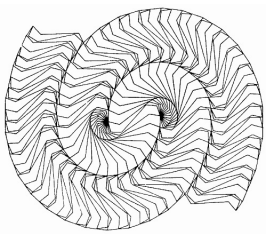
ERNST MORITZ ARNDT
UNIVERSITÄT GREIFSWALD

Wissen
lockt.
Seit 1456

Resampling for *U*-Statistics: a New and Fast Method

Martin Wendler
joint work with O.Sh. Sharipov and J. Tewes

German Probability and Statistics Days 2016



Institut für
Mathematik und Informatik

Outline

One-Sample-U-Statistics

- $(X_n)_{n\in\mathbb{N}}$ stationary sequence of random variables
- $h : \mathbb{R}^2 \to \mathbb{R}$ measurable and symmetric

Definition

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

called (bivariate) U**-statistic** $U_n(h)$ with kernel h

Examples

1. $h(x, y) = |x - y|$

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

Gini's mean difference

2. $h(x, y) = \frac{1}{2}(x - y)^2$

$$U_n(h) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

sample variance

Hoeffding-Decomposition

decompose $U_n(h)$ into **linear part** and **degenerate part**

$$U_n(h) = \theta + \frac{2}{n}\sum_{i=1}^{n} h_1(X_i) + U_n(h_2)$$

with

$$\begin{aligned} \theta &:= Eh(X_1, X_2) \\ h_1(x) &:= Eh(x, X_2) - \theta \\ h_2(x, y) &:= h(x, y) - h_1(x) - h_1(y) - \theta. \end{aligned}$$

degeneracy: for $i < j$ and X_i, X_j independent

$$E[h_2(X_i, X_j)|X_i] = 0.$$

CLT under Independence

summands of $U_n(h_2)$: uncorrelated

$$\operatorname{Var} U_n(h_2) = \frac{4}{n^2(n-1)^2} \sum_{1 \le i < j \le n} \operatorname{Var} h_2(X_i, X_j) = O(\frac{1}{n^2})$$

CLT for partial sums

$$\frac{2}{n} \sum_{i=1}^{n} h_1(X_i) \xrightarrow{\mathcal{D}} N(0, 4 \operatorname{Var} h_1(X_1))$$

Theorem (Hoeffding, 1948)

$(X_n)_{n \in \mathbb{N}}$ i.i.d. random variables and $\operatorname{Var} h_1(X_1) < \infty$, then

$$\sqrt{n}(U_n(h) - \theta) \xrightarrow{\mathcal{D}} N(0, 4 \operatorname{Var} h_1(X_1))$$

CLT under Dependence

Theorem

$$\sqrt{n}\left(U_n(h) - \theta\right) \xrightarrow{\mathcal{D}} N\left(0, 4\sigma_\infty^2\right)$$

with $\sigma_\infty^2 = \mathrm{Var}\left[h_1(X_1)\right] + 2\sum_{k=2}^{\infty} \mathrm{Cov}\left[h_1(X_1), h_1(X_k)\right]$

- Hoeffding (1948): independence, second moments
- Yoshihara (1976): absolute regularity, $(2+\delta)$-moments
- Denker, Keller (1986): near epoch dependence on absolutely regular sequences, $(2+\delta)$-moments, continuity condition
- Dehling, Wendler (2010): strong mixing, $(2+\delta)$-moments, continuity condition

if $\mathrm{Var}\left[h_1(X_1)\right]$: degenerate *U*-statistic, other limit

Strong Mixing

$(X_n)_{n\in\mathbb{N}}$ stationary process

Definition

strong mixing coefficient given by

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

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

- goes back to Rosenblatt (1956)
- standard assumption
- time series models like ARMA coverd under extra conditions

Continuity Condition

Definition

$h(x, y)$ satisfies the **variation condition** with constant L, if for all $\epsilon > 0$

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

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

holds for

- Lipschitz-conitnuous kernels, e.g. $h(x, y) = |x - y|$
- $h(x, y) = \frac{1}{2}(x - y)^2$ (variance estimation)
- discontinuous kernels under extra conditions, e.g. $h(x, y) = \mathbb{1}_{\{|x-y|\leq t\}}$

Outline

Efron's Bootstrap

- $X_1, \ldots, X_n$ i.i.d., real-valued
- distribution $F(t) = P(X_i \leq t)$ not known
- distribution needed for tests and confidences intervals
- $P((X_1, \ldots, X_n) \in A) = ?$

