SFB 823

Asymptotic distribution of two-sample empirical U-quantiles with applications to robust tests for structural change

Herold Dehling, Roland Fried

Nr. 43/2010

Discussion Paper

Asymptotic Distribution of Two-Sample Empirical U-Quantiles with Applications to Robust Tests for Structural Change

Herold Dehling[1] and Roland Fried[2]

We derive the asymptotical distributions of two-sample U-statistics and two-sample empirical U-quantiles in the case of weakly dependent data. Our results apply to observations that can be represented as functionals of absolutely regular processes, including e.g. many classical time series models as well as data from chaotic dynamical systems. Based on these theoretical results we propose a new robust nonparametric test for the two-sample location problem, which is constructed from the median of pairwise differences between the two samples. We inspect the properties of the test in the case of weakly dependent data and compare the performance with classical tests such as the t-test and Wilcoxon's two-sample rank test with corrections for dependencies. Simulations indicate that the new test offers better power even than the Wilcoxon test in case of skewed and heavy tailed distributions, if at least one of the two samples is not very large. The test is then applied for detecting shifts of location in some weakly dependent time series, which are contaminated by outliers.

KEY WORDS: Functionals of absolutely regular processes; Hodges-Lehmann estimator; Two-sample location problem; U-statistics; Weak dependence.

1 Introduction

Consider the classical two-sample problem where $X_1, \ldots, X_{n_1}$ are F-distributed and $Y_1, \ldots, Y_{n_2}$ are G-distributed with unknown distribution functions F, G. In the case when all the observations are independent and when F and G are normal distributions with means μ_1, μ_2, respectively, and common variance σ^2, the minimum variance unbiased estimator for $\mu_1 - \mu_2$ is given by the difference of the sample means $\bar{X} - \bar{Y}$. Denoting the pooled variance of the samples by s_p^2, the uniformly most powerful unbiased test for the hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $\mu_1 > \mu_2$ under these assumptions rejects the hypothesis for large values of the t-test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(\frac{1}{n_1} + \frac{1}{n_2})s_p^2}}.$$

[1]Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany, herold.dehling@rub.de
[2]Fakultät Statistik, TU Dortmund, 44221 Dortmund, Germany, fried@statistik.tu-dortmund.de

A robust alternative to the estimator $\bar{X} - \bar{Y}$ was proposed by Hodges and Lehmann (1963). The two-sample Hodges-Lehmann estimator for the difference in location is defined as

$$\text{median}\{X_i - Y_j, 1 \le i \le n_1, 1 \le j \le n_2\}$$

The goal of this paper is to investigate the asymptotic distribution of estimators of the Hodges-Lehmann type and of related statistics in the case of dependent data. For this we will study the two-sample empirical U-process and empirical U-quantiles. Given a kernel $f(x, y)$, we define the empirical U-distribution function

$$U_{n_1,n_2}(t) = \frac{1}{n_1\, n_2} \#\{1 \le i \le n_1, 1 \le j \le n_2 : f(X_i, Y_j) \le t\}.$$

The empirical U-quantile function is defined as the generalized inverse of $U_{n_1,n_2}(t)$, i.e.

$$Q_{n_1,n_2}(p) = \inf\{t \in \mathbb{R} : U_{n_1,n_2}(t) \ge p\}.$$

The Hodges-Lehmann estimator is a special case of the empirical U-quantile, namely $Q_{n_1,n_2}(\frac{1}{2})$ for the kernel $f(x, y) = x - y$. We will investigate the asymptotic behavior of the empirical U-process and the empirical U-quantiles under rather weak assumptions on the dependence structure of the processes $(X_i)_{i\ge 1}$ and $(Y_i)_{i\ge 1}$. The population analogues of $U_{n_1,n_2}(t)$ and $Q_{n_1,n_2}(p)$ are

$$U(t) = P(f(X, Y) \le t) \quad \text{and} \quad Q(p) = \inf\{t \in \mathbb{R} : U(t) \ge p\}.$$

We will show asymptotic normality of the empirical U-process and the empirical U-quantile process, defined by

$$\begin{aligned} &\sqrt{n_1 + n_2}(U_{n_1,n_2}(t) - U(t)) \\ \text{and} \quad &\sqrt{n_1 + n_2}(Q_{n_1,n_2}(p) - Q(p)), \end{aligned}$$

respectively.

In the course of our work, we will further study the asymptotic distribution of two-sample U-statistics with kernel $h(x, y)$,

$$U_{n_1,n_2}(h) = \frac{1}{n_1\, n_2} \sum_{1\le i\le n_1} \sum_{1\le j\le n_2} h(X_i, Y_j),$$

again in the case of dependent data. For one-sample U-statistics of dependent data, there are results by Yoshihara (1976), Denker and Keller (1983, 1986), Borovkova, Burton and Dehling (2001) and Dehling and Wendler (2010). To the best of our knowledge, general two-sample U-statistics of dependent data have not been investigated before.

In this paper, we allow the sequences $(X_i)_{i\ge 1}$ and $(Y_j)_{j\ge 1}$ to be weakly dependent. Specifically, we will assume that $(X_i)_{i\ge 1}$ and $(Y_j)_{j\ge 1}$ are both stationary ergodic processes that can be represented as functionals of absolutely regular processes. Given a probability space $(\Omega, \mathcal{F}, P)$ and two sub-σ-fields $\mathcal{A}$ and $\mathcal{B}$ of $\mathcal{F}$, let

$$\beta(\mathcal{A}, \mathcal{B}) = \sup \sum_{i=1}^{m} \sum_{j=1}^{n} |P(A_i \cap B_j) - P(A_i)\, P(B_j)|,$$

where the supremum is taken over all partitions of Ω into sets $A_1, \ldots, A_m \in \mathcal{A}$, all partitions of Ω into sets $B_1, \ldots, B_n \in \mathcal{B}$ and all $m, n \geq 1$. A stochastic process $(X_i)_{i\in\mathbb{Z}}$ is called absolutely regular, if

$$\beta(k) = \sup_n \beta(\mathcal{F}_{-\infty}^n, \mathcal{F}_{n+k}^{\infty}) \to 0,$$

as $k \to \infty$. Here $\mathcal{F}_k^l$ denotes the σ-field generated by the random variables $X_k, \ldots, X_l$. A process $(X_i)_{i\geq 1}$ is called a functional of an absolutely regular sequence if there exists an absolutely regular process $(Z_n)_{n\in\mathbb{Z}}$ and a function $f : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ such that $X_i = f((Z_{i+n})_{n\in\mathbb{Z}})$. The process $(X_i)_{i\in\mathbb{Z}}$ is called a 1-approximating functional with coefficients $(a_k)_{k\geq 1}$ if

$$E\left(|X_i - E(X_i|Z_{i-k}, \ldots, Z_{i+k})|\right) \leq a_k.$$

Analogously we call $(X_i)_{i\geq 1}$ a one-sided functional of the absolutely regular process $(Z_n)_{n\geq 1}$ if $X_i = f(Z_i, Z_{i+1}, \ldots)$.

The concept of functionals of absolutely regular processes is wide enough to cover all relevant examples from statistics except long memory processes, as well as a large class of examples from dynamical systems. E.g., let $(X_i)_{i\geq 1}$ be an infinite order moving average process, i.e. $X_i = \sum_{n=0}^{\infty} \psi_n Z_{i-n}$ where $(Z_n)_{n\in\mathbb{Z}}$ is an i.i.d. process with finite mean. If the coefficients $(\psi_n)_{n\geq 0}$ are absolutely summable, we get

$$\begin{aligned} E\left(|X_i - E(X_i|Z_{i-k}, \ldots, Z_{i+k})|\right) &= E\left(|\sum_{n=0}^{\infty} \psi_n Z_{i-n} - \sum_{n=0}^{k} \psi_n Z_{i-n}|\right) \\ &\leq E(|Z_1|) \sum_{n=k+1}^{\infty} |\psi_n|. \end{aligned}$$

Thus, in this case $(X_i)_{i\geq 1}$ is a 1-approximating functional of an i.i.d. process with coefficient $a_k = E(|Z_1|) \sum_{n=k+1}^{\infty} |\psi_n|$. A large number of further examples of processes that can be represented as functionals of absolutely regular processes can be found, e.g. in Borovkova et al. (2001).

In the two-sample case, there are essentially two different settings in which dependent observations can arise.

1. The two samples $(X_i)_{i\geq}$ and $(Y_j)_{j\geq 1}$ are independent of each other, but there is dependence within the samples. E.g. both processes might be functionals of two independent absolutely regular processes.

2. We have a single process $(X_i)_{i\geq 1}$ that is a functional of an absolutely regular process, and $Y_j = X_{n_1+j}$, $1 \leq j \leq n_2$.

In the course of this work we will develop the theory under the model (2). This model is especially relevant for applications to change-point problems. All our results continue to hold under the first model, albeit with slightly different proofs.

The rest of the paper is organized as follows. Section 2 states the main theoretical results. Section 3 outlines asymptotical inference based on the Hodges-Lehmann estimator. We derive a test for a difference in location, which can be seen as a variant of the modified Wilcoxon test for strongly mixing processes (Serfling 1968), which is a smaller class of processes than the functionals of absolutely regular processes considered here. Section 4 compares the performance of the new test to a modified version of the t-test for dependent data. Section 5 provides some applications. Section 6 states some conclusions. The proofs are deferred to an appendix.

2 Statement of main theoretical results

This section presents the main theoretical results of our paper. Details of the proofs will be given in the appendix. We present our results for the case when we have a single underlying process $(X_i)_{i\geq 1}$. The two samples are thus the initial segment $X_1, \ldots, X_{n_1}$ and the following segment $X_{n_1+1}, \ldots, X_{n_1+n_2}$ of this one sequence of observations. Identical results hold in the case of two processes $(X_i)_{i\geq 1}$ and $(Y_i)_{i\geq 1}$ that are independent of each other.

The first fundamental result of this paper concerns two-sample U-statistics, defined as

$$U_{n_1,n_2}(h) = \frac{1}{n_1\, n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} h(X_i, X_j)$$

for some kernel $h(x, y)$. In order to formulate the theorem, we introduce the Hoeffding decomposition of the kernel $h(x, y)$, given by

$$h(x, y) = \theta + h_1(x) + h_2(y) + g(x, y), \tag{1}$$

where the constant θ and the functions $h_1(x), h_2(y)$ and $g(x, y)$ are defined by

$$\begin{aligned} \theta &= Eh(X, Y) \\ h_1(x) &= Eh(x, Y) - \theta \\ h_2(y) &= Eh(X, y) - \theta \\ g(x, y) &= h(x, y) - h_1(x) - h_2(y) + \theta, \end{aligned}$$

with X, Y being independent random variables with the same distribution as X_1.

