Robust change point tests by bounded transformations

Alexander Dürre*a, Roland Frieda

*aFakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany

Abstract
Classical moment based change point tests like the cusum test are very powerful in case of Gaussian time series with one change point but behave poorly under heavy tailed distributions and corrupted data. A new class of robust change point tests based on cusum statistics of robustly transformed observations is proposed. This framework is quite flexible, depending on the used transformation one can detect for instance changes in the mean, scale or dependence of a possibly multivariate time series. Simulations indicate that this approach is very powerful in detecting changes in the marginal variance of ARCH processes and outperforms existing proposals for detecting structural breaks in the dependence structure of heavy tailed multivariate time series.

Keywords: short range dependence, covariance inequalities, median absolute deviation, psi functions, Robustness
2010 MSC: 62G35, 62G20, 62M10

1. Introduction
There is a fast growing literature on robust change-point detection. Maybe the first contribution is by Page (1955) who proposed a sign based test, interestingly not with robustness in mind, but to get a distribution free procedure. Popular robust procedures are based on signs (McGilchrist and Woodyer, 1975; Vogel and Fried, 2015), ranks (Bhattacharyya and Johnson, 1968; Horvath and Parzen, 1994; Gombay and Hušková, 1998; Antoch et al., 2008), U-statistics (Pettitt, 1979; Csörgő and Horváth, 1988; Gombay and Horváth, 1995; Horváth and Shao, 1996; Dehling et al., 2015a,b; Vogel and Wendler, 2015), quantiles (Csörgő and Horváth, 1987), M-estimators (Kumar Sen, 1984; Hušková, 1996; Fiteni, 2002; Hušková and Marnišiaková, 2012a) or distribution functions (Deshayes and Picard, 1986).
We propose a general framework for robust change-point detection, using cusum statistics on robustly transformed observations. Denote therefore $X_1, \ldots, X_T$ a $p$-dimensional time series. We assume it to be stationary and short range dependent under the null hypothesis, see Section 2 for the technical assumptions. If one is interested in a change of location and expects at most one change point it is common to look at the cusum statistic which for $p = 1$ is defined as

$$W_T(x) = \frac{1}{\hat{v}} \left( \sum_{i=1}^{\lfloor T x \rfloor} X_i - \frac{\lfloor T x \rfloor}{T} \sum_{i=1}^{T} X_i \right)$$

where $\hat{v}$ is a consistent estimator of the long run variance $v = \sum_{h=-\infty}^{\infty} \text{Cov}(X_i, X_{i+h})$ and rejects the null-hypothesis if functionals of $(W_T(x))_{x \in [0,1]}$ such as $\sup_{x \in [0,1]} |W_T(x)|$ are unusually large.

*corresponding author: alexander.duerre@udo.edu, +49-231-755-4288
If \( p > 1 \) one often modifies (1) to the quadratic form

\[
W_T^2(x) = \frac{1}{T} \left( \sum_{i=1}^{[T_x]} X_i - \frac{T_x}{n} \sum_{i=1}^{n} X_i \right)^T \hat{V}^{-1} \left( \sum_{i=1}^{[T_x]} X_i - \frac{T_x}{n} \sum_{i=1}^{T} X_i \right)
\]  

(2)

where \( \hat{V} \) consistently estimates \( V = \sum_{h=-\infty}^{\infty} \text{Cov}(X_i, X_{i+h}) \). To gain robustness we will transform the observations by a function \( \Psi : R^p \rightarrow R^q \) which can be of very different kind to detect a shift for example in location, scale, skewness or dependence. Formal conditions on \( \Psi \) can be found in Section 2. The main feature they have in common is the boundedness, we require that \( |\Psi(x)|_\infty \leq b \) for some \( b \in R \), where \( |.|_\infty \) denotes the maximum norm. This property is more an advantage than a restriction. By applying a bounded \( \Psi \) we also bound the influence of outlying observations from another probability model than the bulk of the observations which could ruin the inference otherwise. An individual outlier can either fake a change-point although the other observations are stationary or also hide a true change-point. Furthermore a bounded \( \Psi \) is also advantageous under heavy tailed distributions. Table 1 contains a selection of useful \( \Psi \)-functions.

Note that these transformations need to be applied to properly standardized random variables. For example the univariate sign function applied to strictly positive random variables will destroy any information of the dataset. To prohibit such behavior we standardize the data marginally by location and scale estimators \( \hat{\mu}^{(j)}, \hat{\sigma}^{(j)}, j = 1, \ldots, p \) and then apply the transformation \( \Psi \):

\[
Y_{i,T} = \Psi(\{X_i^{(1)} - \hat{\mu}^{(1)}\}/\hat{\sigma}^{(1)}, \ldots, \{X_i^{(k)} - \hat{\mu}^{(k)}\}/\hat{\sigma}^{(k)})' = \Psi(D_{\sigma}^{-1}[X_i - \hat{\mu}]),
\]

(3)

where \( D_{\sigma} \) is the diagonal matrix containing \( \hat{\sigma}_1, \ldots, \hat{\sigma}_p \) and \( \hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_p)' \). For now we only assume that these estimators converge in probability to some population values \( \mu_j \) and \( \sigma_j \) for \( j = 1, \ldots, k \) and postpone the discussion of the theoretical properties to Section 2. In practice we recommend the highly robust median and median absolute deviation. After the transformation we apply the quadratic cusum statistic (2) to the transformed values:

\[
W_T^2(x) = \frac{1}{T} \left( \sum_{i=1}^{[T_x]} Y_{i,T} - \frac{T_x}{T} \sum_{i=1}^{T} Y_{i,T} \right)^T \hat{U}^{-1} \left( \sum_{i=1}^{[T_x]} Y_{i,T} - \frac{T_x}{T} \sum_{i=1}^{T} Y_{i,T} \right).
\]

(4)

Here \( \hat{U} \) is an estimator for the respective long run variance \( U = \text{Cov} \sum_{h=-\infty}^{\infty} \text{Cov}(Y_i, Y_{i+h}) \), where

\[
Y_i = \Psi(\{X_i^{(1)} - \mu^{(1)}\}/\sigma^{(1)}, \ldots, \{X_i^{(k)} - \mu^{(k)}\}/\sigma^{(k)})' = \Psi(D_{\sigma}^{-1}[X_i - \mu])
\]

Table 1: Selection of useful transformations \( \Psi \). The function \( s : R^p \rightarrow R^p \) with \( s(x) = x/|x| \) for \( x \neq 0 \) and \( s(x) = 0 \) for \( x = 0 \) denotes the spatial sign.
with \( D_\sigma = \text{diag}(\sigma_1, \ldots, \sigma_p) \) and \( \mu = (\mu_1, \ldots, \mu_p)' \). One can either choose bootstrap methods (Härdle et al., 2003; Bühlmann, 2002; Kreiss and Paparoditis, 2011), sub-sampling (Dehling et al., 2013) or kernel estimators (Parzen, 1957; Andrews, 1991). We propose the latter. For a bandwidth \( b_T \in \mathbb{R}_+ \) and a kernel function \( k: \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) it is defined as

\[
\hat{U} = \frac{1}{T} \sum_{i,j=1}^{T} (Y_{i,T} - \bar{Y}_T)(Y_{j,T} - \bar{Y}_T)^T k \left( \frac{|i-j|}{b_T} \right)
\]

where \( \bar{Y}_T \) is the arithmetic mean of \( Y_{1,T}, \ldots, Y_{T,T} \). We use the flat-top-kernel

\[
k_F(x) = \begin{cases} 
1 & |x| \leq 0.5 \\
2 - 2|x| & 0.5 < |x| \leq 1 \\
0 & |x| > 1
\end{cases}
\]

which was first proposed in Politis and Romano (1993). Our simulations reveal that the optimal bandwidth crucially depends on the dimension \( s \) of \( Y_{1,T}, \ldots, Y_{T,T} \). For \( s = 1 \) we get good results with \( b_T = 0.9T^{1/3} \); if \( s > 1 \) the right choice of \( b_T \) gets more complicated. Generally, cusum tests get conservative under positive serial correlation and we notice that this effect gets stronger with increasing \( s \). We propose a very short bandwidth \( b_T = \log_{1.8+40/(T/50)} \) which compensates indirectly by underestimating the serial correlation. A comparable bandwidth is also chosen in Aue et al. (2009) where only a very light serial correlation is considered in the simulations.

We reject the null hypothesis of a stationary time series if \( M_T = \sup_{x \in [0,1]} W_T^2(x) \) is unusually large. We will show in Section 2 that under stationarity \( W_T^2(x) \) converges weakly to a well investigated stochastic process \( (W^2(x))_{x \in [0,1]} \), often called Bessel bridge (Pitman et al., 1999), see Theorem 1 of Section 2 for the necessary assumptions. Suitable asymptotical critical values of \( M_T \) were first tabulated by Kiefer (1959) for \( p \leq 5 \) and are implemented in the R-package `robcp` (Göerz and Dürr, 2019) up to \( p = 5000 \). Crude approximations for large \( p \) are given in Aue et al. (2009). A small selection of critical values can be found in Table 2.

| \( \alpha/p \) | 1     | 2     | 3     | 4     | 5     | 10    | 20    | 50    | 100   |
|--------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 0.9          | 1.500 | 2.114 | 2.623 | 3.083 | 3.514 | 5.450 | 8.885 | 18.172| 32.624|
| 0.95         | 1.844 | 2.508 | 3.053 | 3.543 | 4.000 | 6.041 | 9.626 | 19.219| 34.022|
| 0.975        | 2.191 | 2.894 | 3.469 | 3.984 | 4.464 | 6.595 | 10.310| 20.168| 35.276|
| 0.99         | 2.649 | 3.396 | 4.004 | 4.548 | 5.053 | 7.288 | 11.154| 21.321| 36.783|

Table 2: Quantiles of \( \sup_{0 < x < 1} W_T^2(x) \).