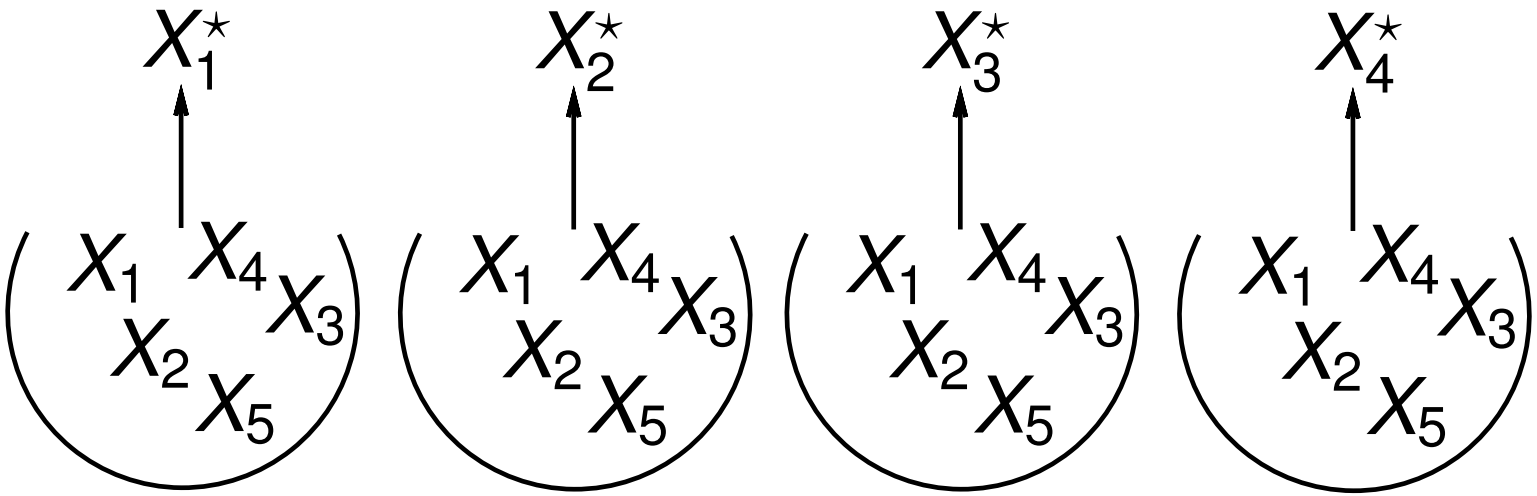
Efron's Bootstrap:

- estimate F by $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}$
- bootstrap sample: $X_1^\star, \ldots, X_n^\star$ i.i.d. conditional on $X_1, \ldots, X_n$ with distribution function F_n
- estimate $P((X_1, \ldots, X_n) \in A)$ by

$$P^\star((X_1^\star, \ldots, X_n^\star) \in A) = P((X_1^\star, \ldots, X_n^\star) \in A | X_1, \ldots, X_n)$$

Efron's Bootstrap II

- $X_1^\star, \ldots, X_n^\star$: draw with replacement from $X_1, \ldots, X_n$
- $P\left[X_i^\star = X_j\right] = \frac{1}{n}$ for $\quad i, j = 1, \ldots, n$
- use Monte Carlo simulation



Plug-in Bootstrap for U-Statistics

- bootstrap $X_1, \ldots, X_n$
- calculate U_n for $X_1^\star, \ldots, X_n^\star$

$$U_n^\star(h) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} h\left(X_i^\star, X_j^\star\right) = \theta + \frac{2}{n} \sum_{i=1}^{n} h_1(X_i^\star) + U_n^\star(h_2)$$

for the linear part (Singh, Bickel, Freedman, 1981):

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(h_1(X_i^\star) - E^\star[h_1(X_1^\star)]\right) \Rightarrow N\left(0, 4\sigma_1^2\right) \quad \text{a.s.}$$

remains to show: bootstrapped degenerate part converges to zero.

