ERNST MORITZ ARNDT
UNIVERSITÄT GREIFSWALD

Wissen
lockt.
Seit 1456

Bootstrap and Change-Point Detection in Functional Time Series and Random Fields

Martin Wendler
coauthors: B. Bucchia, H. Dehling, O.Sh. Sharipov, J. Tewes

SMSA 2017

Institut für
Mathematik und Informatik

Hilbert Space

Hilbert space:

- H vector space
- $\langle\cdot,\cdot\rangle$ inner product (symmetric, bilinear, positive definite)
- $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$ norm
- H complete

assumptions:

- H separable
- $(X_n)_{n\in\mathbb{N}}$ H-valued
- $(X_n)_{n\in\mathbb{N}}$ stationary
- $E\,\|X_n\|^2 < \infty$

Absolute Regularity

Definition

absolute regularity defined as

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

with $\mathcal{F}^a_b = \sigma(Z_a, \ldots, Z_b)$, and process $(Z_n)_{n\in\mathbb{N}}$ called *absolutely regular*, if $\beta_k \to 0$ for $k \to \infty$

- dates back to Volkonskii, Rozanov (1959)
- not easy to establish for time series model
- not all time series covered, no dynamical systems

Near Epoch Dependence

Definition

$(X_n)_{n\in\mathbb{N}}$ called L_1 -*near epoch dependent* (NED) with approximation constants (a_l) on a process $(Z_n)_{n\in\mathbb{Z}}$ if

$$E\left\|X_n - E\left(X_n \middle| \mathcal{F}_{n-l}^{n+l}\right)\right\| \leq a_l$$

and $a_l \to 0$ für $l \to \infty$, with $\mathcal{F}_b^a = \sigma(Z_a, \ldots, Z_b)$

- introduced by Ibragimov (1962)
- holds for many time series models (GARCH, ARMA)
- holds for dynamical systems $X_{n+1} = T(X_n)$

Brownian Motion in Hilbert Space

Definition

A random variable X with values in H is called Gaussian, if for all $a \in H$, the random variable $\langle a, X \rangle$ has a normal distribution.

- $\langle a_1, X \rangle, \ldots, \langle a_d, X \rangle$ d-dimensional normal distribution
- covariance operator $V : H \to H$ defined by $\langle Va, b \rangle = \mathrm{Cov}\left(\langle a, X \rangle, \langle b, X \rangle\right)$

Definition

A random function $W : [0, \infty) \to H$ is called Brownian motion, if for all $t_1, \ldots, t_d$ the increments $W(t_1) - W(0)$, $W(t_2) - W(t_1)$,...,$W(t_d) - W(t_{d-1})$ are independent, centered, Gaussian with covariance operator $(t_i - t_{i-s})V$.

Central Limit Theorem

Theorem (Sharipov, Tewes, Wendler 2016)

1. $E\,\|X_1\|^{4+\delta} < \infty$ *for a* $\delta > 0$,
2. $(X_n)_{n\in\mathbb{Z}}$ L_1*-NED with* $\sum_{m=1}^{\infty} m^2 a_m^{\delta/(3+\delta)} < \infty$,
3. $(Z_n)_{n\in\mathbb{Z}}$ *absolutely regular with* $\sum_{m=1}^{\infty} m^2 \beta_m^{\delta/(4+\delta)} < \infty$

weak convergence in $D_H[0,1]$*:*

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{[nt]}(X_i - EX_i)\right)_{t\in[0,1]} \xrightarrow{W} W,$$

W Brownian motion with covariance operator V

$$\langle V(x), y\rangle = \sum_{j=-\infty}^{\infty} \mathrm{Cov}\left(\langle X_0, x\rangle, \langle X_j, y\rangle\right).$$

Efron's Bootstrap

- $X_1, \ldots, X_n$ i.i.d. real-valued
- unknown distribution function $F(t) = P(X_i \leq t)$
- how to construct tests and confidence intervals?
- needed: $P((X_1, \ldots, X_n) \in A)$

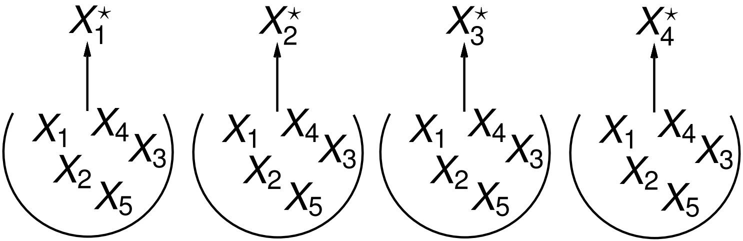
Efron's bootstrap:

- estimation of 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

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

Alternative Description

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



inconsistent under dependence!

