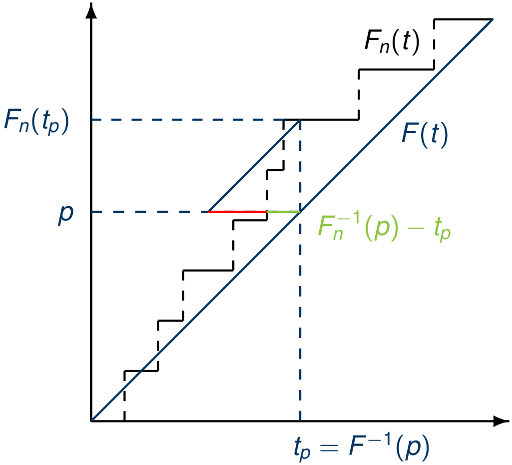
$$EG(t,s)G(t',s') = \min\{s,s'\}(\min\{t,t'\} - tt').$$

Theorem (Kiefer, 1972)

There exists (after enlarging the probability space) a Gaussian process $(G(t,s))_{t,s\in[0,1]}$ such that almost surely

$$\sup_{t,s\in[0,1]} \frac{1}{\sqrt{n}} |ns(F_{ns}(t)-t) - G(t,s)| = O(n^{-\frac{1}{6}} \log^{\frac{2}{3}} n).$$

Bahadur Representation



$$F_n^{-1}(p) := \inf\{t | F_n(t) \geq p\}$$

$$\frac{F_n(t_p) - p}{t_p - F_n^{-1}(p)} \approx f(t_p) := F'(t_p)$$

$$F_n^{-1}(p) - t_p = \frac{p - F_n(t_p)}{f(t_p)} + R_n$$

Bahadur Representation II

$$F_n^{-1}(p) - t_p = \frac{p - F_n(t_p)}{f(t_p)} + R_n$$

How to find bounds for R_n?

$$R_n = \frac{F_n(t_p) - p}{f(t_p)} + F_n^{-1}(p) - t_p$$

$$\approx \frac{F_n(t_p) - F_n(F_n^{-1}(p))}{f(t_p)} - (t_p - F_n^{-1}(p))$$

Theorem (Bahadur, 1966)

Let $(X_n)_{n\in\mathbb{N}}$ be iid. Then almost surely

$$R_n = O(n^{-\frac{3}{4}}(\log n)^{\frac{1}{2}}(\log\log n)^{\frac{1}{4}})$$

Hoeffding Decomposition

$g : \mathbb{R}^2 \to \mathbb{R}$ measurable and symmetric

Definition

The (bivariate) *U-statistic* $U_n(g)$ with kernel h is defined as

$$U_n(g) := \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} g(X_i, X_j).$$

Example: Gini's mean difference

$$G_n := \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|$$

Hoeffding Decomposition II

$U_n\,(g)$ can be decomposed into a *linear part* and a *degenerate part*

$$U_n(g) = \theta + \frac{2}{n}\sum_{i=1}^{n} g_1(X_i) + U_n(g_2)$$

with $\operatorname{Var} U_n(g_2) = O(\frac{1}{n^2})$. The CLT for partial sums together with Slutzky's lemma imply:

Theorem (Hoeffding, 1948)

If $(X_n)_{n\in\mathbb{N}}$ *is a sequence of iid random variables and* $\operatorname{Var} g(X, Y) < \infty$*, then*

$$\sqrt{n}\,(U_n\,(h) - \theta) \xrightarrow{\mathcal{D}} N\,(0, 4\operatorname{Var} g_1(X_1))\,.$$

U-Distribution Function

$h : \mathbb{R}^3 \to \mathbb{R}$ bounded, measurable function, symmetric in first two arguments, nondecreasing in third argument,

Definition (Empirical U-distribution function)

We define

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

$(U_n(t))_{t \in \mathbb{R}}$ is called empirical U-distribution function.