In order to prove limit theorems for U-statistics of functionals of absolutely regular processes, we have to require continuity properties of the kernel h. The kernel $h(x,y)$ is called p-continuous, if there exists a function $\phi : [0,\infty) \to [0,\infty)$ such that for all $\epsilon > 0$

$$E\left(|h(X,Y') - h(X,Y)|\, 1_{|Y-Y'|\le\epsilon}\right) \le \phi(\epsilon), \tag{2}$$

for all random variables X, Y, Y' such that X, Y, Y' all have the same marginal distribution as X_1 and such that (X,Y) have joint distribution P_{X_1,X_k} for some k or $P_{X_1} \times P_{X_1}$. The notion of p-continuity was introduced by Borovkova et al. (2001), who present many examples of p-continuous kernels. Note that p-continuous kernels are not necessarily continuous functions. E.g., the kernel $1_{x\le y}$ is p-continuous provided that the marginal distribution of the process has a bounded density; for details see Borovkova et al. (2001).

Theorem 2.1 *Let $(X_i)_{i\ge1}$ be a 1-approximating functional with constants $(a_k)_{k\ge1}$ of an absolutely regular process with mixing coefficients $(\beta_k)_{k\ge1}$ and assume that $\sum_{k=1}^\infty k^2\,(a_k+\beta_k) < \infty$. Suppose moreover that $\sup_{i,j} E|h(X_i,X_j)|^{2+\epsilon} < \infty$ for some $\epsilon > 0$, and assume that the kernel $h(x,y)$ is 1-continuous. Then, as $n_1, n_2 \to \infty$ in such a way that $\frac{n_1}{n_1+n_2} \to \lambda \in (0,1)$, we get*

$$\sqrt{n_1+n_2}(U_{n_1,n_2} - \theta) \to N(0,\sigma^2),$$

where

$$\begin{aligned}\sigma^2 &= \frac{1}{\lambda}\left(\mathrm{Var}(h_1(X)) + 2\sum_{i=1}^\infty \mathrm{Cov}(h_1(X_1), h_1(X_{i+1}))\right)\\ &\quad + \frac{1}{1-\lambda}\left(\mathrm{Var}(h_2(X)) + 2\sum_{i=1}^\infty \mathrm{Cov}(h_2(X_1), h_2(X_{i+1}))\right).\end{aligned}$$

Next we will study the asymptotic distribution of the empirical U-distribution function

$$U_{n_1,n_2}(t) = \frac{1}{n_1\,n_2}\#\{1\le i\le n_1, n_1+1\le j\le n_1+n_2 : f(X_i,X_j)\le t\},$$

where $f(x,y)$ is a given kernel. We define the auxiliary functions

$$H_{t,1}(x) = P(f(x,X_1)\le t) \quad \text{and} \quad H_{t,2}(y) = P(f(X_1,y)\le t).$$

Theorem 2.2 *Let $(X_i)_{i\ge1}$ be a 1-approximating functional with constants $(a_k)_{k\ge1}$ of an absolutely regular process with mixing coefficients $(\beta_k)_{k\ge1}$ and assume that $\sum_{k=1}^\infty k^2\,(a_k+\beta_k) < \infty$. Suppose moreover that $U(t)$ is Lipschitz-continuous. Let $n_1, n_2 \to \infty$ in such a way that $\frac{n_1}{n_1+n_2} \to \lambda \in (0,1)$. Then*

$$\sqrt{n_1+n_2}(U_{n_1,n_2}(t) - U(t))_{t\in\mathbb{R}} \to (W_t)_{t\in\mathbb{R}},$$

in the sense of weak convergence of all finite dimensional marginals. Here $(W_t)_{t\in\mathbb{R}}$ is a mean-zero Gaussian process with variance

$$\mathrm{Var}(W_t) = \frac{\sigma_1^2(t)}{\lambda} + \frac{\sigma_2^2(t)}{1-\lambda}$$

where

$$\sigma_1^2(t) = \mathrm{Var}(H_{t,1}(X_1)) + 2\sum_{k=2}^{\infty}\mathrm{Cov}(H_{t,1}(X_1), H_{t,1}(X_k)) \tag{3}$$

$$\sigma_2^2(t) = \mathrm{Var}(H_{t,2}(X_1)) + 2\sum_{k=2}^{\infty}\mathrm{Cov}(H_{t,2}(X_1), H_{t,2}(X_k)) \tag{4}$$

and with autocovariance structure

$$\mathrm{Cov}(W_s, W_t) = \frac{\rho_1(s,t)}{\lambda} + \frac{\rho_2(s,t)}{1-\lambda}$$

where

$$\begin{aligned} \rho_1(s,t) &= \mathrm{Cov}(H_{s,1}(X_1), H_{t,1}(X_1)) + \sum_{k=2}^{\infty}\mathrm{Cov}(H_{s,1}(X_1), H_{t,1}(X_k)) \\ &\quad + \sum_{k=2}^{\infty}\mathrm{Cov}(H_{s,1}(X_k), H_{t,1}(X_1) \end{aligned} \tag{5}$$

$$\begin{aligned} \rho_2(s,t) &= \mathrm{Cov}(H_{s,2}(X_1), H_{t,2}(X_1)) + \sum_{k=2}^{\infty}\mathrm{Cov}(H_{s,2}(X_1), H_{t,2}(X_k)) \\ &\quad + \sum_{k=2}^{\infty}\mathrm{Cov}(H_{s,2}(X_k), H_{t,2}(X_1)) \end{aligned} \tag{6}$$

Remark 2.3 (i) Convergence in the above theorem holds in the function space $D([0,1])$, too. The proof, however, is technically quite involved and uses empirical process techniques as developed by Borovkova et al. (2001) for the one sample empirical U-process.

(ii) Empirical U-processes in the one sample case have been introduced by Serfling (1984). Serfling also proved the first invariance principle for empirical U-processes in the case of i.i.d. observations. Arcones and Yu (1994) proved the empirical U-process central limit theorem for absolutely regular observations. Later, this was extended to functionals of absolutely regular observations by Borovkova et al. (2001).

(iii) In the classical case when the process $(X_i)_{i\geq 1}$ is i.i.d., the formulae for the variance and covariance simplify to

$$\begin{aligned} \sigma_1^2(t) &= \mathrm{Var}(H_{t,1}(X_1)), & \sigma_2^2(t) &= \mathrm{Var}(H_{t,2}(X_1)) \\ \rho_1(s,t) &= \mathrm{Cov}(H_{s,1}(X_1), H_{t,1}(X_1)), & \rho_2(s,t) &= \mathrm{Cov}(H_{s,2}(X_1), H_{t,2}(X_1)) \end{aligned}$$

(iv) In case of the Hodges-Lehmann estimator, i.e. for the kernel $f(x,y) = x - y$, we obtain

$$\begin{aligned} H_{t,1}(x) &= P(x - X_1 \leq t) = P(X_1 \geq x - t) = 1 - F(x-t) \\ H_{t,2}(y) &= P(X_1 - y \leq t) = P(X_1 \leq t + y) = F(y+t), \end{aligned}$$

where F is the distribution function of X_1. In this case, the formulae for the limiting variance and covariance become

$$\begin{aligned} \sigma_1^2(t) &= \mathrm{Var}(F(X_1 - t)) + 2\sum_{k=2}^{\infty} \mathrm{Cov}(F(X_1 - t), F(X_k - t)) \\ \sigma_2^2(t) &= \mathrm{Var}(F(Y_1 + t)) + 2\sum_{k=2}^{\infty} \mathrm{Cov}(F(Y_1 + t), F(Y_k + t)) \end{aligned}$$

$$\begin{aligned} \rho_1(s,t) &= \mathrm{Cov}(F(X_1 - s), F(X_1 - t)) + \sum_{k=2}^{\infty} \mathrm{Cov}(F(X_1 - s), F(X_k - t)) \\ &\quad + \sum_{k=2}^{\infty} \mathrm{Cov}(F(X_k - s), F(X_1 - t)) \\ \rho_2(s,t) &= \mathrm{Cov}(F(Y_1 + s), F(Y_1 + t)) + \sum_{k=2}^{\infty} \mathrm{Cov}(F(Y_1 + s), F(Y_k + t)) \\ &\quad + \sum_{k=2}^{\infty} \mathrm{Cov}(F(Y_k + s), F(Y_1 + t)), \end{aligned}$$

Now we investigate the asymptotic distribution of the empirical U-quantiles, defined as generalized inverses of the empirical U-distribution function, i.e. $Q_{n_1,n_2}(p) = U^{-1}_{n_1,n_2}(p)$.

Theorem 2.4 *Let $(X_i)_{i\geq 1}$ be a 1-approximating functional with constants $(a_k)_{k\geq 1}$ of an absolutely regular process with mixing coefficients $(\beta_k)_{k\geq 1}$ and assume that $\sum_{k=1}^{\infty} k^2 \left(a_k^{2/3} + \beta_k^{1/2}\right) < \infty$. Suppose moreover that $U(t)$ is differentiable in $Q(p)$ and Lipschitz-continous. Let $n_1, n_2 \to \infty$ in such a way that $\frac{n_1}{n_1+n_2} \to \lambda \in (0,1)$. Then*

$$\sqrt{n_1+n_2}(Q_{n_1,n_2}(p) - Q(p)) \longrightarrow \frac{1}{U'(Q(p))} W_{Q(p)},$$

where W_t is defined as in Theorem 2.2. The limit random variable has a normal distribution with mean zero and variance

$$\frac{\mathrm{Var}(W_{Q(p)})}{(U'(Q(p)))^2} = \frac{1}{(U'(Q(p)))^2}\left(\frac{\sigma_1^2(Q(p))}{\lambda} + \frac{\sigma_2^2(Q(p))}{1-\lambda}\right),$$

where $\sigma_1^2(Q(p))$ and $\sigma_2^2(Q(p))$ are defined as in Theorem 2.2.

The quantile process can be studied with the help of a Bahadur representation,

$$Q_{n_1,n_2}(p) = Q(p) + \frac{p - U_{n_1,n_2}(Q(p))}{U'(Q(p))} + R_n,$$

where $R_n(t)$ is a remainder term which has to be controlled. The Bahadur representation is motivated by the approximation of the derivative of the empirical U-distribution by the derivative of its limit, which is $U(t)$.