Nonoverlapping Block Bootstrap

- divide sample into blocks of length $p = p(n)$
- $k = \left[\frac{n}{p}\right]$-times drawing with replacement from the blocks. new sample $X_1^\star, \dots, X_{kp}^\star$ with

$$P\left[\left(X_{(i-1)p+1}^\star, \dots, X_{ip}^\star\right) = \left(X_{(j-1)p+1}, \dots, X_{jp}\right)\right] = \frac{1}{k}$$

conditions on block length

- $p(n) \to \infty$
- $p(n) \leq Cn^{1-\epsilon}$ for a $0 < \epsilon < 1$
- $p(n) = p(2^l)$ for $2^l < n \leq 2^{l+1}, \qquad l = 1, 2, \dots$

Bootstrap Central Limit Theorem

Theorem (Sharipov, Tewes, Wendler 2014)

1. $E\,\|X_1\|^{4+\delta} < \infty$ *for a* $\delta > 0$,
2. $(X_n)_{n\in\mathbb{Z}}$ L_1*-NED with* $\sum_{m=1}^{\infty} m^2 a_m^{\delta/(3+\delta)} < \infty$,
3. $(Z_n)_{n\in\mathbb{Z}}$ *absolutely regular with* $\sum_{m=1}^{\infty} m^2 \beta_m^{\delta/(4+\delta)} < \infty$

almost surely weak convergence of conditional distribution:

$$\left(\frac{1}{\sqrt{kp}}\sum_{i=1}^{[kpt]}\left(X_i^\star - \frac{1}{kp}\sum_{i=1}^{kp} X_i\right)\right)_{t\in[0,1]} \xrightarrow{W^\star} W,$$

W Brownian motion with covariance operator V

$$\langle V(x), y\rangle = \sum_{j=-\infty}^{\infty} \mathrm{Cov}\left(\langle X_0, x\rangle, \langle X_j, y\rangle\right).$$

Ideas of Proof

- tightness and finite dimensional convergence
- tightness: use 4.-moments inequalities
- finite dimensional convergence: consider blocks
- enough to show (Varadarajan 1958): almost surely for countable many Lipschitz-continuous and bounded functions $(f_m)_{m\in\mathbb{N}}$

$$E\left[f_m\left(\frac{1}{\sqrt{p}}\left(\sum_{i=1}^{p} X_i^\star - \frac{1}{k}\sum_{i=1}^{kp} X_i\right)\right)\middle| X_1,\ldots,X_n\right]$$
$$= \frac{1}{k}\sum_{j=1}^{k} f_m\left(\frac{1}{\sqrt{p}}\left(\sum_{i=(j-1)p+1}^{jp} X_i - \frac{1}{k}\sum_{i=1}^{kp} X_i\right)\right) \xrightarrow{n\to\infty} E\left[f_m\left(N\right)\right]$$

N Gaussian with covariance operator V

Test Problem

- $(Y_n)_{n\in\mathbb{N}}$ stationary, centered process (not observed)
- $X_n = Y_n + \mu_n$ (observed, μ_n not known)

$$\begin{aligned} &\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_n \\ \text{against } &\mathrm{H}_1 : \exists k \in 1, \ldots, n-1 : \ \mu_1 = \ldots = \mu_k \neq \mu_{k+1}, \ldots, \mu_n. \end{aligned}$$

- if k known: two sample problem
- if k not known: change point problem

Change Point Statistic

k not known, CUSUM statistic

- comparing of partial sums (**cu**mulated **sum**)
- maximal difference for all time points k

$$T_n := \frac{1}{\sqrt{n}} \max_{k=1,\dots,n} \left\| \sum_{i=1}^{k} X_i - \frac{k}{n} \sum_{i=1}^{n} X_i \right\|$$

$$= \sup_{t\in[0,1]} \frac{1}{\sqrt{n}} \left\| \sum_{i=1}^{[nt]} X_i - t \sum_{i=1}^{n} X_i \right\| \xrightarrow{D} \sup_{t\in[0,1]} \left\| W(t) - tW(1) \right\| =: \sup_{t\in[0,1]} \|B(t)\|$$