- example: $h(x, y, t) = \mathbb{1}_{\{g(x,y) \le t\}}$
  $U_n(t)$ empirical distribution function of the sample $(g(X_i, X_j))_{1 \le i < j \le n}$
- natural estimator for U-distribution function
  $U(t) = E[h(X, Y, t)] = P[g(X, Y) \le t]$, where X, Y independent

U-Quantiles

Generalization of quantiles:

Definition (U-Quantile)

Let be $p \in (0,1)$. $t_p = U_n^{-1}(p) := \inf\left\{t \middle| U_n(t) \geq p\right\}$ is called the p-th empirical U-quantile.

- natural estimator of the U-quantile $t_p := U^{-1}(p)$
- for $h(x,y,t) = \mathbb{1}_{\{g(x,y)\leq t\}}$: smallest p-quantile t_p of the sample $(g(X_i,X_j))_{1\leq i<j\leq n}$
- example: median of absolute differences

$$Q_n = \text{median}\left\{ \left|X_i - X_j\right| \middle| 1 \leq i < j \leq n \right\}$$

Generalized Linear Statistics

Definition (*GL*-Statistic)

Let be $p_1, \ldots, p_d \in (0,1)$, $b_1, \ldots, b_d \in \mathbb{R}$ and J a bounded function, continuous a.e. and vanishes outside of I.

$$T_n = T\left(U_n^{-1}\right) := \int_I J(p)\, U_n^{-1}(p)dp + \sum_{j=1}^{d} b_j U_n^{-1}(p_j)$$

$$= \sum_{i=1}^{\frac{n(n-1)}{2}} \int_{\frac{2(i-1)}{n(n-1)}}^{\frac{2i}{n(n-1)}} J(t)\, dt \cdot U_n^{-1}\left(\frac{2i}{n(n-1)}\right) + \sum_{j=1}^{d} b_j U_n^{-1}(p_j)$$

is called generalized linear statistic (*GL*-statistic).

Generalized Linear Statistics II

- Let $h(x,y,t) := \mathbb{1}_{\{|x-y| \leq t\}}$, $p_1 = 0.5$, $b = 1$. The related *GL*-statistic is the median of absolute differences
- Let $h(x,y,t) := \frac{1}{2}\left(\mathbb{1}_{\{x \leq t\}} + \mathbb{1}_{\{y \leq t\}}\right)$, $p_1 = 0.25$, $p_2 = 0.75$, $b_1 = -1$, $b_2 = 1$, and $J = 0$.

$$T_n = F_n^{-1}(0.75) - F_n^{-1}(0.25)$$

  is the inter quartile distance.
- Let $h(x,y,t) := \mathbb{1}_{\{\frac{1}{2}(x-y)^2 \leq t\}}$, $p_1 = 0.75$, $b_1 = 0.25$ and $J(x) = \mathbb{1}_{\{x \in [0,0.75]\}}$. The related *GL*-statistic is called winsorized variance.

Mixing Conditions

Definition (Strong mixing)

$$\alpha(k) := \sup_{n\in\mathbb{N}} \sup_{A\in\mathcal{F}_1^n, B\in\mathcal{F}_{n+k}^\infty} |P(A\cap B) - P(A)P(B)|,$$

where $\mathcal{F}_b^a$ is the σ-field generated by r.v.'s $X_a, \ldots, X_b$
$(X_n)_{n\in\mathbb{N}}$ is called strongly mixing, if $\alpha(k) \to 0$ as $k \to \infty$.

Definition (Absolute Regularity)

$$\beta(k) := \sup_{n\in\mathbb{N}} E \sup\{\left|P(A|\mathcal{F}_{-\infty}^n) - P(A)\right| : A \in \mathcal{F}_{n+k}^\infty\}$$

$(X_n)_{n\in\mathbb{N}}$ is called absolutely regular, if $\beta(k) \to 0$ as $k \to \infty$.