3 Statistical inference

Now we come back to the two-sample problem stated in the introduction, assuming that G is a shifted version of F, such that a $\Delta \in \mathbb{R}$ exists for which $G(x) = F(x + \Delta)$ for all $x \in \mathbb{R}$. This two-sample location problem generalizes the homoscedastic normal situation mentioned at the beginning of the introduction. We are interested in statistical inference for Δ and want to test the null hypothesis of equal levels, $H_0 : \Delta = 0$.

The classical estimator of Δ is the difference of the sample means, for which asymptotically

$$T^{(c)}(\Delta) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \frac{\bar{Y} - \bar{X} - \Delta}{\sqrt{\sum_{h=-\infty}^{\infty} \mathrm{Cov}(X_1, X_h)}} \overset{a}{\sim} N(0,1) \tag{7}$$

for a large class of weakly dependent stationary processes. In the case of our model (1), i.e. for two independent samples, this follows from the central limit theorem of Ibragimov and Linnik (1971), Theorem 18.6.2. In the case of model (2), one can apply an invariance principle of Philipp and Stout (1975), Theorem 7.1. Both references require the underlying processes to be functionals of strongly mixing processes, which is weaker than the assumptions made in this paper. A pivotal quantity for Δ, which generalizes the pivot underlying the two-sample t-test, can be derived by plugging in an estimator of the denominator $\sigma^2 = \sum_{h=-\infty}^{\infty} \mathrm{Cov}(X_1, X_h)$. In the setting of one sample $X_1, \ldots, X_n$, i.e. $n = n_1$, Peligrad and Shao (1995) suggest the estimators

$$\hat{\sigma}_p = \left(\frac{c_p}{n - l_n + 1} \sum_{j=0}^{n-l_n} \left(\frac{|S_j(l_n) - l_n \bar{X}|}{\sqrt{l_n}} \right)^p \right)^{1/p}$$

with $S_j(l_n) = \sum_{i=j+1}^{j+l_n} X_i$ and $c_p = 2^{-p/2}\sqrt{\pi}/\Gamma((p+1)/2)$ (e.g. $c_1 = \sqrt{\pi/2}$ and $c_2 = 1$). Assuming $(X_i : i \geq 1)$ to be a stationary ρ-mixing process with finite moments of order $2 \vee p$, $p \geq 1$, they show $\hat{\sigma}_p$ to be a consistent estimator of σ, provided that $l_n \to \infty$ and $l_n/n \to 0$. Their results indicate that the estimator $\hat{\sigma}_2$ with $p = 2$ has the smallest asymptotical variance within this class of estimators if X_1 possesses fourth moments. We call the resulting procedure a corrected t-test or a corrected mean difference test, with the latter name resembling the median difference test introduced below. Note that the condition of ρ-mixing is stronger than necessary for the derivation of the theory in Section 2. Doukhan, Jakubowicz and León (2010) treat the asymptotics of subsampling variance estimators related to $\hat{\sigma}_p$ under a broad range of weakly dependent processes.

Using Theorem 2.4, we can alternatively base our inference on the median difference $Q_{n_1,n_2}(0.5)$ between the samples. If $n_1, n_2 \to \infty$ and $\frac{n_1}{n_1+n_2} \to \lambda \in (0,1)$, we have

$$\begin{aligned} T^{(r)}(\Delta) &= H'(0)\sqrt{\frac{\lambda(1-\lambda)}{\sigma_1^2(0)}}\sqrt{n_1+n_2}(Q_{n_1,n_2}(0.5)-\Delta) \\ &= H'(0)\sqrt{\frac{n_1 n_2}{n_1+n_2}}\frac{Q_{n_1,n_2}(0.5)-\Delta}{\sqrt{\sigma_1^2(0)}} \overset{a}{\sim} N(0,1) \end{aligned}$$

where $\sigma_1^2(0) = \sum_{k=-\infty}^{\infty} \text{Cov}(F(X_1), F(X_k))$, which needs to be estimated. Again in the one sample setting, Dewan and Prakasa Rao (2003) show in case of an associated sequence that a consistent estimator of four times this sum is obtained replacing $S_j(l_n)$ by $\sum_{i=j+1}^{j+l_n}(1-2F_n(X_i))$ and $\bar{X}$ by $n^{-1}\sum_{i=1}^{n}(1-2F_n(X_i))$ in the above definition of $\hat{\sigma}_1$. Although the consistency of this estimator of $\sigma_1^2(0)$ has only been proved in case of positive dependencies, we apply it also in case of negative dependencies in the next section and obtain good empirical results.

In case of two independent large samples, model (1), we can pool the estimates $\hat{\sigma}_1^2(0)$ and $\hat{\sigma}_2^2(0)$ derived from $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$, respectively, and use their weighted sum $n_1\hat{\sigma}_1^2(0)/(n_1+n_2) + n_2\hat{\sigma}_2^2(0)/(n_1+n_2)$ as a pooled estimate of $\sigma_1^2(0) = \sigma_2^2(0)$ under H_0. In case of two dependent samples, model (2), we found a joint estimator derived from all observations without level correction, constructed under H_0, to give test sizes closer to the nominal significance levels.

For appropriate scaling, we need to estimate additionally the density $H'(0)$ at zero. We tried several kernel density estimates and will comment on three of them in the next section, all constructed using the R function *density* (R Development Core Team 2010) with the default Gaussian kernel and bandwidth selection as proposed by Sheather and Jones (1991). One of the estimates is constructed under H_0 and uses all pairwise differences within the full sample $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$ without level correction, while the other two are valid also under the alternative and based on all pairwise differences within the two samples separately or on all pairwise differences between the samples, correcting both samples by their sample medians.

The above pivots with $\Delta = 0$ and the mentioned scalings can be used to construct asymptotical significance tests for the null hypothesis H_0 of equal levels, $\Delta = 0$, and identical distributions, $F = G$. The arising test statistics $T^{(r)} = T^{(r)}(0)$ and $T^{(c)} = T^{(c)}(0)$

are compared to appropriate percentage points of the standard normal distribution. $T^{(r)}$ and $T^{(c)}$ are both derived from unbiased estimators of Δ if $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are identically distributed samples from F and G, respectively, and $n_1 = n_2$ or F and G are symmetric. Comparing the asymptotical powers of the two tests against local alternatives is then equivalent to comparing the asymptotical standard deviations of the underlying estimators. In case of independent observations, the asymptotic efficiency of the test based on $T^{(r)}$ is the same as that of the Wilcoxon-Mann-Whitney (WMW) test, which is locally more powerful than the test based on $T^{(c)}$ iff $\sqrt{12\gamma_0}H'(0) > 1$ (see e.g. Lehmann 1963). This is no surprise since the median difference is the location estimator corresponding to the WMW test. In the case of weakly dependent data considered here the comparison is more difficult since we need to take additionally the double infinite sums $\sum_{k=-\infty}^{\infty} \text{Cov}(X_1, X_k)$ and $\sum_{k=-\infty}^{\infty} \text{Cov}(F(X_1), F(X_k))$ into account.

Therefore we perform a simulation study in the next section, considering different distributions, different strengths of autocorrelation and situations with and without outliers. Additionally we include a modified version of the WMW test for dependent data into the comparison, which has been developed by Serfling (1968) for strongly mixing processes and verified by Dewan and Prakasa Rao (2003) for associated sequences. Although the test based on $T^{(r)}$ is asymptotically equivalent to the WMW test under our basic assumptions, we will observe some differences in case of finite samples and particularly in the presence of strongly deviating observations caused by skewness, heavy tails or outliers.

Another advantage of our approach based on the median difference is that an explicit asymptotical confidence interval for Δ can be constructed from the pivot $T^{(r)}(\Delta)$, namely

$$\left[Q_{n_1,n_2}(0.5) - z_{1-\alpha/2}\sqrt{\frac{n_1+n_2}{n_1n_2}}\frac{\hat{\sigma}_1(0)}{\hat{H}'(0)}, Q_{n_1,n_2}(0.5) + z_{1-\alpha/2}\sqrt{\frac{n_1+n_2}{n_1n_2}}\frac{\hat{\sigma}_1(0)}{\hat{H}'(0)}\right],$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$-quantile of the standard normal. Application of this asymptotic confidence interval is more convenient than an implicit approach, where we include all values of Δ not rejected by a WMW test applied to the modified samples, with Δ subtracted from $Y_1, \ldots, Y_{n_2}$.

4 Simulations

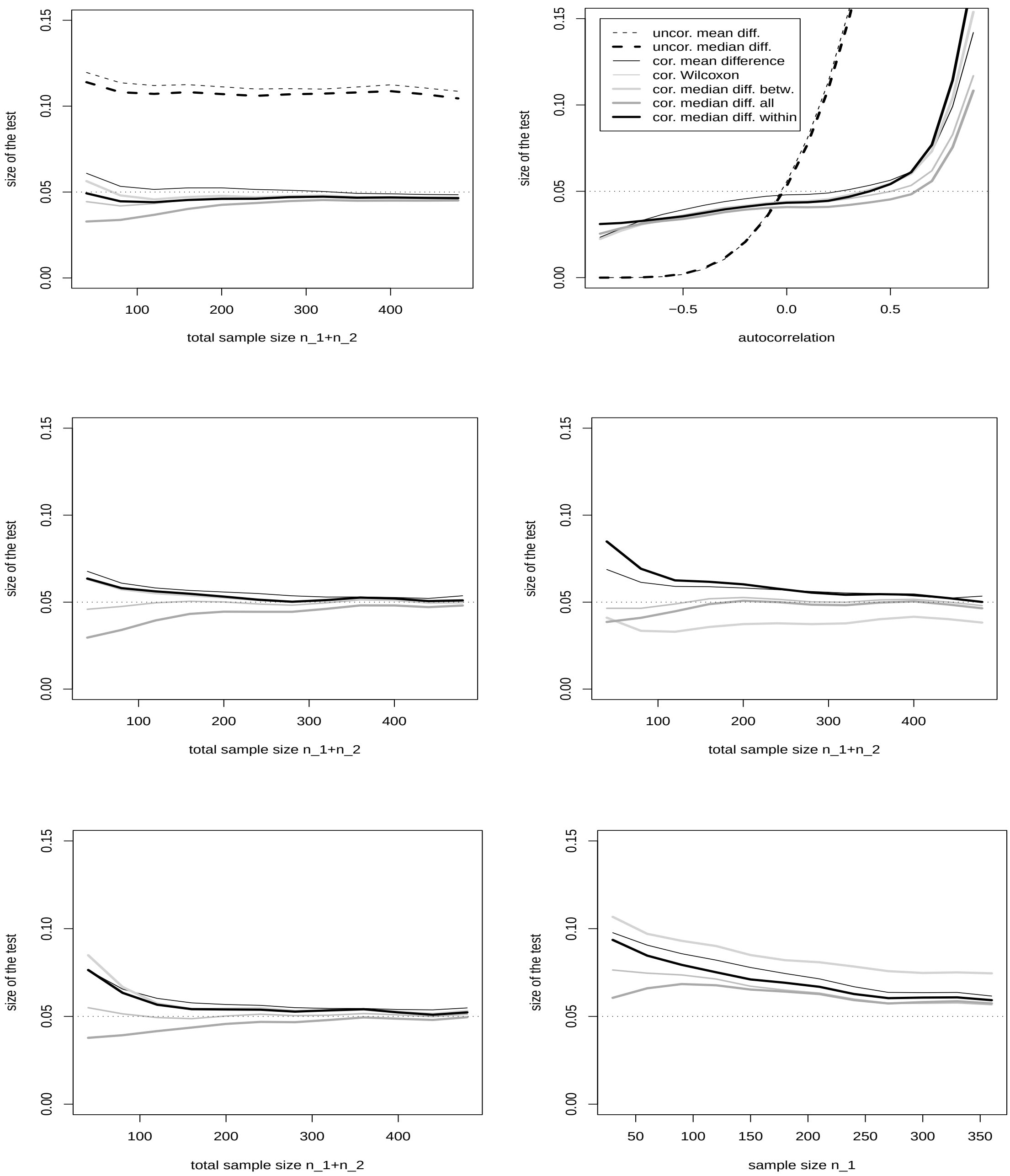
We concentrate on model (2) presented in the introduction, where $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ are subsequent stretches from the same time series. To check whether the asymptotical tests described above keep their nominal significance levels, we consider first order autoregressive (AR(1)) processes with different lag one autocorrelations ϕ and different continuous innovation distributions, generating 20000 time series for each of several lengths $n_1 + n_2$. We split the time series into two dependent samples of sizes n_1 and n_2 and derive the empirical rejection rates to estimate the sizes of the tests at a nominal .05 level, see Figure 1.