- continuous mapping theorem
- $$\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} (X_i - \mu) \right)_{t\in[0,1]} \xrightarrow{W} (W(t))_{t\in[0,1]}$$

Bootstrap for CUSUM Statistic

$$\left(\frac{1}{\sqrt{kp}}\sum_{i=1}^{[kpt]}\left(X_i^\star - \frac{1}{kp}\sum_{i=1}^{kp} X_i\right)\right)_{t\in[0,1]} \xrightarrow{W^\star} W,$$

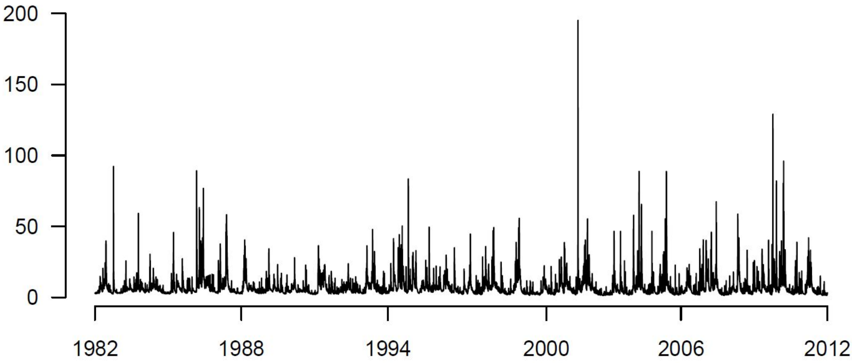
- W Brownian motion in H
- same covariance operator V as partial sum process

bootstrapped CUSUM statistic: with continuous mapping theorem

$$T_n^\star := \max_{m=1,\ldots,pk} \frac{1}{\sqrt{pk}}\left\|\sum_{i=1}^{m} X_i^\star - \frac{m}{n}\sum_{i=1}^{n} X_i^\star\right\| \xrightarrow{D^\star} \sup_{t\in[0,1]} \|W(t) - tW(1)\|$$

almost surely

Data : River Discharge



daily discharge of river Chemnitz at Göritzhain

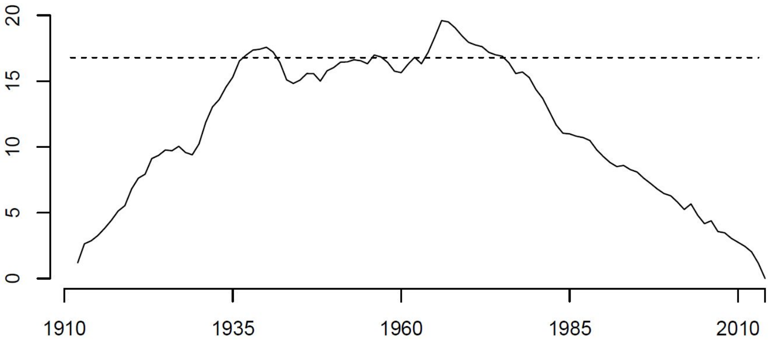
Data : River Discharge

- daily discharges of river Chemnitz at Göritzhain
- 1910 to 2012
- skewness, heavy tails
- dependence
- seasonality

use Hilbert-space theory:

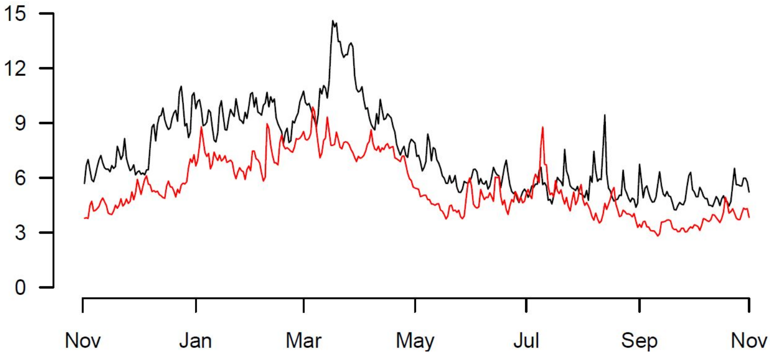
- avoid parametric models, smoothing
- treat as time series in $\mathbb{R}^{365}$
- bootstrap CUSUM test with block length 10
- logarithms