Near Epoch Dependence

Definition (Near epoch dependent sequence)

Assume that $X_n = f\left((Z_{n+k})_{k\in\mathbb{Z}}\right)$ for a stationary process $(Z_n)_{n\in\mathbb{Z}}$. $(X_n)_{n\in\mathbb{Z}}$ is called a Near epoch dependent, if

$$E\left|X_1 - E(X_1 \middle| Z_{-l},\ldots,Z_l)\right| \leq a_l \qquad l = 0,1,2\ldots$$

with $a_n \to 0$.

examples:

1. linear processes (with absolutely regular innovations)
2. data from dynamical systems $X_{n+1} = T(X_n)$

General Assumptions

Assume that one of the following two mixing conditions hold:

(M1) $(X_n)_{n\in\mathbb{N}}$ is strongly mixing with mixing coefficients $\alpha(n) = O(n^{-\alpha})$ for $\alpha \geq 8$ and $E|X_i|^r < \infty$ for a $r > \frac{1}{5}$.

(M2) $(X_n)_{n\in\mathbb{N}}$ is near epoch dependent on an absolutely regular process with mixing coefficients $\beta(n) = O(n^{-\beta})$ for $\beta \geq 8$ with approximation constants $a(n) = O(n^{-a})$ for $a = \max\{\beta + 3, 12\}$.

$U(t) := Eh(X, Y, t)$ differentiable on (C_1, C_2) with $0 < \inf_{t\in(C_1,C_2)} U'(t) \leq \sup_{t\in(C_1,C_2)} U'(t) < \infty$ and

$$\sup_{s,t\in(C_1,C_2):\ |t-s|\leq x} \left|U(t) - U(s) - U'(t)(t-s)\right| = O\left(x^{\frac{5}{4}}\right)$$

General Assumptions II

continuity condition:

Definition (Variation Condition)

h satisfies the variation condition with constant L, if for all $\epsilon > 0$

$$E\left[\sup_{\|(x,y)-(X,Y)\|\leq\epsilon} |h(x,y,t) - h(X,Y,t)|\right] \leq L\epsilon,$$

where X, Y are independent with same distribution as X_1.

$\mathbb{1}_{\{|x-y|\leq t\}}$ satisfies the variation condition if X_1 has bounded density.

Empirical U-Process

Theorem (W., 2011)

There exists a centered Gaussian process $(K(t,s))_{t,s\in\mathbb{R}}$ *with*

$$EK(t,s)K(t',s') = \min\{s,s'\}\Big(4\,\mathrm{Cov}\left[h_1(X_1,t),h_1(X_1,t')\right] + 4\sum_{k=1}^{\infty}\mathrm{Cov}\left[h_1(X_1,t),h_1(X_{k+1},t')\right] + 4\sum_{k=1}^{\infty}\mathrm{Cov}\left[h_1(X_{k+1},t),h_1(X_1,t')\right]\Big).$$

such that almost surely

$$\sup_{t\in\mathbb{R},\ s\in[0,1]} \frac{1}{\sqrt{n}}\left|\lfloor ns\rfloor(U_{\lfloor ns\rfloor}(t)-U(t)) - K(t,ns)\right| = o(1).$$

Empirical U-process II

Theorem (W. 2010)

A.s. as $n \to \infty$

$$\sup_{t,t':\ |t-t'|\leq C\sqrt{\frac{\log\log n}{n}}} \left| U_n(t) - U_n(t') - u(t)(t-t') \right| = o\left(n^{-\frac{1}{2}-\frac{1}{8}\gamma} \log n\right)$$

for some $\gamma = (0,1)$.

Proof is based on

- Hoeffding decomposition
- 4th moment inequalities

Generalized Bahadur Representation

$$R_n(p) := \frac{U_n(t_p) - p}{u(t_p)} + U_n^{-1}(p) - t_p$$

$$\approx \frac{U_n(t_p) - U_n(U_n^{-1}(p))}{u(t_p)} - (t_p - U_n^{-1}(p))$$

- $\sup_{t \in \mathbb{R}} |U_n(t) - U(t)| = O\left(\sqrt{\frac{\log\log n}{n}}\right)$ a.s.
- $\sup_{p \in I} R_n(p) \leq C \sup_{t,t'\colon |t-t'| \leq C\sqrt{\frac{\log\log n}{n}}} \left| U_n(t) - U_n(t') - u(t)(t-t') \right|$