We observe that the two-sample t-test and the median difference test without correction for dependencies are severely oversized even in case of small autocorrelations, $\phi = 0.2$, and Gaussian innovations, whereas the asymptotic corrections work well already in case of moderate sample sizes $n_1 = n_2 = 50$ under these circumstances. Only the t-test for dependent data is slightly liberal even for $n_1 = n_2 = 100$ if $\phi = 0.2$. All tests become increasingly liberal with increasing value of ϕ. The median difference test with density estimation from all differences performs best in this respect and becomes liberal only for $\phi \geq 0.7$ in case of $n_1 = n_2 = 100$ and Gaussian innovations, while the WMW test becomes liberal slightly earlier for $\phi \geq 0.6$. In case of skewed χ_1^2-distributed innovations and $\phi = 0.5$, the corrected t-test and the test based on the median difference with density estimation from the pairwise differences within the samples perform slightly liberal even if $n_1 + n_2 = 400$. In case of heavy tailed t_3-distributed innovations or unequal sample sizes $n_1 = \lfloor 5n_2/3 \rfloor$, this also applies for the median difference test with density estimation from the corrected differences between the samples. When fixing $n_2 = 11$ and increasing only the size n_1 of the first sample, all tests perform liberal, but the size of the tests gradually decreases with n_1. The median difference test with density estimation from all differences without correction performs best then, jointly with the corrected WMW test, while the uncorrected tests (not shown here) lead to empirical sizes of about 0.25 throughout.

In summary, the median difference test with density estimation from all differences between the full set of all data points $x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$, without correcting for a difference in location, keeps the nominal significance level better than its competitors, including the corrected WMW test, but is sometimes somewhat conservative. Similarly, using all the observations without corrections gave the best scaling by the double infinite sum of autocovariances $\sum_{k=-\infty}^{\infty} \mathrm{Cov}(X_1, X_k)$ and $\sum_{k=-\infty}^{\infty} \mathrm{Cov}(F(X_1), F(X_k))$, respectively. An explanation is that we can use larger subsamples for the estimations when using the full data set. The correction appears to be more difficult for the t-test, as it is slightly oversized except for zero and negative values of ϕ. These results for the t-test corrected with $\hat{\sigma}_2$ still look better than those with correction by $\hat{\sigma}_1$ (not shown here).

To compare the power of the corrected tests in case of different values $\Delta \neq 0$, different innovation distributions, and different numbers and sizes of outliers, we generate 2000 time series for each of several scenarios. Figure 2 illustrates that the corrected t-test is only slightly more powerful than the corrected tests based on the median difference or the corrected WMW test in case of a Gaussian AR(1) model with $\phi = .5$. The small differences between the several versions of the median difference and the WMW test agree with their different empirical sizes, i.e. density estimation from the corrected pairwise differences between the samples or from the differences within the samples leads to slightly

Figure 1: Empirical sizes of the tests in case of AR(1) processes with different lag one correlations ϕ: $\phi = 0.2$ and increasing sample sizes $n_1 = n_2$ (top left) or different values of $\phi = -0.9, -0.8, \ldots, 0.9$, $n_1 = n_2 = 100$ (top right), both with Gaussian innovations; $\phi = 0.5$ with t_3- (center left) or χ_1^2-innovations (center right), both with growing sample sizes $n_1 = n_2$; $\phi = 0.5$ with Gaussian innovations and $n_1 = \lfloor 5n_2/3 \rfloor$ (bottom left) or $n_2 = 11$ fixed and n_1 increasing (bottom right).



more power than the WMW test, which in turn is slightly more powerful than the median difference test with density estimation from all pairs of observations. The t-test loses its superiority in case of a single outlier of size 10, heavy tailed t_3-innovations or skewed χ_1^2-innovations. The small differences between the median difference and the WMW tests are similar to the pure Gaussian case, except for the skewed case, where the median difference with density estimation from the corrected pairwise differences between the samples seems to be the best choice due to its better power than the WMW test and smaller size. Moreover, about five percent outliers of size 10 in one of the samples destroy the power of the t-test almost completely even if $\Delta = 1$ is rather large. The WMW and the median difference tests are more robust and show some power even if there are about 15 percent outliers in one sample, with the median difference tests performing somewhat better than the WMW test in case of isolated outliers. Note that about 20% outliers can lead to detection of a location difference into the wrong direction when using the corrected t-test test in this scenario.

Finally, Figure 3 illustrates the power of the tests in case of sample sizes $n_1 = 100$ and $n_2 = 11$, for $\phi = 0.5$. The small differences in case of Gaussian innovations again correspond to the somewhat different empirical sizes of the tests. As for the case $n_1 = n_2 = 100$, the corrected t-test is outperformed by the other tests in case of heavy tailed t_3-innovations. It loses all its power in case of a single very large outlier, and also in case of 5% moderately large patchy outliers in the larger sample. The median difference test outperforms the WMW test w.r.t. robustness against outliers here, particularly, if the corrected pairwise differences between the samples are used for density estimation. Similar to the case of two large samples, this option again leads to the best size-power behavior in case of skewed innovations: it is less oversized than the other tests in case of χ_1^2-distributed innovations, but nevertheless leads to larger power than the WMW test.

5 Application

We illustrate the performance of the tests by two applications: one for model (1) presented in the introduction and two large samples, and one where we apply the tests for sequential monitoring of a time series comparing the data in a short test window to a large reference window.

5.1 Application to climate data

The investigation of climate change is a major research topic nowadays. We analyze climate data from Potsdam, Germany, where one of the earliest weather stations founded reports daily weather data since 1893, with only a four day break in april 1945 because of the second world war. The temperature in Potsdam is known to match the average temperature

Figure 2: Power of the tests in case of jumps of increasing height in an AR(1) process with $\phi = 0.5$, $n_1 = n_2 = 100$: Gaussian innovations without (top left) and with one additive outlier of size 10 (top right); scaled t_3- (center left) and χ^2_1-innovations (center right); $\Delta = 1$ and Gaussian innovations with an increasing number of patchy (bottom left) or isolated outliers (bottom right) of size 10 in one sample.