Note that the supremum of \( (W_T^2(x))_{x \in [0,1]} \) is most sensitive to changes in the middle of the time series. There are other functionals like \( \int_0^1 W_T(x)^2 \, dx \) or weighted suprema \( \sup_{0 < x < 1} W_T^2(x)/q(x) \) with \( q(x) \rightarrow 0 \) for \( x \rightarrow 0 \) and \( x \rightarrow 1 \) which are more powerful if changes occur at the beginning or at the end.

We conclude this section by mentioning some similar approaches in the literature. Koul et al. (2003) are among the first who considered M-estimators in the change point context, more precisely for estimating a change point in a regression model with iid errors. In Fiteni (2002) additionally to Koul et al. (2003) short range dependence is considered and the use of a standardization by a scale estimator \( \hat{\sigma} \). Han and Tian (2006) use truncated observations to estimate the time of a location change of a strongly mixing time series with heavy tails. M-estimators are closely connected to our theory. In the one-dimensional location context they are defined as
solution of
\[ \arg \min_{\mu \in \mathbb{R}} \sum_{t=1}^{T} \rho \left( \frac{X_t - \mu}{\hat{\sigma}} \right) \]  
(6)

where \( \hat{\sigma} \) is a scale estimator and \( \rho \) a positive usually symmetric and and often convex function. If the latter is true one can reformulate (6) such that \( \hat{\mu} \) is the unique solution of
\[ \sum_{t=1}^{T} \Psi \left( \frac{X_t - \mu}{\hat{\sigma}} \right) = 0 \]  
(7)

where \( \Psi \) is the derivative of \( \rho \). So if we choose \( \hat{\mu} \) as M-estimator defined by \( \Psi \) in (3), (4) is the square (or quadratic form) of the cusum statistic of M-residuals. Looking at univariate M-residuals for change point testing has already been proposed in Hušková and Picek (2005) and Hušková and Marušiaková (2012b) but without the beneficial scale standardization. Multiple change point detection under iid noise using M residuals (also without standardization) is considered in Fearnhead and Rigaill (2017).

To detect changes in the dependence structure of a multivariate time series Vogel and Fried (2015) basically propose (4) with \( \Psi(x) = s(x)s(x)^T \) but do not consider a standardization by location and scale.

The rest of the article is structured as follows. Section 2 consists of the technical assumptions and the theoretical results, while Section 3 contains some tests based on specific \( \Psi \)-functions and their performance in small simulation studies. All proofs can be found in the appendix.

2. Theoretical results

In this section we give theoretical justification of the asymptotic critical values and compile all conditions which are necessary for the asymptotical results. We start by defining the type of short range dependence we impose on the time series. We assume that it is strongly mixing.

**Assumption 1.** Let \( (X_t)_{t \in \mathbb{N}} \) be strongly mixing with mixing constants \( (a_k)_{k \in \mathbb{N}} \) fulfilling
\[ \sum_{k=1}^{\infty} k^2 a_k < \infty. \]

**Remark 1.** Linear and GARCH processes are strongly mixing if the innovations possess a Lebesgue density (Chanda, 1974; Lindner, 2009). For an overview on different concepts of mixing and its properties see Bradley (2005).

Next we take a closer look at the transformation \( \Psi : \mathbb{R}^p \to \mathbb{R}^s \). In the following \( D_x \in \mathbb{R}^{p \times p} \) denotes the diagonal matrix with the elements of \( x \) on its diagonal. Besides the boundedness of \( \Psi \) we demand that the map is not too "irregular" and do not lead to a degenerate long run covariance matrix:

**Assumption 2.** Let \( \Psi : \mathbb{R}^p \to \mathbb{R}^s \) be a function fulfilling:

a) there exists \( C \in \mathbb{R} \) such that \( |\Psi(x)|_\infty < C \ \forall x \in \mathbb{R}, \)

b) \( \det(U) > 0 \) where
\[ U = \text{Var}(\Psi \{ D^{-1}_\sigma (X_1 - \mu) \}) + \sum_{h=1}^{\infty} \text{Cov}(\Psi \{ D^{-1}_\sigma (X_1 - \mu) \}, \Psi \{ D^{-1}_\sigma (X_{1+h} - \mu) \}) \]

c) Every component of \( \Psi \) is two times continuous differentiable in \( \mathbb{R} \setminus D \) and there exists \( C_1, C_2 > 0 \) with \( |\Psi^{(i)}(x)^T x| \leq C_1 \) and \( x^T \Psi^{(i)T}(x)x \leq C_2 \), \( \forall x \in \mathbb{R}^p \setminus C \) and \( i = 1, \ldots, s \)
d) The exemption set $D$ can have the following properties:

i) It can contain balls $A_r(a_1), \ldots, A_r(a_n)$ with radius $\epsilon > 0$ around finitely many singularities $a_1, \ldots, a_k$ such that there exists $C_3 > 0$ and $\delta > 0$ with \( \sup_{T} \sup_{r \geq |x-a_j| \geq T \rightarrow 2p} \Psi(i)'(x) = C_3 T^{p-\delta} \) and \( \sup_{T} \sup_{r \geq |x-a_j| \geq T \rightarrow 2p} \Psi(i)''(x) \leq C_3 T^{2p-\delta} \) for $j = 1, \ldots, k$ and $i = 1, \ldots, p$.

ii) If $\Psi$ is Lipschitz continuous, it can contain a bounded hypersurface $B$ where $\Psi$ is not differentiable.

iii) If $\Psi$ is Lipschitz continuous and fulfills the following condition: \( \exists K > 0 \) such that for arbitrary but fixed $x_1, x_{k-1}, x_{k+1}, \ldots, x_p$

\[ \Psi(x_1, \ldots, x_{k-1}, a, x_{k+1}, \ldots, x_p) = \Psi(x_1, \ldots, x_{k-1}, b, x_{k+1}, \ldots, x_p), \forall a, b > K, \]

than $D$ can contain unbounded sets $B_1, \ldots, B_l$ where $B_i$ are hyperplanes of the form $B_i = \{(x_1, \ldots, x_{m-1}, a_i, x_{m+1}, \ldots, x_p) : x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p \in \mathbb{R}\}$ where $\Psi$ is not differentiable.

iv) It can contain unbounded sets $E_1, \ldots, E_r$ where $E_i$ are hyperplanes of the form $E_i = \{(x_1, \ldots, x_{m-1}, a_i, x_{m+1}, \ldots, x_p) : x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_p \in \mathbb{R}\}$ where $\Psi$ is discontinuous as long as $\Psi(x) = 0$ for $x \in \mathbb{R}^p \setminus (E_1 \cup \ldots \cup E_r)$.

Remark 2.  
- In Assumption a) we actually need only finite essential supremum and infimum of $\Psi$. One can even drop the boundedness condition completely and demand finite moments and a faster decrease of the mixing constants $(a_k)_{k \in \mathbb{N}}$. However boundedness is a necessary assumption for robustness.

- Assumption b) guarantees that we do not have a degenerate limit process which could be a result of a degenerate $(X_t)_{t \in \mathbb{R}}$ or an improper choice of $\Psi$, $\bar{\mu}$ or $\bar{\sigma}$.

- If one estimates $\mu$ and $\sigma$ one needs a type of continuity of $\Psi$. In the more specific situation of $M$-residuals without scale standardization Husková and Marušiakova (2012) assume Lipchitz continuity of the derivative of $\Psi$ and Lipschitz continuity of some $2 + \delta$ moment. We assume two times differentiability at almost all points but allow for exemption sets which enables the use of discontinuous $\Psi$.

- The multivariate sign function $\Psi_s(x)$ has a singularity in 0. Allowing for the exemption points $a_1, \ldots, a_k$ in d) i) enables the use this function and also the spatial sign cross product $\Psi_{SCov}(\bar{x})$.

- The multivariate Huber function $\Psi_H(\bar{x})$ is not differentiable at $\{x \in \mathbb{R} : |x| = k\}$. This exemption set is allowed because of d) ii) which also enables the use of the Huber cross product $\Psi_{HCov}(\bar{x})$.

- The marginal Huber functions defined by $\Psi(x)^{(i)} = x^{(i)} I_{x^{(i)} \leq k} + k \cdot s(x^{(i)}) I_{x^{(i)} > k}$ for $i = 1, \ldots, p$ are allowed due to d) iii) likewise marginal Huber cross products.

- Marginal signs are allowed due to d) iv).

Furthermore we need that both the location and the scale estimator be $\sqrt{T}$ consistent for their theoretical counterparts.