Plug-in Bootstrap for U-Statistics II

for $i_1 < i_2 < i_3 < i_4$:

$$E^\star\left[h_2\left(X_{i_1}^\star, X_{i_2}^\star\right) h_2\left(X_{i_3}^\star, X_{i_4}^\star\right)\right]$$
$$= \frac{1}{n^4} \sum_{i_1,i_2,i_3,i_4=1}^{n} h_2\left(X_{i_1}, X_{i_2}\right) h_2\left(X_{i_3}, X_{i_4}\right) = o\left(\frac{1}{n}\right) \quad \text{a.s.}$$

consequently

$$E^\star\left[\left(\sqrt{n} U_n^\star\left(h_2\right)\right)^2\right] \xrightarrow{\text{a.s.}} 0$$

Theorem (Bickel, Freedman, 1981)

$$\sqrt{n}\left(U_n^\star(h) - E^\star U_n^\star(h)\right) \Rightarrow N\left(0, 4\sigma_1^2\right) \quad a.s.$$

Circular Block Bootstrap

- Politis, Romano (1992)
- overlapping blocks of length $l = l(n)$
- drawing blocks with replacement $k = \left[\frac{n}{l}\right]$ times
- for $j = 1, \ldots, k$, $i = 1, \ldots, n$

$$P^\star(X^*_{l(j-1)+1} = X_i, ..., X^*_{jl} = X_{i+l-1}) :=$$

$$P\left(X^*_{l(j-1)+1} = X_i, ..., X^*_{jl} = X_{i+l-1} \middle| X_1, \ldots, X_n\right) = \frac{1}{n}$$

conditions on the block length:

- $l(n) \to \infty$, $l(n) \leq Cn^{1-\epsilon}$ for some $\epsilon > 0$
- $l(n) = p(2^k)$ for $2^k < n \leq 2^{k+1}$, $\quad l = 1, 2, \ldots$

Plug-in Bootstrap for U-Statistics

calculate U-statistic for $X_1^\star, \ldots, X_{lk}^\star$:

$$U_n^\star(h) = \frac{1}{\binom{lk}{2}} \sum_{1 \le i < j \le lk} h\left(X_i^\star, X_j^\star\right) = \theta + \frac{2}{lk} \sum_{i=1}^{lk} h_1(X_i^\star) + U_n^\star(h_2)$$

Theorem (Dehling, Wendler, 2010)

under the conditions of CLT for U-statistics

$$\sup_{x \in \mathbb{R}} \left| P^\star \left[\sqrt{lk} \left(U_n^\star(h) - E^\star \left[U_n^\star(h) \right] \right) \le x \right] - P\left[\sqrt{n} \left(U_n(h) - \theta \right) \le x \right] \right| \to 0 \quad a.s.$$

for degenerate U-statistic: different method, see Dehling, Sharipov, Wendler (2015)

Subsampling

- overlapping blocks of length $l = l(n)$
- calculate U-statistic for blocks: for $j = 1, \ldots, n$:

$$\hat{u}_{j,n} := \frac{2}{l(l-1)} \sum_{j \le i_1 < i_n \le j+l-1} h(X_{i_1}, X_{i_2})$$

subsampling estimator:

$$\hat{F}_{l,n}(t) = \frac{1}{n-l+1} \sum_{j=1}^{n-l+1} 1_{\{\sqrt{l}(\hat{u}_{j,n} - U_n(h)) \le t\}}$$

estimator for $F_{U_n}(t) = P\left(\sqrt{n}(U_n(h) - \theta) \le t\right)$

Subsampling Consistency

Theorem (Politis, Romano, 1994)

- *conditions for U-statistics CLT under strong mixing*
- *block length* $l \to \infty$, $l/n \to 0$

then

$$\sup_{t \in \mathbb{R}} \left(\hat{F}_{l,n}(t) - F_{U_n}(t) \right) \xrightarrow{\mathcal{P}} 0$$

sketch of proof:

- $\mathrm{Var}(\hat{F}_{l,n}(t)) \to 0$ because of strong mixing property
- $F_{U_n}(t) \to \Phi(\frac{t}{\sigma_\infty})$ because $\sqrt{n}(U_n - \theta) \Rightarrow N(0, \sigma_\infty)$
- $\mathrm{E}\, \hat{F}_{l,n}(t) \approx F_{U_l}(t) \to \Phi(\frac{t}{\sigma_\infty})$ because $l \to \infty$

Partial Bootstrap for U-statistics

- plug-in bootstrap: long computing times
- partial bootstrap: compromise between bootstrap and subsampling
- calculate U-statistic for blocks: for $j = 1, \ldots, n$:

$$\hat{u}_{j,n} := \frac{2}{l(l-1)} \sum_{j \le i_1 < i_2 \le j+l-1} h(X_{i_1}, X_{i_2})$$

bootstrap version:

- drawing from $\hat{u}_{1,n}, \ldots, \hat{u}_{n,n}$ with replacement $k = [n/l]$ times to get $\hat{u}^\star_{1,n}, \ldots, \hat{u}^\star_{k,n}$
-

$$\tilde{U}^\star_n(h) := \frac{1}{k} \sum_{j=1}^{k} \hat{u}^\star_{j,n}$$

Partial Bootstrap Consistency

$$\hat{u}^{\star}_{j,n} := \theta + \frac{2}{l} \sum_{i=(j-1)l+1}^{jl} h_1(X_i^{\star}) + \frac{2}{l(l-1)} \sum_{(j-1)l+1 \le i_1 < i_2 \le jl} h_2(X_{i_1}^{\star}, X_{i_2}^{\star})$$

for linear part:

$$\frac{1}{k} \sum_{j=1}^{k} \frac{2}{l} \sum_{i=(j-1)l+1}^{jl} h_1(X_i^{\star}) = \frac{2}{kl} \sum_{i=1}^{kl} h_1(X_i^{\star})$$

for degenerate part:

$$E^{\star}\left[\left(\frac{2}{l(l-1)} \sum_{(j-1)l+1 \le i_1 < i_2 \le jl} h_2(X_{i_1}^{\star}, X_{i_2}^{\star}) \right)^2 \right] = o_P(\frac{1}{l})$$

Partial Bootstrap Consistency II

Theorem (Sharipov, Tewes, Wendler, 2015)

under the conditions of CLT for U-statistics

$$\sup_{x\in\mathbb{R}}\left|P^\star\left[\sqrt{lk}\left(\tilde{U}_n^\star(h)-\overline{\hat{u}_{j,n}}\right)\le x\right]\right.$$
$$\left.-P\left[\sqrt{n}\left(U_n(h)-\theta\right)\le x\right]\right|\to 0$$

in probability

conjecture: under technical conditions, consistency holds for other statistics T_n if $\sqrt{n}T_n \to N(0,\sigma^2)$

Computing Time

M: number of Monte Carlo iterations

plug-in bootstrap (naively programmed): $\approx C_1 M n^2$
plug-in bootstrap (well programmed): $\approx C_2 n^2 l^2 + C_3 M \left(\frac{n}{l}\right)^2$
partial bootstrap: $\approx C_4 n l^2 + C_5 M \frac{n}{l}$
subsampling: $\approx C_6 n l^2$

Simulations I

U-statistic: sample variance (kernel $h(x,y) = \frac{1}{2}(x-y)^2$)

observations:

- AR-process: $X_n = \alpha X_{n-1} + \epsilon_n$
- $\alpha = 0.2, 0.4$
- $(\varepsilon_n)_{n \in \mathbb{N}}$ sequence of i.i.d. standard normal
- sample size $n = 100, 200$

resampling methods:

- plug-in bootstrap, partial bootstrap, subsampling
- block length $l = 3, 5, 7, 10, 15$

Simulations II

coverage probability of 95% confidence intervals
plug-in bootstrap / partial bootstrap / subsampling

n	α / l	0.2	0.4
100	3	0.908/0.922/0.823	0.894/0.817/0.730
	5	0.911/0.910/0.852	0.896/0.838/0.770
	7	0.909/0.909/0.849	0.874/0.857/0.793
	10	0.890/0.902/0.851	0.880/0.859/0.815
200	5	0.926/0.930/0.873	0.908/0.859/0.813
	7	0.925/0.926/0.881	0.905/0.870/0.831
	10	0.924/0.920/0.889	0.911/0.885/0.857
	15	0.910/0.910/0.884	0.902/0.893/0.849

Conclusion

Results:

- partial bootstrap as alternative for *U*-statistics
- partial bootstrap outperfornes subsampling

Open Questions:

- choice of block length?
- consistency of partial bootstrap for other statistics?

Thank you
for your attention!

References

H. DEHLING, O.SH. SHARIPOV, M. WENDLER (2015). Bootstrap for dependent Hilbert space-valued random variables with application to von Mises statistics. *JMVA* 133, 200-215.

H. DEHLING, M. WENDLER (2010). Central limit theorem and the bootstrap for U-statistics of strongly mixing data. *JMVA* 100, 126-137.

O.SH. SHARIPOV, M. WENDLER (2012) Bootstrap for the sample mean and for U-statistics of mixing and near-epoch dependent processes. *Journal of Nonparametric Statistics* 24, 317-342.

O.SH. SHARIPOV, J. TEWES, M. WENDLER (2015). Bootstrap for U-Statistics: A new approach. *preprint* arXiv: 1505.07260.