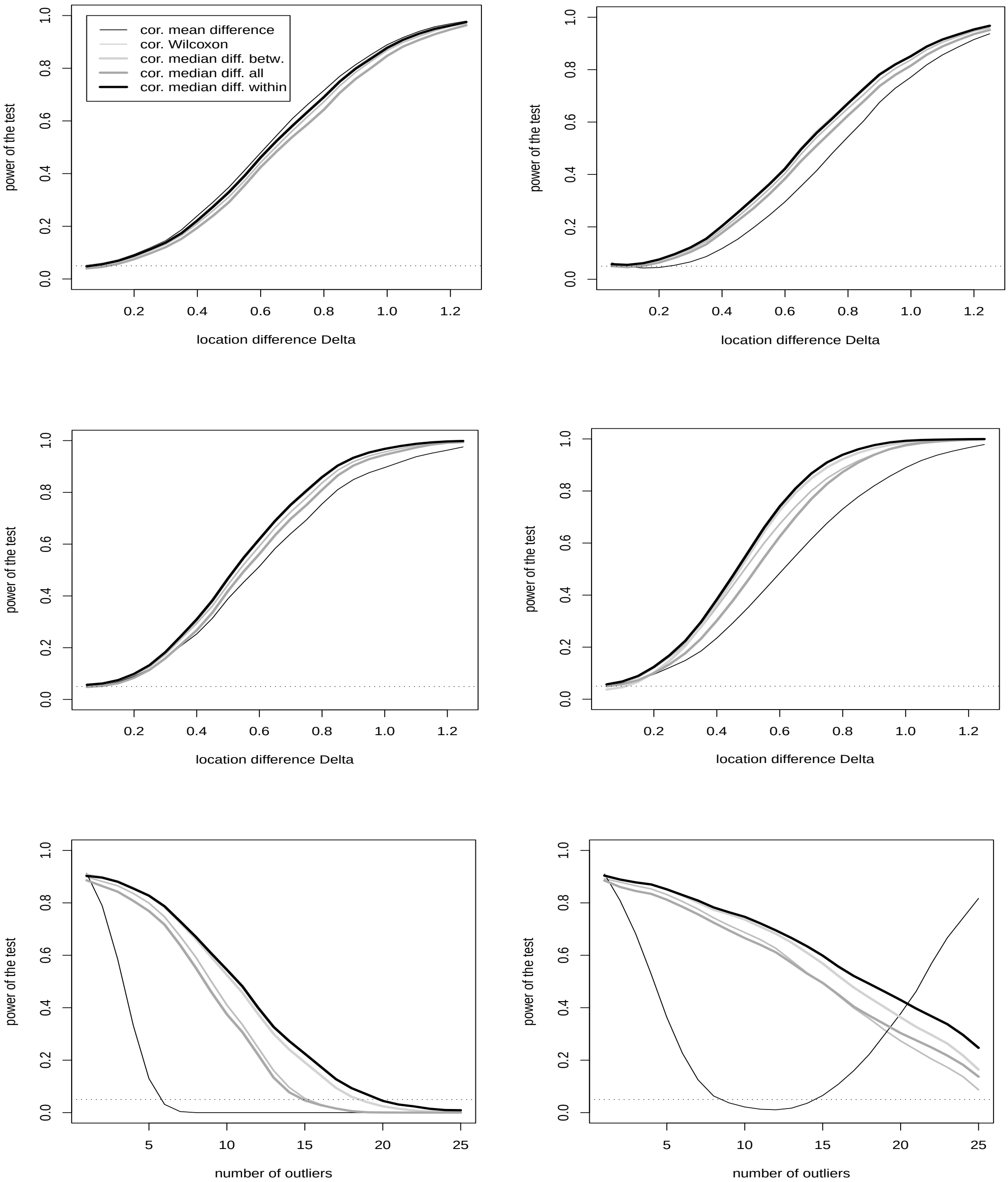
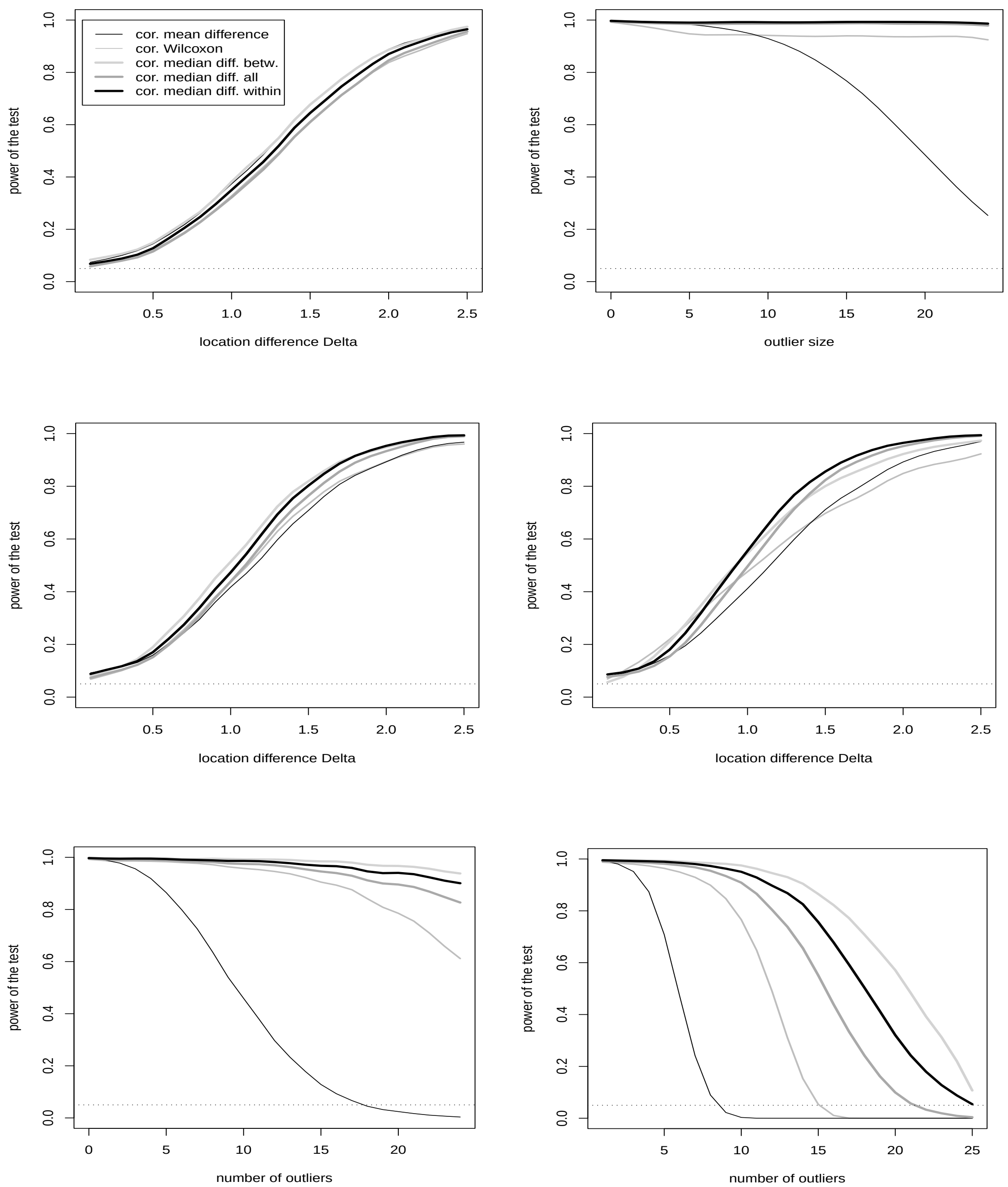


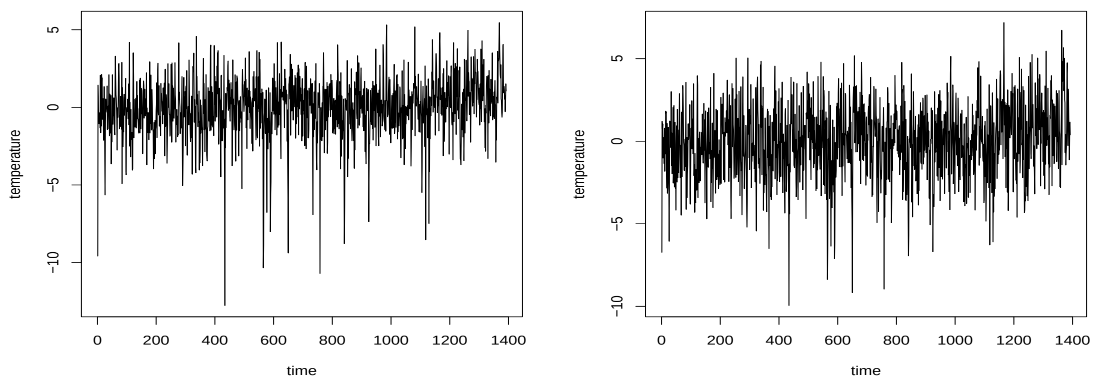
Figure 3: Power of the tests in case of jumps of increasing height in an AR(1) process with $\phi = 0.5$, $n_1 = 100$, $n_2 = 11$: Gaussian without (top left) and with one additive outlier of increasing size (top right); scaled t_3- (center left) or scaled χ_1^2-innovations (center right); Gaussian innovations with an increasing number of isolated (bottom left) or patchy outliers (bottom right) in the larger sample.



on the northern hemisphere quite well, so that these data are particularly interesting.

In a first step we aggregate the daily time series into monthly averages and deseasonalize the monthly data by subtracting the average of all values for the same calendar month. Figure 4 depicts the deseasonalized monthly temperatures resulting from the minimum and the maximum daily temperature. as well as the daily amplitude.

Figure 4: Saisonally adjusted monthly averages of daily temperature maxima (left) and minima (right).



In the following we test these two time series as well as the average daily amplitudes, i.e. the difference between the maximum and the minimum temperature, whether its level has changed significantly from the early to the late period of industrialization, comparing the 50 years of data corresponding to the period 1895-1944 and the 50 years from 1960 to 2009. The autocorrelation and partial autocorrelation functions suggest AR(1) models for both periods in each case, with AR(1) coefficients of about 0.25. The median difference test with any scaling (results for between sample scaling are shown here) gives results very similar to the WMW test for all three comparisons, and the results for the corrected t-test are neither very different, see Table 1: we find significant increases of both the maximum and minimum temperature of about 0.5 degrees on average, while there does not seem to be a change in the daily temperature amplitude. The confidence intervals for Δ obtained from the mean and from the median difference test closely agree. Confidence intervals would have been harder to obtain from the WMW test. Note that the uncorrected versions of the tests would have indicated highly significant differences due to ignoring the dependencies in the data.

To challenge the robustness of the tests, we replace some observations by the value -9.9. In the original data set, this value stands for missingness. This had been overlooked in our first data analysis, what resulted in some strange findings. Replacing an increasing number of observations at the start of the second period by the value -9.9 artificially creates some not very large outliers, which could be caused by encoding missingness or by measurement

Table 1: Test results and confidence intervals for the deseasonalized Potsdam climate time series at an asymptotic $\alpha = .05$ error level

Variable	$\bar{y} - \bar{x}$	CI	$T^{(c)}$	$Q(0.5)$	CI	$T^{(r)}$	WMW	$T^{(c)}_{unc.}$	$T^{(r)}_{unc.}$
Max.Temp.	.48	[.04 ; .92]	2.12	.47	[.04 , .90]	2.17	2.16	3.74	3.66
Min.Temp.	.52	[.12 ; .92]	2.55	.50	[.15 ; .86]	2.86	2.83	4.79	5.25
Amplitude	-.45	[-.21 ; .12]	-0.54	-.02	[-.17 ; .14]	-0.19	-0.19	-0.71	0.25

artifacts. In case of the average maximum temperatures, the corrected t-test no longer detects a shift at the 5% significance level, when two observations have been replaced. For the WMW test, six replacements are needed to make it non-significant, while the median difference test still resists this number and needs one more outlier. In case of the average minimum temperatures, the corrected t-test resists five such outliers before it becomes non-significant, while the WMW and the median difference test resist even 21 outliers.

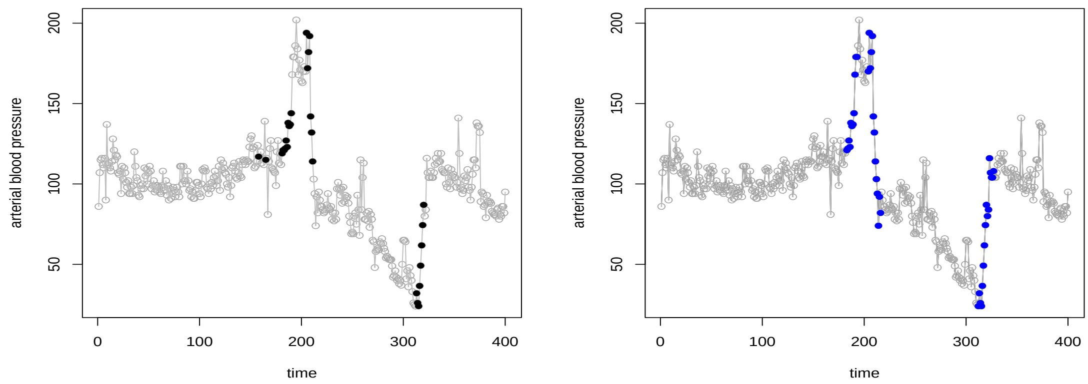
These results confirm that the corrected median difference test, like the corrected WMW test, performs very similarly to the corrected t-test in case of uncontaminated data sets which are not far from normality. Here, qqplots of the deseasonalized temperature data point at tails somewhat heavier than the Gaussian, so that the results of the nonparametric tests are more reliable. The next example illustrates that the differences between the methods are larger if at least one of the samples is small.