Assumption 3. Let $(\mu_T)_{T \in \mathbb{N}}$ and $(\sigma_T)_{T \in \mathbb{N}}$ be two stochastic sequences fulfilling

$\mu_T - \mu = O(T^{-\frac{1}{2}})$ and $\sigma_T - \sigma = O(T^{-\frac{1}{2}})$. 

5
Remark 3. Assumption 3 is rather weak. We actually do not need to know the population values $\mu$ and $\sigma$ and in practice we rarely will. This would require knowledge of the marginal distribution of $X_1$. Consistency of quantile based estimators like median and MAD follows from the existence of a Bahadur representation which was shown under strong mixing and continuous density of the innovations in [Yoshihara (1995)].

Finally we need assumptions for the kernel estimator. They are nearly identical with these in [Jong and Davidson (2000)]:

Assumption 4. Let $k : \mathbb{R} \to [-1,1]$ be a function which is continuous in 0 and has only finitely many discontinuities. Furthermore it fulfills $k(0) = 1$, $k(x) = k(-x)$ for $x \in \mathbb{R}$ and $\int_{-\infty}^{\infty} |k(x)| \, dx < \infty$ as well as $\int_{-\infty}^{\infty} | \int_{-\infty}^{\infty} k(x) e^{-2\pi x \xi} \, dx | \, d\xi < \infty$. For the bandwidth $b_T$ it holds that: $b_T \to \infty$ and $b_T/T^{1-\epsilon} \to 0$ for $T \to \infty$ and some $\epsilon > 0$.

The next theorem contains the main result of this paper and describes the asymptotic distribution of $(W_T(x)^2)_{x \in [0,1]}$ under the null hypothesis.

Theorem 1. Let Assumptions 1-3 hold then

$$(W_T(x)^2)_{x \in [0,1]} \xrightarrow{w} \left( \sum_{i=1}^{p} BB_i(x)^2 \right)_{x \in [0,1]}$$

where $(BB_i(x))_{x \in [0,1]}$, $i = 1, \ldots, p$ are mutually independent standard Brownian Bridges.

The result concerns weak convergence in the Skorohod space $D[0,1]$ which consists of all functions which are right continuous with left-hand limits. The continuous mapping theorem therefore yields the validity of the proposed asymptotic critical values. The linear structure of the test statistic makes it feasible to derive also asymptotics under the alternative. Nevertheless there are two challenges. The first one is the estimation of the location and scale standardization and the second is the estimation of the long run covariance. Especially if $s > 1$ this gets tricky since we need to ensure positive definiteness of $V$ which depends on the direction of the alternative. If one looks at local alternatives this poses less problems, nevertheless we postpone the theory under alternatives to future work and investigate the properties of our approach by simulations.

3. Application

Theorem 1 enables various kinds of change point tests. As mentioned in the introduction one can detect amongst others changes in location scale and dependence. If one does not assume a specific model one will usually use truncated moments like in Table 1. Otherwise truncated scores of the log-likelihood can be used. For example, if $X_1, \ldots, X_T$ is assumed to be a sequence of independent exponentially distributed random variables under the null hypothesis, one can use

$$\Psi(X_i) = \begin{cases} \frac{\ln(2)X_i}{\hat{\sigma}}, & \text{if } \frac{\ln(2)X_i}{\hat{\sigma}} \leq k \\ \frac{\ln(2)X_i}{\hat{\sigma}}, & \text{if } \frac{\ln(2)X_i}{\hat{\sigma}} > k \end{cases}$$

where $\hat{\sigma} = \text{Median}(X_1, \ldots, X_T)$. Note that we do not center the observations here and drop the term $-\log(2)/\hat{\sigma}$ since it does not change the cusum statistic. Even if the model is incorrectly specified, the observations are serially correlated or follow a different distribution the test is still valid with respect to the asymptotic size under the null hypothesis of no change but will lose power under the alternative.
In the following we present some useful non-parametric tests. If one tries to detect a change in location and \( p = 1 \) we recommend the Huber-\( \Psi \)-function
\[
\Psi_H(x) = \begin{cases} 
  x, & \text{if } |x| \leq k \\
  ks(x), & \text{if } |x| > k 
\end{cases}
\]
originally proposed for location estimation in \cite{Huber1964}. Different authors propose different choices of \( k \) in the estimation context. In \cite{Huber2011} \( k \in [1, 2] \) is recommended and \( k = 1.5 \) suggested, but there are also other proposals, for example \( k = 1.2 \) \cite{CantoniRonchetti2001}, \( k = 1.25 \) \cite{Street1988} and \( k = 1.25 \) \cite{Wang2007} propose a data dependent \( k \). In general a larger value of \( k \) is more efficient under normality but less efficient under heavy tails and outliers.

We investigate the impact of \( k \) in the related problem in detecting a scale or scatter shift. There are two straightforward generalizations for \( p > 1 \). One can use \( p \) univariate Huber-M-\( \Psi \)-functions
\[
\tilde{\Psi}_H(x) = (\Psi_H(x)^{(1)}, \ldots, \Psi_H^{(p)}(x))'
\]
and a multivariate one
\[
\Psi_H(x) = \begin{cases} 
  x, & \text{if } |x| \leq k \\
  s(x)k, & \text{if } |x| > k 
\end{cases}
\]
One expects that the later version is more powerful but less robust in case of elliptical marginal distributions as this applies for the estimation problem \cite{MaronnaYohai1976}. If \( p \) is large compared to \( T \), tests based on projections might be preferable though the power of the test crucially depends on the direction \( a \in \mathbb{R}^p \) chosen for the projection. The choice
\[
\Psi_P(x) = (\Psi_H(x)^{(1)}, \ldots, \Psi_H^{(p)}(x))a
\]
gives a robustified projection based test.

The sample variance and covariance are due to their quadratic nature even more influenced by outliers than the arithmetic mean. For symmetric distributions
\[
\Psi_{HVar}(x) = \begin{cases} 
  x^2, & \text{if } |x| \leq k \\
  k^2, & \text{if } |x| > k 
\end{cases}
\]
is an intuitive choice to find a change in the scale of time series with rather symmetric marginal distributions, but there are many other possibilities. One can of course also look at Huberized absolute values or any power of it. If the marginal distribution is very heavy skewed the test is still valid but not optimal. If the distribution is for example positively skewed the \( \Psi \)-function will downweight many valuable observations on the right tail and on the other hand overlook outliers on the left tail. If the time series is additionally positive \( \Psi_{LogHVar}(x) = \Psi_{HVar}(\log(x)) \) might be a suitable choice, since it can symmetrize the marginal distribution.

Like in the location case there are at least two intuitive generalizations for \( p > 1 \). The first one Huberizes observations in every component individually
\[
\tilde{\Psi}^{i,j}_{HCov}(x) = \Psi_{HVar}(x^{(i)})\Psi_{HVar}(x^{(j)}),
\]
while the second one Huberizes all components at once
\[
\Psi_{HCov}(x) = \begin{cases} 
  xx^T, & \text{if } |x| \leq k \\
  k^2 s(x)s(x)^T, & \text{if } |x| > k 
\end{cases}
\]
Depending on the outlier model one proposal is probably more favourable than the other. The first one is related to the scatter estimator proposed in [Van Aelst (2016)], which is constructed to cope with cellwise outliers. In this setting the elements of an observation vector are corrupted individually. The second approach is similar to scatter estimators recently proposed in [Raymaekers and Rousseeuw (2018)] which show good results in case of rowwise outliers, where whole observations are corrupted. If an observation $X_t$ is corrupted only in a few components one should apply $\tilde{\Psi}_{HCov}$, while if one expects only heavy tailed observations or rowwise no outliers then $\Psi_{HCov}(x)$ should be used.

For transformations measuring dependence we need additional operators, since assumption 2b) is violated by default otherwise. If one simply vectorizes a symmetric matrix one gets duplicated elements and therefore perfectly linear dependence which lead to a singular $V$. We define $DU: \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{(p+1)/2}$ with $DU(M) = (M_{1,1}, M_{2,1}, \ldots, M_{p,1}, M_{2,2}, \ldots, M_{p,p}, \ldots, M_{p,p})^T$ which extracts the diagonal and lower diagonal elements of a matrix. Then also the Huberized covariance transformation $DU(\Psi(x))$ with $\Psi_{HCov}(x)$ fulfills Assumption 2. For the spatial sign covariance matrix we need $dU: \mathbb{R}^{p \times p} \rightarrow \mathbb{R}^{p \times (p+1)-1}$ with $dU(M) = (M_{1,1}, M_{2,1}, \ldots, M_{p,1}, M_{2,2}, \ldots, M_{p,p-1})^T$ which omits the last diagonal element out as opposed to $DU$.

4. Simulations

We want to evaluate advantages and disadvantages of the proposed approach in some simulations. In the following we concentrate on the case of a change in the variance ($p = 1$) respectively covariance ($p > 2$) of a possibly multivariate time series. If one is interested in a change in location we refer to [Dehling et al. (2015b)] which includes extensive simulations. Our approach is called Huberization test there and is quite competitive, though it is beaten by the two sample Hodges-Lehmann test proposed there. However, there are two advantages which might counterbalance a slight disadvantage with respect to power. First, our procedure has a computational complexity of $T \log(T)$ whereas for the current implementation of the Hodges-Lehmann based test it is of $T^3$ prohibiting its application to very large samples and second, our approach based on bounded transformations has a natural extension to the case $p > 1$. Under elliptical distributions it is known that estimators and tests based on the multivariate sign function get more efficient with increasing dimension [Paindaveine et al. (2016)] and this is also observed in the change point context [Vogel and Fried (2015)].