Norm of Differences



- maximum of differences exceeds bootstrapped level of significance, test decision: structural change
- time of change: time if highest difference (1964)

Before and after Change



- red curve: mean course of the years 1910 to 1964
- black curve: mean course of the years 1965 to 2012
- original data (no logarithm)

Simulation Results

comparison: projection on $d = 1$ or $d = 2$ principal component (proposed by Berkes et al. (2009))

- sample size $n = 50, 100$
- observations: independent Brownian motions (BM) oder Brownian bridges (BB)
- case 1: hypothesis (stationarity)
- case 2: mean jumps from $\mu(t) = 0$ to $\mu(t) = 2\sin(t)$
- 1000 simulation runs
- 500 bootstrap iterations
- level $\alpha = 5\%$

Simulation Results

empirical rejections frequencies, asymptotic level $\alpha = 0.05$

sample size	hypothesis	jump in mean
$n = 50$, BB	0.112 / 0.106 / 0.106	0.998 / 1.000 / 1.000
$n = 50$, BM	0.103 / 0.099 / 0.083	0.815 / 0.726 / 0.823
$n = 100$, BB	0.124 / 0.102 / 0.095	1.000 / 1.000 / 1.000
$n = 100$, BM	0.122 / 0.098 / 0.097	0.974 / 0.964 / 0.980

$d = 1$ principal components / $d = 2$ principal components / full Hilbert space

Cramér-von Mises-Statistic

weighted L^2-distance of empirical distribution function F_n with

$$F_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq t\}}$$

statistic: for bounded, integrable weight function w

$$V_n := \int (F_n(t) - F_0(t))^2 \, w(t)dt = \|F_n - F_0\|^2$$

Hilbert-space H: space of measurable functions $f : \mathbb{R} \to \mathbb{R}$ with $\langle f, f \rangle < \infty$, where

$$\langle f, g \rangle := \int f(t)g(t)w(t)dt$$

more general V-statistics possible

Change of Distribution Function

- hypothesis: stationary
- alternative: $\exists k : \; Y_1, \dots, Y_k \sim F, \; Y_{k+1}, \dots, Y_n \sim G$
- distribution functions $F \neq G$

test statistic

$$T_n := \max_{k=1,\dots,n} \frac{k^2}{n} \int (F_k(t) - F_n(t))^2 \, w(t) dt$$

critical values: use bootstrap approximation with

$$F_n^\star(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i^\star \le t\}}$$

Random Fields with Epedemic Changes

- $(Y_{\mathbf{n}})_{\mathbf{n}\in\mathbb{N}^d}$ stationary, centered, H-valued process
- $X_{\mathbf{n}} = Y_{\mathbf{n}} + \mu_{\mathbf{n}}$

$$\mathrm{H}_0 : \exists \mu \in H \forall \mathbf{i} \in \{1, \dots, n\}^d : \ \mu_{\mathbf{i}} = \mu$$

$$\text{against } \mathrm{H}_1 : \exists \mathbf{k}_1 < \mathbf{k}_2 \in \{1, \dots, n\}^d : \ \forall \mathbf{i} \notin [\mathbf{k}_1, \mathbf{k}_2] \ \mu_{\mathbf{i}} = \mu$$

$$\forall \mathbf{i} \in [\mathbf{k}_1, \mathbf{k}_2] \ \mu_{\mathbf{i}} = \tilde{\mu}$$

- $\mu \neq \tilde{\mu}$
- $\leq$: component wise

Test Statistic and Notation

test statistic:

$$T_n = \max_{0 \leq \mathbf{k} < \mathbf{m} \leq \mathbf{n}} \frac{1}{n^{d/2}} \left\| \sum_{\mathbf{k} < \mathbf{j} \leq \mathbf{m}} X_{\mathbf{j}} - \frac{\prod_{j=1}^{d}(m_j - k_j)}{n^d} \sum_{\mathbf{1} \leq \mathbf{j} \leq \mathbf{n}} X_{\mathbf{j}} \right\|.$$

notation:

- $\mathbf{n} = (n, \ldots, n)^t$
- $W(\mathbf{t})$, $\mathbf{t} \in [0,1]^d$: H-valued Brownian sheet (Chentsov process)
- increments:

$$W((\mathbf{s}, \mathbf{t}]) = \sum_{\varepsilon_1 = 0,1} \cdots \sum_{\varepsilon_d = 0,1} (-1)^{d - \sum_{i=1}^{d} \varepsilon_i} W(s_1 + \varepsilon_1(t_1 - s_1), \ldots, s_d + \varepsilon_d(t_d - s_d))$$

Limit Theorem

Theorem (Bucchia, Wendler, 2015)

If for a $\delta > 0$*:*

1. $E\|X_{\mathbf{1}}\|^{2+\delta} < \infty$
2. $\sum_{m\geq 1} m^{d-1}\alpha_{1,1}(m)^{\delta/(2+\delta)} < \infty$

$$\textit{then} \quad T_n \Rightarrow \sup_{\mathbf{0}\leq\mathbf{s}<\mathbf{t}\leq\mathbf{1}} \left\| W((\mathbf{s},\mathbf{t}]) - \left(\prod_{j=1}^{d} |s_j - t_j|\right) W(\mathbf{1}) \right\| = T$$

idea of proof:

- limit theorem for partial sum process $\frac{1}{n^{d/2}}\sum_{\mathbf{1}\leq\mathbf{j}\leq\lfloor n\mathbf{t}\rfloor}(X_{\mathbf{j}} - \mu)$, $\mathbf{t} \in [0,1]^d$
- continuous mapping theorem

Dependent Wild Bootstrap

partial sum process:

$$S_n(\mathbf{t}) = n^{-d/2} \sum_{\mathbf{1} \le \mathbf{i} \le \lfloor n\mathbf{t} \rfloor} (X_{\mathbf{i}} - \mu)$$

bootstrap version:

$$S_n^{\star}(\mathbf{t}) = n^{-d/2} \sum_{\mathbf{1} \le \mathbf{i} \le \lfloor n\mathbf{t} \rfloor} V_n(\mathbf{i}) \left(X_{\mathbf{i}} - \hat{\mu}(\mathbf{i})\right),$$

- $\hat{\mu}(\mathbf{i})$: estimator of mean
- $V_n(\mathbf{i})$, $\mathbf{i} \in \{1, \ldots, n\}^d$: triangular scheme of dependent, Gaussian random variables

Dependent Wild Bootstrap II

Theorem (Bucchia, Wendler, 2015)

under assumptions of CLT and addtional

- $\operatorname{Cov}\left(V_n(\mathbf{i}), V_n(\mathbf{j})\right) = \omega((\mathbf{i}-\mathbf{j})/q)$
- $\sum_{-\mathbf{n}\leq\mathbf{j}\leq\mathbf{n}} |\omega(\mathbf{j}/q)| = \mathcal{O}(q^d)$
- $q_n \to \infty$ *und* $q_n = o(\sqrt{n})$

we have

$$(S_n(\mathbf{t}), S_n^\star(\mathbf{t}))_{\mathbf{t}\in[0,1]^d} \Rightarrow (W(\mathbf{t}), W^\star(\mathbf{t}))_{\mathbf{t}\in[0,1]^d}$$

where $W^\star$ *is an independent copy of* W

corollary: bootstrap gives asymptotically correct critical values

Conclusion

- change point detection in function spaces without dimension reduction
- for time series and random fields
- functional central limit theorems in Hilbert spaces
- nonparametric bootstrap methods

References

- I. BERKES, R. GABRYS, L. HORVÁTH, P. KOKOSZKA (2009): Detecting changes in the mean of functional observations, *J. Royal Statistical Society: Series B* **71** 927-946
- DEHLING, FRIED, GARCIA, WENDLER (2014): Change-point detection under dependence based on two-sample U-statistics, *in: Asymptotic Methods in Stochastics - Festschrift in Honor of Miklos Csörgő*.1.
- 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.
- O.SH. SHARIPOV, J. TEWES, M. WENDLER (2016): Sequential block bootstrap in a Hilbert space with application to change point analysis, *Canadian Journal of Statistics* 44(3), 300-322.
- B. BUCCHIA, M. WENDLER (2017): Change-Point Detection and Bootstrap for Hilbert Space Valued Random Fields, *JMVA* 155, 344-368.