5.2 Application to intensive care data

As a second example we consider a monitoring time series representing the arterial blood pressure of a patient in intensive care, which is observed once a minute, see Figure 5. The interest here is detection of abrupt changes in the varying level of the time series with only short delays. Since there are periods with monotone trends and some outliers in the data, we first detrend the series by subtracting the online trend estimate provided by the weighted repeated median (Fried, Einbeck and Gather 2007) with triangular weights and a window width of $n = 60$ observations, corresponding to one hour of measuring. Then we apply the tests sequentially to the detrended time series in order to detect a jump after any of the time points not too close to the start or the end of the series. For this we compare the data in a moving test window consisting of the most recent $n_2 = 11$ observations to the data in a reference window with the $n_1 = 60$ immediately preceding observations. In this way the testing procedure can adapt to only locally valid stationarity conditions. A very small significance level of $\alpha = .001$ is chosen for the tests at each time point because of the multiple testing. Keeping the overall significance level is not the goal here, we just want to avoid getting many false alarms. Our primary objective is not to miss relevant jumps.

The corrected t-test detects not only the three obvious relevant shifts in the time series, but also alarms close to the occurrence of a few outliers at about time 170. As opposed to this, the corrected median difference test detects only the relevant changes. Given that the shifts occur in several subsequent steps, it even detects almost all of the time points where steps occur. The uncorrected tests (not shown here) give many false alarms even at this very small significance level, while the corrected WMW test does not detect any change at $\alpha = .001$, due to its inferior power and robustness when one of the windows is short.

Figure 5: Time series representing arterial pressure and time points where a corrected t-test (left) or a corrected median difference test (right) applied to the detrended series using moving windows of widths $n_1 = 60$ and $n_2 = 11$ detect a significant change at an $\alpha = .001$ significance level.



6 Conclusion

We investigate for the first time the asymptotic distribution of general two-sample U-statistics and empirical U-quantiles for dependent data. Our results hold for data that can be represented as functionals of an absolutely regular process. This class includes e.g. all short memory time series models as well as many chaotic dynamical systems.

Based on the derived asymptotical distributions and following Lehmann (1963), we have constructed an asymptotically distribution-free two-sample test for a difference in location between weakly dependent data sets based on the Hodges-Lehmann two-sample estimator, which is the median of all pairwise differences. This test, like Serfling's (1968) version of the Wilcoxon Mann Whitney test for dependent data, performs very similarly to a corrected t-test in case of uncontaminated data sets which are not far from normality. However, our simulations indicate that the median difference test is to be preferred to the other methods in case of heavy tails, skewness or outliers, although the differences to the WMW test become small if both samples are very large, according to the asymptotic equivalence of the tests. If at least one of the samples is only moderately large, the statistic of the WMW test takes a moderate number of different values only. Changing a single measurement a lot can strongly affect the significance of this test statistic. This is different for the median difference, so that the test based on it is more reliable for heavy tailed and skewed distributions, because observations largely deviating from the others are common in such situations. The price to be paid is an increase of computational costs, what is not a big problem in most applications nowadays.

Acknowledgements
The financial support of the Deutsche Forschungsgemeinschaft (SFB 823, "Statistical modelling of nonlinear dynamic processes") is gratefully acknowledged.

Appendix: Proofs of Theorems

6.1 Proof of Theorem 2.1

Here we study two-sample U-statistics of the type

$$U_{n_1,n_2}(h) = \frac{1}{n_1\, n_2} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n_1+n_2} h(X_i, X_j),$$

where $(X_i)_{i\geq 1}$ is a stationary process that can be represented as a functional of an absolutely regular process. From the Hoeffding decomposition (1) of the kernel $h(x,y)$ we obtain the Hoeffding decomposition of the two-sample U-statistic

$$U_{n_1,n_2} = \theta + \frac{1}{n_1} \sum_{i=1}^{n_1} h_1(X_i) + \frac{1}{n_2} \sum_{j=n_1+1}^{n_1+n_2} h_2(X_j) + \frac{1}{n_1\, n_2} \sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n_1+n_2} g(X_i, X_j). \tag{8}$$

Observe that, by definition, the functions $h_1(x)$, $h_2(y)$ and $g(x,y)$ have the properties

$$Eh_1(X) = Eh_2(Y) = 0, \tag{9}$$

$$Eg(X,y) = Eg(x,Y) = 0, \tag{10}$$

where X, Y are random variables with the same distribution as X_1. A kernel satisfying property (10) is called degenerate.

The proof of Theorem 2.1 consists of two parts. First we will show that the last term in the Hoeffding decomposition, i.e. $\sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n_1+n_2} g(X_i, X_j)$ is negligible compared to the other terms. In the case of independent observations, this follows from the fact that the summands $g(X_i, X_j)$ are uncorrelated and thus the variance of the last term equals $n_1\, n_2 \mathrm{Var}(g(X_1, X_2)$. In the case of dependent observations, estimation of the variance of $\sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n_1+n_2} g(X_i, X_j)$ is more difficult and requires subtle calculations. In the second part of the proof, we will apply the central limit theorem for partial sums of dependent data to the terms $\sum_{i=1}^{n_1} h_1(X_i)$ and $\sum_{i=n_1+1}^{n_1+n_2} h_2(X_i)$. Note that by (9), both are sums of mean-zero random variables.

Computing the 2nd moment of $\sum_{i=1}^{n_1} \sum_{j=n_1+1}^{n_1+n_2} g(X_i, X_j)$ leads to terms of the type

$$E(g(X_{i_1}, X_{j_1})g(X_{i_2}, X_{j_2}))$$

with indices $1 \leq i_1, i_2 \leq n_1$, $n_1 + 1 \leq j_1, j_2 \leq n_1 + n_2$. For independent processes $(X_i)_{i\geq1}$, these expectations are equal to zero unless $i_1 = i_2$ and $j_1 = j_2$. The following proposition provides a bound on these expectations in case $(X_i)_{i\geq1}$ is a functional of an absolutely regular process.

Proposition 6.1 *Let $(X_i)_{i\geq1}$ be a 1-approximating functional with constants $(a_k)_{k\geq1}$ of an absolutely regular process with mixing coefficients $(\beta_k)_{k\geq1}$ and let $g(x, y)$ be a 1-continuous bounded degenerate kernel. Then we have for all $1 \leq i_1, i_2 \leq n_1$, $n_1 + 1 \leq j_1, j_2 \leq n_1 + n_2$*

$$\begin{aligned} |E\left(g(X_{i_1}, X_{j_1})g(X_{i_2}, X_{j_2})\right)| \leq\ & 2M\phi(a_{[\max(|j_2-j_1|,|i_2-i_1|)/3]}) \\ & +4M^2(\sqrt{a_{[\max(|j_2-j_1|,|i_2-i_1|)/3]}} + \beta_{[\max(|j_2-j_1|,|i_2-i_1|)/3]}) \\ & +4\phi(\sqrt{a_{[(\min(j_1,j_2)-max(i_1,i_2))/3]}}) \\ & +8M^2(\sqrt{a_{[(\min(j_1,j_2)-\max(i_1,i_2))/3]}} + \beta_{[(\min(j_1,j_2)-\max(i_1,i_2))/3]}) \end{aligned}$$

where $M = \sup_{x,y} |g(x, y)|$.

PROOF. Let $(Y_j)_{j\geq1}$ be a copy of the $(X_j)_{j\geq1}$-process that is independent of $(X_j)_{j\geq1}$. Then

$$\begin{aligned} E\left(g(X_{i_1}, X_{j_1})g(X_{i_2}, X_{j_2})\right) =\ & E\left(g(Y_{i_1}, X_{j_1})g(Y_{i_2}, X_{j_2})\right) \\ & +E\left(g(X_{i_1}, X_{j_1})g(X_{i_2}, X_{j_2}) - g(Y_{i_1}, X_{j_1})g(Y_{i_2}, X_{j_2})\right). \end{aligned} \tag{11}$$

We now treat the two terms on the right hand side of (11) separately, beginning with the first term. Define $k := \max(|j_2 - j_1|, |i_2 - i_1|)$. Without loss of generality, we may assume that $k = j_2 - j_1$. According to Proposition 2.16 of Borovkova et al. (2001) copies $(X'_i)_{i\geq1}$ and $(X''_i)_{i\geq1}$ of the process $(X_i)_{i\geq1}$ exist with the following properties

(12) $(X''_i)_{i\geq1}$ is independent of $(X_i)_{i\geq1}$

(13) There exists a set A with $P(A) \geq 1 - \beta_{[k/3]}$ and $E(|X_{j_2} - X'_{j_2}|1_A) \leq 2a_{[k/3]}$