4.1. Change in marginal scale of a one dimensional time series

We first want to take a closer look at the case $p = 1$ and a change in the scale of a one dimensional time series. There are some tests in the literature for this situation which are robust to some degree or appropriate under heavy tails, see [Gerstenberger et al. (2016)]. All three of them are cusum type tests based on different estimators of scale, using

$$\max_{k=1, \ldots, n} \frac{k}{\sqrt{n}} |s_{1:k} - s_{1:n}|$$

where $s_{1:k}$ is a scale estimator based on $X_1, \ldots, X_k$ and $\hat{v}$ is an estimator of the asymptotic variance of $s_{1:n}$. In [Gerstenberger et al. (2016)] three estimators are investigated. The mean absolute deviation (abbreviated as MD) is defined as

$$s_{1:n} = \frac{1}{n} \sum_{i=1}^{n} |x_i - q_{0.5}(X)|$$

where $q_{\alpha}(X)$ denotes the $\alpha-$Quantile of $X_1, \ldots, X_n$. The MD has an asymptotic efficiency of 0.88 compared to the standard deviation under independent and identically normal distributed
data (Fisher, 1921), but its breakdown point is 0. Ginis mean difference (GMD) is a $U -$statistic of the following form

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

It has an efficiency of 0.98 (Gerstenberger and Vogel, 2015) and also a breakdown point of 0. The quantile of pairwise differences ($Q_{n}^{(\alpha)}$) is a U-Quantile defined as

$$ Q_{n}^{(\alpha)} = q_{\alpha}(|X_i - X_j| : 1 \leq i < j \leq n). $$

To get a robust estimator, Rousseeuw and Croux (1993) propose $\alpha = (\lfloor n/2 \rfloor + 1) \approx 0.25$ which yields an optimal breakdown point of 0.5 and an efficiency of 0.82. However, Gerstenberger et al. (2016) explore that $\alpha = 0.8$ is more appropriate in the change point setting. This yields an efficiency of 0.96 and a breakdown point of 0.08.

Our test is based on the psi-function $\Psi_{H_{\text{Cov}}}$. It remains to choose a suitable value of $k$. A larger value transforms less values and leads to more powerful tests under Gaussianity whereas a smaller $k$ is more efficient under heavy tails. In the following we will choose $k$ as a quantile of the $\chi^2_1$ distribution (with one degree of freedom). So $k = q_{\chi^2_1}(0.5) \approx 0.45$ means that under independent normally distributed data approximately 50% of the observations is transformed. We choose the 0.5, 0.8 and 0.95 quantile and denote the resulting estimators as $M_{0.5}$, $M_{0.8}$ respectively $M_{0.95}$.

In our simulations we concentrate on ARCH models of order one:

$$ X_t = \sigma_t \varepsilon_t, \quad t = 1, \ldots, T $$

(8)

where $(\varepsilon_t)_{t \in \mathbb{Z}}$ is a series of iid Gaussian random variables and $\sigma_t^2 = \pi_0 + \pi_1 X_{t-1}^2$. This comparatively simple model allows us to investigate the effect of heavy tails as well as serial correlation. A large value of $\pi_1$ yields a strong dependence and heavy tails, while $\pi_1 = 0$ produces an iid Gaussian time series.

There are some specific tests to detect changes in ARCH models. We included some of these in our simulations. The first proposal is by Kokoszka et al. (2002) who propose a test on estimated GARCH residuals $\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_T$. We abbreviate it as GARCH-res-edcF. Denote $\hat{F}_{1:k}$ the empirical distribution function of $\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_k^2$ and by $\hat{F}_{k+1:n}$ the one based on $\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2$. Then the test statistic is computed as

$$ B_T = \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} \left( \frac{\sqrt{T}}{T} \left[ 1 - \frac{i}{T} \right] |\hat{F}_{1:i}[\hat{\varepsilon}_j] - \hat{F}_{i+1:T}[\hat{\varepsilon}_j]| \right)^2. $$

Asymptotical critical values of $B_T$ can be found in Kokoszka et al. (2002) which date back to Blum et al. (1961).

Also residual based is the approach of Kulperger et al. (2005) which uses

$$ D_T = \max_{t=1, \ldots, T} \frac{\sum_{i=1}^{t} (\hat{\varepsilon}_i - \bar{\varepsilon})^2 - i \hat{\sigma}^2}{\sqrt{i \hat{\nu}}}, $$

where

$$ \bar{\varepsilon} = \frac{1}{n} \sum_{t=1}^{T} \hat{\varepsilon}_t, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (\hat{\varepsilon}_t - \bar{\varepsilon})^2 \quad \text{and} \quad \hat{\nu}^2 = \frac{1}{n} \sum_{t=1}^{T} (\hat{\varepsilon}_t - \bar{\varepsilon})^2 - \bar{\varepsilon}^2. $$

The cusum-type statistic $D_T$ is asymptotically distributed like the maximum of the absolute value of a Brownian bridge. We abbreviate this test by GARCH-res.
There are two more tests based on the likelihood function. Both require the parameter $\pi_1$ to be strictly larger than 0. The first one is by Berkes et al. (2004) who propose using functionals of
\[
E(t) = \frac{1}{T} \left[ \sum_{i=1}^{\left\lfloor nt \right\rfloor} \hat{l}_i' (\hat{\pi}) \right] D^{-1} \left[ \sum_{i=1}^{\left\lfloor nt \right\rfloor} \hat{l}_i (\hat{\pi}) \right]^T.
\]  
(9)

Thereby $l'_i$ denotes the estimated score. If one assumes normally distributed errors $\epsilon_t$, $t = 1, \ldots, T$ for the stated model it equals
\[
l'_i (\hat{\pi}) = \left( \frac{X_i^2}{2(\hat{\pi}_0 + \hat{\pi}_1 X_{i-1})^2} - \frac{1}{2(\hat{\pi}_0 + \hat{\pi}_1 X_{i-1})} \cdot \frac{X_i^2 X_{i-1}^2}{2(\hat{\pi}_0 + \hat{\pi}_1 X_{i-1})^2} - \frac{X_i^2}{2(\hat{\pi}_0 + \hat{\pi}_1 X_{i-1})} \right).
\]

and $D$ can be estimated by $\frac{1}{T} \sum_{t=1}^{T} \hat{l}_i (\hat{\pi})^T \hat{l}_i (\hat{\pi})$. It is shown that $(E(t))_{t \in [0,1]}$ converges to a squared Bessel bridge with parameter 2. To simplify the computational complexity we use the maximum of $E_t$. We call the Test GARCH-LM3 in the following.

A Lagrange multiplier test which enables detecting changes of single parameters is proposed by Galeano and Tsay (2009) and abbreviated as GARCH-LM1. The test statistic reads

\[LM_T = \max_{t_{\min} \leq t \leq t_{\max}} T \frac{T}{(1-t)} \frac{[\hat{l}_i B_T(\hat{\pi}) A_T(\hat{\pi})]_{[2]}^2}{A_T(\hat{\pi})_{[2]} B_T(\hat{\pi})^{-1} A_T(\hat{\pi})_{[2]}^T}.\]

where $\hat{l}_i = \frac{1}{T} \sum_{t=1}^{T} l_i (\hat{\theta})$. In our model the scores $l_i (\hat{\theta})$ have the following form

\[l_i (\hat{\theta}) = \frac{X_i - \hat{\mu}}{\hat{\pi}_0 + \hat{\pi}_1 (X_{i-1} - \hat{\mu})} (1,0,0)^T + \frac{1}{2} \left( \frac{(X_i - \hat{\mu})^2 - \hat{\pi}_0 - \hat{\pi}_1 (X_{i-1} - \hat{\mu})^2}{\hat{\pi}_0 + \hat{\pi}_1 (X_{i-1} - \hat{\mu})^2} \right) h'_i (\hat{\theta})^T\]

with $h'_i (\hat{\theta}) = [ -2(X_i - \hat{\mu}), 1, (X_{i-1} - \hat{\mu})^2 ]$.

Furthermore we have

\[A_T(\hat{\theta})_{[2]} = -\frac{1}{T} \sum_{t=1}^{T} \left[ \frac{1}{2[\hat{\pi}_0 + \hat{\pi}_1 (X_{i-1} - \hat{\mu})^2] h'_i (\hat{\theta})^T h'_i (\hat{\theta})} \right]_{[2]} \quad \text{and} \quad B_T = \frac{1}{T} \sum_{t=1}^{T} l_i (\hat{\theta}) l_i (\hat{\theta})^T\]

with $(\cdot)_{[i]}$ denoting the $i$-th column of a matrix. By choosing the second column of $A_T$ we test against a change in $\pi_0$ which reflects a change in the marginal variation. In De Long (1981) the following asymptotic formula to compute $p$-values of $LM_T$ is derived

\[P(LM_T > x) \approx \frac{1}{\Gamma\left(\frac{1}{2}\right)} \sqrt{2xe^{-\frac{x}{2}}} \left[ \log \left( \frac{t_{\max}}{t_{\min}} \right) \left( \frac{1}{x} \right)^{\frac{1}{2}} + \frac{2}{x} \right].\]

The values $t_{\min} = 0.05$ and $t_{\max} = 0.95$ are chosen as boundary values.