Theorem (W., 2011)

A.s. as $n \to \infty$

$$\sup_{p \in I} R_n = o\left(n^{-\frac{1}{2} - \frac{1}{8}\gamma} \log n\right)$$

Empirical *U*-Quantile Process

Theorem (W., 2011)

Under the technical assumptions above there exists a centered Gaussian process $(K'(p,s))_{p\in I, s\in\mathbb{R}}$ *with covariance function*

$$EK'(p,s)K'(p',s') = \frac{1}{u(t_p)u(t'_p)}EK(t_p,s)K(t_{p'},s').$$

such that

$$\sup_{p\in I,\ s\in[0,1]} \frac{1}{\sqrt{n}}\left|\lfloor ns\rfloor(U^{-1}_{\lfloor ns\rfloor}(p)-t_p)-K'(p,ns)\right| = o(1).$$

Generalized Linear Statistics

$$T_n = T\left(U_n^{-1}\right) := \int_I J(p)\, U_n^{-1}(p)dp + \sum_{j=1}^{d} b_j U_n^{-1}(p_j)$$

$$\sigma^2 = \int_I \int_I EK'(p,1)K'(q,1)J(p)J(q)dpdq$$
$$+ 2\sum_{j=1}^{d} b_j \int_I EK'(p,1)K'(p_j,1)J(p)dp + \sum_{i,j=1}^{d} b_i b_j EK'(p_i,1)K'(p_j,1)$$

Theorem (W., 2011)

There exists a Brownian motion B such that

$$\sup_{s\in[0,1]} \frac{1}{\sqrt{n}} \left| \lfloor ns \rfloor (T_{\lfloor ns \rfloor} - T(U^{-1})) - \sigma B(ns) \right| = o(1).$$

Generalized Linear Statistics II

Consequently, $\left(\frac{\sqrt{ns}}{\sigma}(T_{\lfloor ns \rfloor} - T(U^{-1}))\right)_{s\in[0,1]}$ converges weakly to a Brownian Motion.
Furthermore, the sequence

$$\left(\frac{\lfloor ns \rfloor}{\sigma\sqrt{2n \log\log n}}(T_{\lfloor ns \rfloor} - T(U^{-1}))_{s\in[0,1]}\right)_{n\in\mathbb{N}}$$

is almost surely relatively compact in the space of bounded continuous functions $C[0,1]$ (equipped with the supremum norm) and the limit set is

$$\left\{f : [0,1] \to \mathbb{R} \middle| f(0) = 0, \int_0^1 f'^2(s)ds \leq 1\right\}.$$

References

W. Hoeffding (1948). A class of statistics with asymptotically normal distribution, *Ann. Math. Stat.* **19** 293-325.

R.R. Bahadur (1966). A note on quantiles in large samples, *Ann. Math. Stat.* **37** 577-580.

J. Kiefer (1972). Skorohod embedding of multivariate RV's, and the Sample DF, *Probab. Theory Related Fields* **24** 1-35.

H. Dehling, M. Wendler (2010). Central limit theorem and the bootstrap for *U*-statistics of strongly mixing data. *J. Multivariate Ana.* **101** 126-137.

H. Dehling, M. Wendler (2010). Law of the iterated logarithm for *U*-statistics of weakly dependent observations. in: Berkes et al. (Eds): *Dependence in Probability, Analysis and Number Theory*. Kendrick Press.

M. Wendler (2011). Bahadur Representation for *U*-Quantiles of Mixing Data. In press: *J. Multivariate Ana.*

M. Wendler (2011). *U*-Processes, *U*-Quantile processes and generalized linear statistics. *preprint.*

I would like to thank

- the organizers of this conference for the opportunity to present my work
- Prof. H. Dehling for his supervision
- German Academic Foundation and Sonderforschungsbereich (Research Center): Dynamische Strukturen (Dynamical Structures) for financial support

Thank you for listening!