(14) $E(|X'_{j_1} - X''_{j_1}|) \leq 2a_{[k/3]}$

The processes $(X_i')_{i\geq 1}$ and $(X_i'')_{i\geq 1}$ can moreover be chosen independently of the process $(Y_j)_{j\geq 1}$. Then we get

$$\begin{aligned} E\left(g(Y_{i_1},X_{j_1})g(Y_{i_2},X_{j_2})\right) &= E\left(g(Y_{i_1},X'_{j_1})g(Y_{i_2},X'_{j_2})\right) \\ &= E\left(g(Y_{i_1},X''_{j_1})g(Y_{i_2},X_{j_2})\right) \\ &\quad +E\left(g(Y_{i_1},X'_{j_1})(g(Y_{i_2},X'_{j_2})-g(Y_{i_2},X_{j_2}))\right) \\ &\quad +E\left(g(Y_{i_2},X_{j_2})(g(Y_{i_1},X'_{j_1})-g(Y_{i_1},X''_{j_1}))\right) \\ &= E\left(g(Y_{i_1},X'_{j_1})(g(Y_{i_2},X'_{j_2})-g(Y_{i_2},X_{j_2}))\right) \\ &\quad +E\left(g(Y_{i_2},X_{j_2})(g(Y_{i_1},X'_{j_1})-g(Y_{i_1},X''_{j_1}))\right), \end{aligned}$$

because $g(x,y)$ is a degenerate kernel. Concerning the first term on the right hand side, we obtain

$$\begin{aligned} &E\left(g(Y_{i_1},X'_{j_1})(g(Y_{i_2},X'_{j_2})-g(Y_{i_2},X_{j_2}))\right) \\ &\leq ME|g(Y_{i_2},X'_{j_2})-g(Y_{i_2},X_{j_2})| \\ &\leq ME|g(Y_{i_2},X'_{j_2})-g(Y_{i_2},X_{j_2})|1_{\{|X'_{j_2}-X_{j_2}|\leq\sqrt{2a_{[k/3]}}\}}+2M^2P(|X'_{j_2}-X_{j_2}|>\sqrt{2a_{[k/3]}}) \\ &\leq M\phi(\sqrt{2a_{[k/3]}})+2M^2(\sqrt{2a_{[k/3]}}+\beta_{[k/3]}), \end{aligned}$$

where we have made use of the fact that

$$\begin{aligned} P(|X'_{j_2}-X_{j_2}|>\sqrt{2a_{[k/3]}}) &\leq P(|X'_{j_2}-X_{j_2}|1_A>\sqrt{2a_{[k/3]}})+P(A^c) \\ &\leq \sqrt{2a_{[k/3]}}+\beta_{[k/3]}. \end{aligned}$$

In a similar way, we obtain for the second term

$$E\left(g(Y_{i_2},X_{j_2})(g(Y_{i_1},X'_{j_1})-g(Y_{i_1},X''_{j_1}))\right)\leq M\phi(\sqrt{2a_{k/3}})+2M^2\sqrt{2a_{k/3}}.$$

We now consider the second term on the r.h.s. of (11). We define $l=\min(j_1,j_2)-\max(i_1,i_2)$, and we assume without loss of generality that $l=j_1-i_2$. Applying Proposition 2.16 of Borovkova et al. (2001), we obtain copies $(X_j')_{j\geq 1}$ and $(X_j'')_{i\geq 1}$ of the original process $(X_j)_{j\geq 1}$ such that $(X_i'')_{i\geq 1}$ is independent of $(X_i)_{i\geq 1}$ satisfying

$$E(|X'_{i_1}-X''_{i_2}| \leq 2a_{[l/3]} \tag{15}$$

$$E(|X'_{i_1}-X''_{i_2}| \leq 2a_{[l/3]} \tag{16}$$

$$E(|X_{j_1}-X_{j_1'}|1_A) \leq 2a_{[l/3]} \tag{17}$$

$$E(|X_{j_2}-X_{j_2'}|1_A) \leq 2a_{[l/3]} \tag{18}$$

for some set A with $P(A)\geq 1-\beta_{[l/3]}$. Thus we obtain

$$
\begin{aligned}
&E\left(g(X_{i_1},X_{j_1})g(X_{i_2},X_{j_2}) - g(Y_{i_1},X_{j_1})g(Y_{i_2},X_{j_2})\right)\\
&= E\left(g(X'_{i_1},X'_{j_1})g(X'_{i_2},X'_{j_2}) - g(X''_{i_1},X_{j_1})g(X''_{i_2},X_{j_2})\right)\\
&= E\left((g(X'_{i_1},X'_{j_1}) - g(X''_{i_1},X_{j_1}))g(X''_{i_2},X_{j_2})\right) + E\left(g(X'_{i_1},X'_{j_1})(g(X'_{i_2},X'_{j_2}) - g(X''_{i_2},X_{j_2}))\right)\\
&\le ME|g(X'_{i_1},X'_{j_1}) - g(X''_{i_1},X_{j_1})| + ME|g(X'_{i_2},X'_{j_2}) - g(X''_{i_2},X_{j_2})|\\
&\le M(E|g(X'_{i_1},X'_{j_1}) - g(X'_{i_1},X_{j_1})| + E|g(X'_{i_1},X_{j_1}) - g(X''_{i_1},X_{j_1})|\\
&\quad +E|g(X'_{i_2},X'_{j_2}) - g(X'_{i_2},X_{j_2})| + E|g(X'_{i_2},X_{j_1}) - g(X''_{i_2},X_{j_2})|).
\end{aligned}
$$

In the final step we consider the four terms on the r.h.s. separately. Each of the terms can be treated with similar arguments; we give the details for the first term. Let $D = \{|X_{j_1} - X'_{j_1}| \ge \sqrt{2a_{[l/3]}}\}$ and note that $P(D) \le \sqrt{2a_{[l/3]}} + \beta_{[l/3]}$. Thus we get

$$
\begin{aligned}
E|g(X'_{i_1},X'_{j_1}) - g(X'_{i_1},X_{j_1})| &= E|g(X'_{i_1},X'_{j_1}) - g(X'_{i_1},X_{j_1})|1_{D^c} + 2\,M\,P(D)\\
&\le \phi(\sqrt{2a_{[l/3]}}) + 2M(\sqrt{a_{[l/3]}} + \beta_{[l/3]}).
\end{aligned}
$$

and thus the proposition is proved. □

Proposition 6.2 *Let $(X_i)_{i\ge1}$ be a 1-approximating functional with constants $(a_k)_{k\ge1}$ of an absolutely regular process with mixing coefficients $(\beta_k)_{k\ge1}$ and let $g(x,y)$ be a 1-continuous bounded degenerate kernel. Assume moreover that $\sum_{k=1}^{\infty} k\,(\beta_k+\sqrt{a_k}+\phi(a_k)) < \infty$. Then we have for some constant C, not depending on n_1 and n_2, that*

$$
E\left(\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2} g(X_i,X_j)\right)^2 \le C\,n_1\,n_2 \tag{19}
$$

PROOF. We can write

$$
\begin{aligned}
E\left(\sum_{j=n_1+1}^{n_1+n_2} g(X_i,X_j)\right)^2 &= \sum_{1\le i_1,i_2\le n_1}\ \sum_{n_1+1\le j_1,j_2\le n_1+n_2} E\left(g(X_{i_1},X_{j_1})g(X_{i_2},X_{j_2})\right)\\
&= \sum_{j=n_1+1}^{n_1+n_2} E(g(X_i,X_j))^2 + {\sum}' E\left(g(X_{i_1},X_{j_1})g(X_{i_2},X_{j_2})\right),
\end{aligned}
$$

where $\sum'$ indicates that we are taking the sum over all indices $1 \le i_1, i_2 \le n_1 < j_1, j_2 \le n_1 + n_2$ satisfying $i_1 \ne i_2$ or $j_1 \ne j_2$. To each of the summands in $\sum'$ we can apply

Proposition 6.1. We thus obtain

$$\begin{aligned}
&\sum{}' |E\left(g(X_{i_1},X_{j_1})g(X_{i_2},X_{j_2})\right)| \\
&\le 2M \sum_{1\le i_1,i_2\le n_1}\ \sum_{n_1+1\le j_1,j_2\le n_1+n_2} \phi(a_{[\max(|j_2-j_1|,|i_2-i_1|)/3]}) \\
&+4M^2 \sum_{1\le i_1,i_2\le n_1}\ \sum_{n_1+1\le j_1,j_2\le n_1+n_2} \left(\sqrt{a_{[\max(|j_2-j_1|,|i_2-i_1|)/3]}}+\beta_{[\max(|j_2-j_1|,|i_2-i_1|)/3]}\right) \\
&+4M \sum_{1\le i_1,i_2\le n_1}\ \sum_{n_1+1\le j_1,j_2\le n_1+n_2} \phi(\sqrt{a_{[(\min(j_1,j_2)-max(i_1,i_2))/3]}}) \\
&+8M^2 \sum_{1\le i_1,i_2\le n_1}\ \sum_{n_1+1\le j_1,j_2\le n_1+n_2} \left(\sqrt{a_{[(\min(j_1,j_2)-\max(i_1,i_2))/3]}}+\beta_{[(\min(j_1,j_2)-\max(i_1,i_2))/3]}\right).
\end{aligned}$$

We now treat the first sum. Define $k = \max(|i_2 - i_1|, |j_2 - j_1|$ and keep for the moment k fixed. Then $k = |i_2 - i_1|$ or $k = |j_2 - j_1|$. In the first case, there are at most n_1 ways to choose i_1 and then exactly 2 ways to choose i_2. Concerning j_1, we have n_2 to pick this index and then at most k ways to choose j_2. Similarly we can bound the number of indices if $k = |j_2 - j_1|$. Finally we get

$$\sum_{1\le i_1,i_2\le n_1}\ \sum_{n_1+1\le j_1,j_2\le n_1+n_2} \phi(a_{[\max(|j_2-j_1|,|i_2-i_1|)/3]}) \quad \le 2\,n_1\,n_2 \sum_{k=1}^{\infty} k\,\phi(a_{[k/3]})$$

Using the assumptions of the proposition, we finally obtain the stated result. □

We can now finish the proof of Theorem 2.1. By the Hoeffding decomposition (8) we have

$$\begin{aligned}
\sqrt{n_1+n_2}(U_{n_1,n_2}-\theta) \;&=\; \sqrt{\frac{n_1+n_2}{n_1}}\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} h_1(X_i) + \sqrt{\frac{n_1+n_2}{n_2}}\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n_1+n_2} h_2(X_i) \\
&+\frac{\sqrt{n_1+n_2}}{n_1\,n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2} g(X_i,X_j).
\end{aligned}$$

By Poposition 6.2 we have

$$E\left(\frac{\sqrt{n_1+n_2}}{n_1\,n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2} g(X_i,X_j)\right)^2 \le C\frac{n_1+n_2}{n_1\,n_2} = C\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\to 0,$$

as $n_1, n_2 \to \infty$. Thus the remainder term in the Hoeffding decomposition converges to 0 in probability. To the linear terms we can apply the central limit theorem for partial sums of functionals of absolutely regular processes; see Theorem 4 in Borovkova et al. (2001). We get

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} h_1(X_i) \to N(0, \mathrm{Var}(h_1(X_1)) + 2\sum_{i=1}^{\infty}\mathrm{Cov}(h_1(X_1),h_1(X_{i+1})),$$

and similarly for the second sum. Moreover, $\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} h_1(X_i)$ and $\frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n_1+n_2} h_2(X_i)$ are asymptotically independent and thus we obtain joint convergence

$$\left(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} h_1(X_i), \frac{1}{\sqrt{n_2}}\sum_{i=n_1+1}^{n_1+n_2} h_2(X_i)\right) \to (Z_1, Z_2),$$

where Z_1 and Z_2 are two independent normally distributed random variables with mean zero and the above variances. □

6.2 Proof of Theorem 2.2

Here we study the empirical U-distribution function

$$\begin{aligned} U_{n_1,n_2}(t) &= \frac{1}{n_1\, n_2}\#\{1 \le i \le n_1, 1 \le j \le n_2 : f(X_i, X_j) \le t\} \\ &= \frac{1}{n_1\, n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} 1_{\{f(X_i,X_j)\le t\}}. \end{aligned}$$

We will first show convergence of the one-dimensional marginals, i.e.

$$\sqrt{n_1+n_2}(U_{n_1,n_2}(t) - U(t)) \xrightarrow{\mathcal{D}} W_t. \tag{20}$$

Convergence of all finite-dimensional marginals follows with the help of the Cramér-Wold device. In order to prove (20), we observe that $U_{n_1,n_2}(t)$ is a two-sample U-statistic with kernel $h_t(x,y) = 1_{\{f(x,y)\le t\}}$. The elements of the Hoeffding decomposition of $h_t(x,y)$ are

$$\begin{aligned} \theta &= E(1\{f(X,Y) \le t\}) = P(f(X,Y) \le t) = U(t) \\ h_{t,1}(x) &= P(f(x,Y) \le t) - U(t) = H_{t,1}(x) - U(t) \\ h_{t,2}(y) &= P(f(X,y) \le t) - U(t) = H_{t,2}(y) - U(t) \end{aligned}$$

Now (20) follows directly from Theorem 2.1.