For all four tests we use the R-package fGarch (Wuertz et al., 2017) to estimate the parameters of the ARCH(1) model.

We have not addressed the specific choice of the kernel and the bandwidth of the long run variance estimation yet. Assumption is very general. From the theory of spectrum estimation we know that a large choices of $b_n$ yield a large variance whereas small ones can produce a large bias. There are optimal rates depending on the time series model and plug-in estimators if the model is unknown. Note that these data dependent estimators are constructed under the null hypothesis. Under the alternative they overestimate the dependence and yield very large values of $b_n$ which seriously affect the power (Vogelsang, 1999). We tried a large number
of different fixed bandwidths and found that the ones in Table 3 to be especially useful for ARCH(1) models. Note that a small bandwidth is always preferable under the alternative, since a level shift results in large estimated autocorrelations and therefore a large estimated long run variance, which enters the denominator of the test statistic. In this regard MD and GMD have a slight advantage over $M_k$ and $Q_{n(0.8)}$. The latter requires an additional tuning parameter for a necessary density estimation. We used the same as Gerstenberger et al. (2016) but noticed that this nuisance parameter is not very influential on the power and the size of the test. Finally we use the finite sample correction proposed in Dürr (2018), leading to

$$\sup_{x \in [0,1]} |W_T(x)| + \frac{\zeta(0.5)}{\sqrt{2\pi \cdot T}}. \quad (10)$$

The summand $\zeta(0.5)/\sqrt{2\pi \cdot T} \approx 0.58/\sqrt{T}$ originally results from the asymptotic expectation of

$$\sqrt{T} \left( \sup_{x \in [0,1]} B_x - \max_{x=0,1/T, \ldots, 1} B_x \right),$$

see Asmussen et al. (1995). But it turns out that this finite sample correction is also valid for the maximum of the absolute values of a Brownian bridge (Dürr, 2018). Simulations indicate that this correction is also useful for robustly transformed tests, see Figure 1. It turns out that the correction is also useful for MD, GMD and $Q_{n(0.8)}$, which is not surprising since the linearization of all these tests is the ordinary cusum test. Therefore we added the correction also to these tests.

First we want to assess the size under the null hypothesis. The first row of Figure 2 reveals that almost all tests hold their size irrespective of $\pi_1$ (the larger $\pi_1$ the larger the serial correlation) and the time series length $T$. For small sample sizes we observe that the non parametric tests (blue and black curves) get a little conservative for moderately large $\pi_1$. This behaviour vanishes as $\pi_1$ increases further. This is typical for cusum type tests, since these tests get more conservative with increasing serial dependence (Dürr, 2018). On the other hand under increasing serial dependence the estimation of the long run variance gets more negatively biased, which first cancels out the conservativeness and then dominates it.

We notice that $Q_{n(0.8)}$ exceeds its size if both $\pi_1$ and $T$ are large. This points to a too small choice of the bandwidth. Since $Q_{n(0.8)}$ has already the largest bandwidth and Gerstenberger et al. (2016) also propose usage $T^{-1/3}$ we decide against enlarging $b_T$ which would considerably harm power under the alternative. It is also noteworthy that $M_\infty$ is very conservative for large $\pi_1$. This is not surprising, since the assumption of finite second moments, which are in fact fourth moments testing the variance testing, is violated. Furthermore we notice that the likelihood based tests have problems if $\pi_1$ is close to 0, which is not surprising since $\pi_1 > 0$ is an assumption for the asymptotics. GARCH-res is rather conservative under any value of $\pi_1$, especially if $T$ is small.

### Table 3: Chosen bandwidths $b_T$ and kernels $k$ for the estimation of the long run variance.

| Estimator   | $b_T$           | Kernel $k(x)$                      |
|-------------|-----------------|-----------------------------------|
| $M_q$       | $0.9T^{1/3}$    | $k(x) = \begin{cases} 1 & x \leq 0.5 \\ 2 - 2x & 0.5 < x < 1 \end{cases}$ |
| MD          | $T^{1/4}$       |                                    |
| GMD         | $T^{1/4}$       |                                    |
| $Q_{n(0.8)}$| $T^{1/3}$       |                                    |
Figure 1: Simulated empirical sizes of $M_q$ tests under $X_T = \epsilon_t$, $t = 1, \ldots, T$. where $\epsilon_t$ are independent and identically distributed following a standard normal distribution (top-left), a Cauchy distribution (top-right), a uniform distribution on $[0, 1]$ (bottom-left) and an exponential distribution with parameter $\lambda = 1$ (bottom right).
Figure 2: Simulation results under ARCH(1) model $X_t = \epsilon_t \sigma_t$ with $\sigma_t = \pi_0 + \pi_1 X_{t-1}, \ t = 1, \ldots, T$ with $T = 100$ (left) and $T = 800$ (right). The first row shows the empirical size under $H_0$ with respect to different $\pi_1$. The remaining rows show the power with respect to different variables, namely the size of the change (2nd row), the time of the change-point (3rd row) and $\pi_1$ (4th row).
In the following we evaluate the power under the alternative. We use the model:

\[ Y_t = \begin{cases} 
X_t, & t \leq bT \\
X_t \cdot \delta, & T > bT 
\end{cases} \]

where \((X_t)_{t \in 1, \ldots, T}\) follows the ARCH(1) model defined in \(8\). Generally we want to compensate for obvious effects and therefore relate the jump height \(\delta\) to the sample size \(T\), the fraction of the data before the change \(b\) and the ARCH-Parameter \(\pi_1\). One expects asymptotically stable power of the tests for jump heights which are proportional to \(1/\sqrt{T}\). Furthermore the power should decrease if the value \(b\) departs from 0.5. It turns out that a jump height proportional to \(1/b/(1-b)\) stabilize power. Finally the power should decrease with increasing \(\pi_1\) due to heavier tails and more serial correlation. The accurate stabilizing function is not known and differs for the different estimators. We account for this with the factor \(1/(1-\pi_1)\) resulting in

\[ \delta = 1 + \frac{s}{\sqrt{T} \cdot b \cdot (1-b) \cdot (1-\pi_1)}. \]  

(11)

Using this more complicated jump heights allows us to see more clearly the differences between the estimators and characteristics of their finite sample behaviour.

First we want to investigate the effect of increasing \(s\). We set \(\pi_1 = 0\), \(b = 0.5\) and \(T = 100\) and visualize the power under varying jump height \(s\) on the left hand side of the second row of Figure 2. We see that the power under a negative jump of size \(s\) is larger than the power under a positive \(s\). This might be due to the fact that the variance respectively scale lives naturally on a log scale. However if \(T\) is very large \(\delta\) differs only marginally from 1 where the derivative of the logarithm is nearly 1, so we do not see this effect for \(T = 800\) on the right hand side of the second row of Figure 2. More surprisingly is the fact that the order of the tests depends on the sign of the jump. For a negative jump MD has the largest power, followed by \(M_{90.95}\) and GMD. The ordering changes to GMD, \(M_{90.95}\) and MD for a positive jump. The variance based test \(M_\infty\) has problems for small \(T\) due to its conservativeness under the null hypotheses but becomes the most powerful test under \(T = 800\). Noteworthy is also the non-monotonic behaviour of \(Q_n^{(0.8)}\) for negative jumps and small \(T\). The specific GARCH-tests are generally not so powerful as the non-parametric ones, which is due to the choice of \(\pi_1 = 0\). We see furthermore that the parametric tests gain more power with increasing \(T\).

In the third row of Figure 2 we see the effect of the fraction of change \(b\). We fix \(\pi_1 = 0.5\) and \(s = 1\) and vary \(b\) between 0.05 and 0.95 in steps of 0.05. Note that one expects the maximal power at \(b = 0.5\). Since we already account for this in the jump height \(11\), one rather expects a straight line instead. Especially for \(T = 100\) we see large deviations from that. GMD, MD, \(M_\infty\) and GARCH-LM1 are more powerful if the change appears at the end of the time series than in the beginning. The other tests show the opposite behaviour. The most extreme here is the GARCH-res-ecdf followed by \(M_{90.5}\) and \(Q_n^{(0.8)}\). The larger \(k\) the less pronounced is the difference in power between early and late changes. For \(T = 800\) we see a larger plateau where the power is indeed constant, but the asymmetry of the tests is still visible. Specific GARCH tests generally outperform their competitors now because of the choice \(\pi_1 = 0.5\).

In the last row of Figure 2 we investigate the tests under increasing serial correlation and heavy tails by varying \(\pi_1\) between 0 and 0.95 in steps of 0.05. We set \(s = 1\) and \(b = 0.5\). Results for \(T = 100\) are shown on the left hand side and for \(T = 800\) on the right. Since the jump is positive GMD is most powerful for small \(\pi_1\). As before \(M_\infty\) is handicapped by its conservativeness under the null and therefore for Gaussian time series only as powerful as GMD for \(T = 800\). We see furthermore that for \(\pi_1 > 0.55\) for \(T = 100\) respectively \(\pi_1 > 0.35\) for \(T = 800\) \(M_{90.95}\) becomes the most powerful non-parametric test. For \(T\) and \(\pi_1\) large it gets beaten by \(M_{90.95}\) though. The parametric tests outperform the non-parametric ones if \(\pi_1 > 0.4\). The difference gets more...
pronounced if \( T \) is large, as we have seen before. Generally the residual based test GARCH-res seems to be the best choice overall.

In summary we have seen that our approach is advantageous under heavy tails compared to other non-parametric methods. The tuning parameter \( k \) should be chosen according to the degree of heavy tailedness. Without a-priori information we recommend using \( k = q_{\chi^2}(0.95) \) since it delivers good results under normality and various degrees of heavy tailedness. If one can assume Gaussianity we prefer MD over GMD because of its computational simplicity. If one expects a GARCH process with severe serial dependence one should use GARCH-res.

4.2. Change in cross sectional dependence of a multivariate time series

Now we want to see how results generalize for \( p > 1 \). To the best of our knowledge there is no other robust test for a change in the cross covariance \( \Sigma = \text{Cov}(X_i) \). But there is a non-robust one by Aue et al. (2009) based on the empirical covariance (abbreviated as Cov)

\[
\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})(X_i - \hat{\mu})^T.
\]

where \( \hat{\mu} \) is the arithmetic mean. Except for using the mean instead of the median this test is a special case of our Huberized covariance test with \( k = \infty \). There is some kind of robustified version which uses roots of absolute values of the original observations:

\[
Y_{i}^{(1)}, \ldots, Y_{i}^{(p)} = (|X_{i1}|^{\delta}, \ldots, |X_{ip}|^{\delta}) \quad \text{with} \quad \delta \in (0, 1]
\]

but in our simulations with random vectors which are not positive this approach does not lead to decent results. This is not surprising since taking absolute values of the components completely destroys the dependence structure.

There are other approaches which concentrate on a change in the dependence structure. In Wied (2016) a test based on the empirical correlation (abbreviated as Cor) is proposed

\[
\hat{\rho}_{k,l} = \frac{\sum_{i=1}^{T} (X_{ik} - \overline{X(k)})(X_{il} - \overline{X(l)})}{\sqrt{\sum_{i=1}^{n} (X_{ik} - \overline{X(k)})^2 \sum_{i=1}^{n} (X_{il} - \overline{X(l)})^2}}
\]

The serial correlation is accounted by a block bootstrap. The sample correlation is known to be efficient under normality but not robust.

A more robust approach is presented in Bücher et al. (2014), who propose a test based on the empirical copula

\[
\hat{C}(u) = \frac{1}{n} \sum_{i=1}^{n} I(U_i \leq u)
\]

where \( U_i = \frac{1}{n}(R_{i1}, \ldots, R_{ip}) \) and \( R_{ik} = \sum_{j=1}^{n} I(X_{ij} \leq X_{ik}) \) (abbreviated as Copula). The authors propose two different versions of multiplier bootstraps to calculate \( p \)-values. We decided to use the computationally faster one, though the other is known to lead to more powerful tests under small and moderate sample sizes. Note that already the faster version is around 50 times slower than our test whereas the other version is about 3000 times slower. We use the implementation in the npcp-package (Kojadinovic, 2015). A test based on the empirical copula possesses power under a broader range of alternatives than the other tests. Covariance, correlation and also huberized covariance test are constructed to have power against changes in the linear dependence, whereas a copula test can detect any kind of change. We therefore expect it to be less powerful than the others if there is indeed a change in the linear dependence.

Kojadinovic et al. (2016) propose to use multivariate generalizations of Spearman’s \( \rho \). Denote \( \hat{\rho}^{ij} \)
| Estimator | \( b_T \) |
|-----------|-----------|
| Cor       | \( b_T = \log_2(T) \) |
| Cov       | \( b_T = \log_{1.8+\left(p(p+1)/40\right)}(T/50) \) |
| Huberg    | \( b_T = \log_{1.8+\left(p(p+1)/40\right)}(T/50) \) |
| Huberm    | \( b_T = \log_{1.8+\left(p(p+1)/40\right)}(T/50) \) |
| Spearman  | \( b_T = T^{1/3} \) |
| Kendall   | \( b_T = T^{1/3} \) |
| Copula    | \( b_T = \log_p(n/12.5) \) |

Table 4: Chosen bandwidths \( b_T \) for different multivariate change-point tests.

Spearmans \( \rho \) of \((X^{(i)}_1, \ldots, X^{(i)}_T)\) and \((X^{(j)}_1, \ldots, X^{(j)}_T)\), then the test (abbreviated as Spearman) is based on

\[
\left(\frac{2}{p}\right)^{-1} \sum_{1 \leq i < j \leq p} \hat{\rho}^{i,j}.
\]

There are two other proposals, but this one seems to have the largest power [Kojadinovic et al., 2016]. There are (at least) two possibilities to estimate the long run variance: a multiplier bootstrap and a kernel estimation. The authors favour the former since the latter has problems to hold it size under strong serial dependence. Nevertheless we choose the kernel estimator since it is considerably faster and we do not look at strong dependences.

Two tests based on generalizations of Kendalls \( \tau \) are proposed in Quessy et al. (2013). We look at the one which is implemented in the Kojadinovic (2015)-Package (abbreviated as Kendall) and based on

\[
\frac{2}{T(T-1)} \sum_{1 \leq i < j \leq T} \left( I\{X_i < X_j\} + I\{X_i > X_j\} \right).
\]

Two different estimators for the long run variance are possible, a multiplier bootstrap (Bücher and Kojadinovic, 2016) and a kernel estimator. We choose the later because of the lower computation times. Note that Spearman and Kendall are type of projection tests. One expects that they have large power if changes are into the same direction (the dependence gets stronger or weaker overall) but low power if changes are into different directions (the dependence between some variables gets stronger but between others gets weaker).

We sample from a multivariate AR(1) model

\[
X_t = \rho X_{t-1} + \epsilon_t, \quad t = 1, \ldots, T
\]  

(12)

where \((\epsilon_t)_{t \in \mathbb{Z}}\) is a series of independent and multivariate t-distributed random vectors with mean \( \mu = 0 \), shape \( V \) and \( k \) degrees of freedom. The smaller \( k \) the more heavy tailed is \((X_t)_{t \in \mathbb{N}}\). The larger \( k \) the more similar is \((X_t)_{t \in \mathbb{N}}\) to a Gaussian distribution. Consequently if we write \( k = \infty \) we sample \((\epsilon_t)_{t \in \mathbb{Z}}\) from a multivariate normal distribution.

First we want to verify if the tests hold their size. We notice that the multivariate tests get even more conservative under serial dependence if the dimension \( p \) increases. The reason behind this is not known yet and content of future research. We try to counterbalance this effect by choosing very short bandwidths \( b_T \) which can be found in Table 4. Note that we choose the somehow arbitrary bandwidths based on plenty of simulations. Since there is no theoretical justification we can only guarantee that these choices are reasonable in our framework of AR(1) processes with moderate \( p \) and \( \rho \). Up to our knowledge there are now profound simulation studies.
Table 5: Empirical size under multivariate AR(1) model: \( X_t = \rho X_{t-1} + \epsilon_t, \ t = 1, \ldots, T \) with \( \rho \in \{0, 0.5\} \), \( T \in \{200, 400, 800\} \), \( \epsilon \sim N(0, I_p) \), \( p \in \{2, 5\} \) and a nominal size of 0.05.

| \( \rho \) | \( p \) | 2  | 0  | 5  | 2  | 0  | 5  | 2  | 0  | 5  |
|---|---|---|---|---|---|---|---|---|---|---|
| \( T \) | 200 | 400 | 800 | 200 | 400 | 800 | 200 | 400 | 800 | 200 | 400 | 800 |
| Cor | 0.04 | 0.06 | 0.06 | 0.00 | 0.01 | 0.02 | 0.07 | 0.06 | 0.08 | 0.02 | 0.02 | 0.05 |
| Cov | 0.02 | 0.03 | 0.04 | 0.20 | 0.03 | 0.02 | 0.01 | 0.03 | 0.04 | 0.02 | 0.01 | 0.02 |
| Huberg0 | 0.04 | 0.04 | 0.06 | 0.03 | 0.04 | 0.04 | 0.06 | 0.07 | 0.06 | 0.01 | 0.05 | 0.06 |
| Huberg0.5 | 0.04 | 0.05 | 0.06 | 0.03 | 0.04 | 0.05 | 0.05 | 0.07 | 0.07 | 0.01 | 0.05 | 0.06 |
| Huberg0.8 | 0.04 | 0.06 | 0.06 | 0.03 | 0.04 | 0.04 | 0.05 | 0.07 | 0.07 | 0.01 | 0.05 | 0.07 |
| Huberm0 | 0.05 | 0.05 | 0.06 | 0.03 | 0.04 | 0.04 | 0.05 | 0.05 | 0.06 | 0.03 | 0.04 | 0.06 |
| Huberm0.5 | 0.05 | 0.05 | 0.05 | 0.04 | 0.04 | 0.03 | 0.04 | 0.06 | 0.06 | 0.02 | 0.03 | 0.06 |
| Huberm0.8 | 0.05 | 0.06 | 0.06 | 0.03 | 0.04 | 0.04 | 0.06 | 0.07 | 0.07 | 0.01 | 0.04 | 0.06 |
| Copula | 0.00 | 0.00 | 0.23 | 0.00 | 0.08 | 0.30 | 0.00 | 0.00 | 0.22 | 0.17 | 0.06 | 0.23 |
| Spearman | 0.03 | 0.05 | 0.04 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 | 0.06 | 0.03 | 0.05 | 0.04 |
| Kendall | 0.04 | 0.05 | 0.05 | 0.05 | 0.04 | 0.05 | 0.05 | 0.05 | 0.06 | 0.07 | 0.06 | 0.05 |