6.3 Bahadur Representation of Two-Sample U-Quantiles

In this section we will consider the two-sample U-quantiles

$$Q_{n_1,n_2}(p) = \inf\{t : U_{n_1,n_2}(t) \ge p\}.$$

Recall that $Q(p)$ is the generalized inverse of $U(t)$. First we derive the Bahadur representation of $Q_{n_1,n_2}(p)$; this will later be our main technical tool in the proof of the asymptotic normality of $Q_{n_1,n_2}(p)$.

Theorem 6.3 *Let $(X_i)_{i\geq 0}$ be a stationary, absolutely regular process with mixing coefficients $\beta(k)$ satisfying $\sum_{k=1}^\infty k\beta(k) < \infty$. Then for any $0 < p < 1$ we have*

$$Q_{n_1,n_2}(p) = Q(p) + \frac{p - U_{n_1,n_2}(Q(p))}{U'(Q(p))} + R_{n_1,n_2} \tag{21}$$

where $R_{n_1,n_2} = o_P(\frac{1}{\sqrt{n_1+n_2}})$.

PROOF. Following the lines of the proof of Ghosh (1971), we have to show that for all $t \in \mathbb{R}$

$$\sqrt{n_1+n_2}\left(\left(U_{n_1,n_2}\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right) - U_{n_1,n_2}(Q(p))\right) - \left(U\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right) - U(Q(p))\right)\right)$$

converges to zero in probability. We introduce for abbreviation the kernel

$$h_{t,p}(x,y) = 1_{\{Q(p)<f(x,y)\leq Q(p)+\frac{t}{\sqrt{n_1+n_2}}\}}$$

and note that

$$\begin{aligned}&\left(U_{n_1,n_2}\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right) - U_{n_1,n_2}(Q(p))\right) - \left(U\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right) - U(Q(p))\right)\\ &= \frac{1}{n_1\,n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2}\left(h_{t,p}(X_i,X_j) - E(h_{t,p}(X,Y))\right).\end{aligned}$$

We then apply the Hoeffding decomposition to the kernel $h_{t,p}(x,y)$ and denote the resulting functions by $h_{t,p,1}(x)$, $h_{t,p,2}(y)$ and $g_{t,p}(x,y)$. Thus we get

$$\begin{aligned}&\left(U_{n_1,n_2}\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right) - U_{n_1,n_2}(Q(p))\right) - \left(U\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right) - U(Q(p))\right)\\ &= \frac{1}{n_1}\sum_{i=1}^{n_1} h_{t,p,1}(X_i) + \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} h_{t,p,2}(X_i) + \frac{1}{n_1\,n_2}\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2} g_{t,p}(X_i,X_j).\end{aligned}$$

Observe that the summands in the linear terms have mean zero and that $g_{t,p}$ is degenerate. Using Lemma 3.1 of Borovkova et al (2001), with $\delta = 2$, we obtain

$$\begin{aligned}E\left(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} h_{t,p,1}(X_i)\right)^2 &\leq \left(E\left(\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1} h_{t,p,1}(X_i)\right)^4\right)^{\frac{1}{2}}\\ &\leq C\,\left(E\left(h_{t,p,1}(X)\right)^4\right)^{\frac{1}{6}} \leq C\,\left(E(h_{t,p,1}(X))^2\right)^{\frac{1}{6}},\end{aligned}$$

where the second inequality follows from the fact that $0 \leq h_{t,p}(x,y) \leq 1$. By definition of $h_{t,p,1}(x)$, we get

$$\begin{aligned}E(h_{t,p,1}(X))^2 &\leq \int\left(\int h_{t,p}(x,y)dP_Y(y)\right)^2 dP_X(x)\\ &\leq P\left(Q(p) < f(X,Y) \leq Q(p) + \frac{1}{\sqrt{n_1+n_2}}\right)\\ &= U\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right) - U(Q(p)).\end{aligned}$$

Thus we get, in total,

$$E\left(\frac{\sqrt{n_1+n_2}}{n_1}\sum_{i=1}^{n_1} h_{t,p,1}(X_i)\right)^2 \le \frac{n_1+n_2}{n_1}\left(U\left(Q(p)+\frac{t}{\sqrt{n_1+n_2}}\right)-U(Q(p))\right)^{1/6} \to 0,$$

as $n_1, n_2 \to \infty$. In the same way, we can show that $E\left(\frac{\sqrt{n_1+n_2}}{n_2}\sum_{i=n_1+1}^{n_1+n_2} h_{t,p,2}(X_i)\right)^2 \to 0$. Concerning the third term in the Hoeffding decomposition, we can apply the same arguments as in the proof of Proposition 6.2 and get $E\left(\sum_{i=1}^{n_1}\sum_{j=n_1+1}^{n_1+n_2} g_{t,p}(X_i, X_j)\right)^2 = O(n_1\, n_2)$ and thus

$$E\left(\frac{\sqrt{n_1+n_2}}{n_1\, n_2}\sum_{j=n_1+1}^{n_1+n_2} g_{t,p}(X_i, X_j)\right)^2 = O\left(\frac{n_1+n_2}{n_1\, n_2}\right).$$

Hence all three terms in the Hoeffding decomposition converge to zero in L_2 and thus in probability. □

6.4 Proof of Theorem 2.4:

By the Bahadur representation (21) we obtain

$$\begin{aligned}\sqrt{n_1+n_2}(Q_{n_1,n_2}(p)-Q(p)) &= -\frac{1}{U'(Q(p))}\sqrt{n_1+n_2}\,(U_{n_1,n_2}(Q(p))-U(Q(p))) \\ &\quad +\sqrt{n_1+n_2}R_{n_1,n_2}\end{aligned}$$

By Theorem 2.2 we obtain $\sqrt{n_1+n_2}(U_{n_1,n_2}(Q(p))-U(Q(p))) \to W_{Q(p)}$. By Theorem 6.3 we get $\sqrt{n_1+n_2}R_{n_1,n_2} \to 0$. Thus Theorem 2.4 follows from an application of Slutzky's lemma. □

REFERENCES

Arcones, M., and Yu, B. (1994), "Central Limit Theorems for Empirical and U-processes of Stationary Mixing Sequences," *Journal of Theoretical Probability*, 7, 47–71.

Borovkova, S. A., Burton, R. M., and Dehling, H. G. (2001), "Limit Theorems for Functionals of Mixing Processes with Applications to U-statistics and Dimension Estimation," *Transactions of the American Mathematical Society*, 353, 4261–4318.

Dehling, H., and Wendler, M. (2010), "Central Limit Theorem and the Bootstrap for U-Statistics of Strongly Mixing Data", *Journal of Multivariate Analysis*, 101, 126–137.

Denker, M., and Keller, G. (1983), "On U-Statistics and Van Mises Statistics for Weakly Dependent Processes", *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 64, 505–522.

Denker, M., and Keller, G. (1986), "Rigorous Statistical Procedures for Data from Dynamic-Systems", *Journal of Statistical Physics*, 44, 67–93.

Dewan, I., and Prakasa Rao, B. L. S. (2003), "Mann-Whitney test for associated sequences", *Ann. Inst. Statist. Math.*, 55, 111–120.

Doukhan, P., Jakubowicz, J., and León, J. R. (2010), "Variance Estimation with Applications", in *Dependence in Probability, Analysis and Number Theory*, eds. I. Berkes, R. C. Bradley, H. Dehling, M. Peligrad, and R. Tichy, Heber City: Kendrick Press, pp. 203–231.

Fried, R., Einbeck, J., and Gather, U. (2007), "Weighted Repeated Median Smoothing and Filtering", *Journal of the American Statistical Association*, 480, 1300–1308.

Ghosh, J. K. (1971), "A New Proof of the Bahadur Representation of Quantiles and an Application", *The Annals of Mathematical Statistics*, 42, 1957–1961.

Hodges, J. L., and Lehmann, E. L. (1963), "Estimates of location based on rank tests", *Ann. Math. Statist.*, 34, 598–611.

Ibragimov, I. A., and Linnik, Yu. V. (1971), *Independent and Stationary Sequences of Random Variables*, Wolters-Noordhoff, Groningen.

Lehmann, E. L. (1963), "Nonparametric Confidence Intervals for a Shift Parameter", *The Annals of Mathematical Statistics*, 34, 1507–1512.

Peligrad, M., and Shao, Q.-M. (1995), "Estimation of the Variance of Partial Sums for ρ-mixing random variables", *Journal of Multivariate Analysis*, 52, 140–157.

Philipp, W., and Stout, W. F. (1975), *Almost sure invariance principles for partial sums of weakly dependent random variables*, Memoirs of the Amer. Math. Soc., 161.

R Development Core Team (2010), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Serfling, R.J. (1968), "The Wilcoxon Two-sample Statistic on Strongly Mixing Processes", *Ann. Math. Statist.*, 39, 1202–1209.

Serfling, R. (1984), "Generalized L-, M- and R-Statistics", *The Annals of Statistics*, 12, 76–86.

Sheather, S. J., and Jones M. C. (1991), "A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation", *J. Roy. Statist. Soc. B*, 683–690.

Yoshihara, K. I. (1976), "Limiting Behavior of U-Statistics for Stationary, Absolutely Regular Processes", *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35, 237–252.