Table 6: Empirical size under multivariate AR(1) model: \( X_t = \rho X_{t-1} + \epsilon_t, \ t = 1, \ldots, T \) with \( \rho \in \{0, 0.5\} \), \( T \in \{200, 400, 800\} \), \( \epsilon \sim t_3(0, I_p) \), \( p \in \{2, 5\} \) and a theoretical size of 0.05.

| \( \rho \) | \( p \) | 2  | 0  | 5  | 2  | 0  | 5  | 2  | 0  | 5  | 2  | 0  | 5  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| \( T \) | 200 | 400 | 800 | 200 | 400 | 800 | 200 | 400 | 800 | 200 | 400 | 800 |
| Cor | 0.07 | 0.04 | 0.04 | 0.12 | 0.09 | 0.07 | 0.08 | 0.07 | 0.07 | 0.19 | 0.14 | 0.13 |
| Cov | 0.01 | 0.01 | 0.01 | 0.21 | 0.04 | 0.01 | 0.01 | 0.01 | 0.03 | 0.09 | 0.02 | 0.01 |
| Huberg0 | 0.04 | 0.06 | 0.05 | 0.03 | 0.04 | 0.05 | 0.06 | 0.06 | 0.06 | 0.01 | 0.05 | 0.07 |
| Huberg0.5 | 0.04 | 0.05 | 0.04 | 0.03 | 0.04 | 0.06 | 0.06 | 0.06 | 0.07 | 0.01 | 0.05 | 0.08 |
| Huberg0.8 | 0.04 | 0.04 | 0.04 | 0.03 | 0.03 | 0.05 | 0.06 | 0.07 | 0.08 | 0.01 | 0.05 | 0.08 |
| Huberm0 | 0.05 | 0.06 | 0.05 | 0.04 | 0.04 | 0.04 | 0.06 | 0.05 | 0.05 | 0.03 | 0.04 | 0.06 |
| Huberm0.5 | 0.03 | 0.05 | 0.03 | 0.04 | 0.04 | 0.04 | 0.07 | 0.05 | 0.06 | 0.01 | 0.03 | 0.05 |
| Huberm0.8 | 0.04 | 0.05 | 0.04 | 0.03 | 0.04 | 0.04 | 0.06 | 0.06 | 0.07 | 0.00 | 0.04 | 0.08 |
| Copula | 0.00 | 0.00 | 0.39 | 0.00 | 0.07 | 0.05 | 0.00 | 0.00 | 0.38 | 0.00 | 0.08 | 0.15 |
| Spearman | 0.02 | 0.05 | 0.04 | 0.03 | 0.04 | 0.05 | 0.03 | 0.04 | 0.06 | 0.05 | 0.05 | 0.06 |
| Kendall | 0.02 | 0.05 | 0.05 | 0.05 | 0.04 | 0.05 | 0.05 | 0.06 | 0.06 | 0.07 | 0.05 | 0.05 |

published for multivariate cusum-type statistics. Aue et al. (2009) choose \( b_T = \log_{10}(T) \), but only consider very weak serial dependence. Wied (2016) prefer \( b_T = \lfloor T^{1/4} \rfloor \) but only looked at MA(1) processes.

First we look at the behaviour under the null hypothesis. Results are based on 1000 runs. Under normally distributed innovations we see that all tests hold their size reasonably well, see Table 5. Multivariate cusum type tests are conservative if \( p = 5 \), \( \rho = 0.5 \) and small \( T \). Furthermore the covariance based tests is very anti conservative for small \( T \), large \( p \) and no serial dependence. There is not much changing if we look at \( t_3 \) distributed innovations, see Table 6. Only the correlation based test behaves different and gets strongly anti conservative, especially if \( p \) is large.

We want to look also at the behaviour under the alternative. Note that there are countless different scenarios one can look at. We concentrate on two. First we want to verify if projection type statistics are indeed given an advantage if changes are uniformly into one direction.
Therefore we look at model (12) with

$$\epsilon_t \sim \begin{cases} N(0, I_p) & t = 1, \ldots, 200 \\ N(0, \Sigma_\Delta) & t = 201, \ldots, 400 \end{cases} \quad \text{with} \quad \Sigma_\Delta = \begin{pmatrix} 1 & 0.2 & \Delta & 0.2 \\ 0.2 & 1 & \Delta & \Delta \\ \Delta & \Delta & 1 & 0.2 \\ 0.2 & \Delta & 0.2, & 1 \end{pmatrix}$$

and $\Delta \in [-0.2, 0.2]$. So depending on $\Delta$ we have a change into the same direction ($\Delta = 0.2$) or into opposite directions ($\Delta = -0.2$). As we can see in Figure 3 projection type tests like Spearman and Kendall outperform truly multivariate ones if the change is into the same direction and have problems to detect a change if it is into different directions. Surprisingly the Copula-test shows similar behaviour. For larger $T$ this test would also be able to detect changes if $\Delta = -0.2$ but the power for $\Delta = 0.2$ stays clearly higher. We do not have any explanation for this. All other tests have surprisingly higher power if $\Delta$ is negative. We do not see significant changes of the order of these tests depending on $\Delta$. The Cor and Huber0.8 are generally the ones with the highest power, slightly outperforming Cov, which is a little more conservative under the null hypothesis.

We want to investigate the influence of heavy tails. Therefore we look at model (12) with

$$\epsilon_t \sim \begin{cases} t_{df}(0, I_p) & t = 1, \ldots, 200 \\ t_{df}(0, \Sigma_2) & t = 201, \ldots, 400 \end{cases} \quad \text{with} \quad \Sigma_\Delta = \begin{pmatrix} 1 & 0.3 & 0 & 0.3 \\ 0.3 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.3 \\ 0.3 & 0 & 0.3, & 1 \end{pmatrix}$$

and $df \in \{1, 2, 3, 5, 8, 10, 20, 50, 200\}$. Note that we choose the special structure with some covariances changing and others not to achieve a fair comparison between projection type tests and truly multivariate ones. Results are based on 2000 runs. We see in Figure 3 that even under 200 degrees of freedom, which is hardly distinguishable from a normal distribution, the Huber08 outperforms all the other tests including Cor. The difference between Huber08 and Cor as well as Cov gets larger for less degrees of freedom. We see that apart from Huber0 all tests loose power if the innovation distribution gets more heavy tailed. But there is a difference how fast the decrease is. The loss for Cov and Cor is the largest and for Huber05 the smallest. The power of the correlation based test increases for 1 degree of freedom since the tests gets extremely anti conservative in this case. Only for very few degrees of freedom Huber0 and Huber0.5 outperform Huber0.8.
Figure 4: Empirical power under one change point and a change of the cross correlation structure of the innovations from $\varepsilon_t \sim t_d(0, I_4)$ for $t = 1, \ldots, 200$ to $\varepsilon_t \sim t_d(0, \Sigma_2)$ for $t = 201, \ldots, 400$ and $df \in \{1, 2, 3, 5, 8, 10, 20, 50, 200\}$.

5. Summary

We develop a new non-parametric and robust approach to detect change-points in possibly multivariate time series. The method can be easily adjusted to the type of change one is interested in. We explicitly propose tests for a change in location, scale or cross dependence. Simulations indicate that these test are almost as powerful as classical cusum type test under Gaussian data. In these settings they have similar power as other robust methods (if they already exist), while having usually a far lower computational complexity. Under heavy tails our new methods clearly outperform cusum type tests and usually also outperform existing robust methods.

Simulations reveal an interesting problem. Our tests become conservative under serial dependence. This effect seems to be more severe if the dimension $p$ of the time series is large. Classical cusum tests show the same behaviour. To the best of our knowledge there are no theoretical results describing this behaviour yet. If it was possible to correct the test statistic, the test would become more powerful under the alternative.

It remains to investigate the behaviour of our method under local and fixed alternatives. In the later case the long run covariance estimation does not converge to the theoretical value under stationarity anymore. In the multivariate case the estimated matrix gets even singular which creates an additional challenge.

Also of interest are the properties of the related change point estimator. The distribution of the cusum estimator can be found in Csörgő and Horváth (1997). There is also a similar result for a robust change point test estimator of a location change based on ranks (Gerstenberger, 2018) which has the same convergence rates as the classical cusum one. The straightforward estimator of change for our method is $\text{argmax}_{W_T(x)} W_T(x)^2$. One would expect that this estimator is more efficient under heavy tailed data but less efficient under Gaussianity.

References

Andrews, D.W., 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. Econometrica: Journal of the Econometric Society 59, 817-858.

Antoch, J., Hušková, M., Janic, A., Ledwina, T., 2008. Data driven rank test for the change point problem. Metrika 68, 1–15.

Asmussen, S., Glynn, P., Pitman, J., 1995. Discretization error in simulation of one-dimensional reflecting brownian motion. The Annals of Applied Probability, 875–896.

Aue, A., Hörmann, S., Horváth, L., Reimherr, M., 2009. Break detection in the covariance structure of multivariate time series models. The Annals of Statistics 37, 4046–4087.
Appendix A. Proofs

We first show that the multivariate cusum statistic converges to a multivariate Brownian motion, if we fix $\mu$ and $\sigma$. Denote by

$$
M_n = \frac{1}{T} \left( \sum_{i=1}^{\lfloor nx \rfloor} Y_i - \frac{\lfloor nx \rfloor}{n} \sum_{i=1}^{n} Y_i \right)
$$

the multivariate cusum statistic with known location $\mu$ and scale $\sigma$.

**Proposition 1.** Let $(X_i)_{i \in \mathbb{N}}$ be a stationary and strongly mixing sequence with approximation constants $(a_k)_{k \in \mathbb{N}}$ fulfilling $a_k = O(k^{-1-\epsilon})$ for some $\epsilon > 0$ and $\Psi : \mathbb{R}^p \to \mathbb{R}^s$ be a bounded function such that $U = \text{Cov} (Y_1, Y_1) + 2 \sum_{h=1}^{\infty} \text{Cov} (Y_i, Y_{i+h})$ has only strictly positive eigenvalues, then

$$(M_n(t))_{t \in [0,1]} \xrightarrow{u} (M(t))_{t \in [0,1]}$$

where $(M_t)_{t \in [0,1]}$ is an $s$-dimensional Brownian motion with $\text{Cov}(W_t) = tU$.

**Proof.** We apply the functional central limit theorem 1.4 of Merlevede and Peligrad (2000). For this we first use a Cramer-Wold-device for elements in the $s$-dimensional Skorohod space, see for example Proposition 4.1 in Wooldridge and White (1988). So let $\lambda = \lambda_1, \ldots, \lambda_s$ be arbitrary with $\lambda^T \lambda = 1$, then

$$(M_n(t))_{t \in [0,1]} \xrightarrow{u} (M(t))_{t \in [0,1]} \quad \text{if and only if} \quad \left( \sum_{j=1}^{s} \lambda_j M_n(t)^{(j)} \right)_{t \in [0,1]} \xrightarrow{u} \left( \sum_{j=1}^{s} \lambda_j M_n(t)^{(j)} \right)_{t \in [0,1]} \quad (A.1)$$

By changing of summation, we see that (A.1) holds if an invariance principle applies to $\frac{1}{\sqrt{T}} \sum_{i=1}^{T} (\lambda^T Y_i)$, which is given in Merlevede and Peligrad (2000) under three conditions:

- $\alpha_T T \to 0$,  
- $\sum_{n=1}^{\infty} \int_{0}^{a_n} Q^2_{|\lambda^T Y_1|}(u) \ du < \infty$ where $Q_W(u) = \inf\{t \geq 0 : P(W > t) \leq u\}$,
- the long run variance of $(\lambda^T Y_i)_{i \in \mathbb{N}}$ is strictly positive.

The first condition is satisfied, since $\alpha_k = O(k^{-1-\epsilon})$. For the second we use that $\lambda^T Y_1$ is bounded such that $Q^2_{|\lambda^T Y_1|}(u) \leq C$ for some $C > 0$. The last condition is fulfilled since $U$ has only positive eigenvalues.

Now we show that estimating location and scale is asymptotically negligible.
Proposition 2. Let Assumptions 1-4 be fulfilled then

\[
(R_T(x))_{0 \leq x \leq 1} = \left[ \frac{1}{\sqrt{T}} \left( \sum_{t=1}^{T} Y_{i,T} - \frac{|T_x|}{T} \sum_{i=1}^{T} Y_{i,T} \right) - \frac{1}{\sqrt{T}} \left( \sum_{t=1}^{T} Y_i - \frac{|T_x|}{T} \sum_{i=1}^{T} Y_i \right) \right]_{0 \leq x \leq 1} \rightarrow 0. 
\]

Proof. The proof consists of two steps. First we show that \( R_T(x) \rightarrow 0 \) \( \forall x \in [0,1] \) and after that we show that \( (R_n(x))_{x \in [0,1]} \) is tight. We use the following covariance inequality

\[
|\text{Cov}(X_1, X_{1+h}) - \mathbb{E}(X_1)\mathbb{E}(X_{1+h})| \leq C \alpha_1^{1/p-1/q} ||X_1||_p ||X_{1+h}||_q \quad \text{(A.2)}
\]

for \( 1/p + 1/q = 1 \) which is derived in Ibragimov (1975) and Deo (1973) and holds also with a slightly different proof for \( p = \infty \), see Borovkova et al. (2001).

First we note that \( \mathbb{E}(R_T(x)) = 0 \) since

\[
\frac{1}{\sqrt{T}} \mathbb{E} \left( \sum_{t=1}^{T} \mathbb{E}(Y_{i,T} - Y_i | \sigma, \mu) - \frac{|T_x|}{T} \sum_{i=1}^{T} \mathbb{E}(Y_{i,T} - Y_i | \sigma, \mu) \right) = 0.
\]

Now we decompose \( \mathbb{R}^p \) into \( A_e(a_1), A_e(a_n), B_T, B_{1T}, \ldots, B_{IT}, E_{1T}, \ldots, E_r \) and the remaining part \( G \). We have not defined sets with the subindex \( T \) yet. For a set \( A \) we define \( A_T \) for some \( \epsilon > 0 \) as:

\[
A_T = \{ x \in \mathbb{R}^p : \exists y \in A : \| x - y \| \leq T^{-\frac{1}{2} - \epsilon} \}.
\]

We can find some \( \epsilon > 0 \) and \( T_0 \) such that these sets are disjoint. Let \( i \in \{1, \ldots, s\} \) be arbitrary. Using the \( c_r \) inequality several times we can split the expectation of \( R^2(x) \) the following way:

\[
\mathbb{E}(R^2(x)) \leq \frac{2q}{T} \sum_{j=1}^{n} \mathbb{E}([1 - \frac{k}{T}] \sum_{t=1}^{k} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in A_e(a_j))^{2} - \frac{k}{T} \sum_{t=k+1}^{T} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in A_e(a_j))^{2}
\]

\[
+ \frac{2q}{T} \sum_{j=1}^{l} \mathbb{E}([1 - \frac{k}{T}] \sum_{t=1}^{k} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in B_T) - \frac{k}{T} \sum_{t=k+1}^{T} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in B_T)^{2}
\]

\[
+ \frac{2q}{T} \sum_{j=1}^{r} \mathbb{E}([1 - \frac{k}{T}] \sum_{t=1}^{k} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in E_{jT}) - \frac{k}{T} \sum_{t=k+1}^{T} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in E_{jT})^{2}
\]

\[
+ \frac{2q}{T} \mathbb{E}([1 - \frac{k}{T}] \sum_{t=1}^{k} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in G_T) - \frac{k}{T} \sum_{t=k+1}^{T} [Y_{i,T}^{(i)} - Y_{i}^{(i)}] | X_{i} \in G_T)^{2}
\]

\[
= A_1 + \ldots + A_n + B + B_1 + \ldots + B_l + E_1 + \ldots + E_r + G
\]

where \( q = n + l + r + 2 \). Let without loss of generality \( D_{\sigma} = I_p \) and \( \mu = 0 \). First we treat \( G \), since \( \Psi \) is 2 times continuous differentiable in \( G_T \) we use a Taylor expansion of first order:

\[
\Psi \left( D_{\sigma}^{-1} |X_i - \hat{\mu}| \right)^{(k)} = \Psi \left( X_i + [D_{\sigma} - I_p]^{-1} X_i - D_{\sigma} \hat{\mu} \right)^{(k)} = \Psi(X_i)^{(k)} (D_{\sigma} - I_p)^{-1} X_i - D_{\sigma} \hat{\mu})
\]

\[
+ ([D_{\sigma} - I_p] X_i - D_{\sigma} \hat{\mu})^{T} R_T(X_i) ([D_{\sigma} - I_p] X_i - D_{\sigma} \hat{\mu})
\]

\[
= (A.3)
\]

\[
= (A.4)
\]

\[
= (A.5)
\]

23
where \( R_T(X_i) \) depends on \( \hat{\mu} \) and \( \hat{\sigma} \). Since \( x^T \Psi^{(i)''}(x) x \leq C_2 \), \( X_i^T R_T(X_i) X_i \) is bounded. In the following we abbreviate \( \partial \Psi^{(i)''}(x) \big|_{x=y} \) as \( \Psi^{(i)'}(y) \)

\[
G = \frac{2q}{T} \mathbb{E} \left\{ \left[ 1 - \frac{k}{T} \right] \sum_{t=1}^{k} \Psi'(X_t) I_{\{X_t \in \mathcal{G}_T\}} [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}] \right. \\
+ \left. \left\{ 1 - \frac{k}{T} \right\} \sum_{t=1}^{k} [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}]^T R_T(X_t) [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}] I_{\{X_t \in \mathcal{G}_T\}} \right. \\
- \left. \frac{k}{T} \sum_{t=1}^{k} \Psi'(X_t) I_{\{X_t \in \mathcal{D}_T\}} [(D_\sigma - I_p) X_t - G_\sigma \hat{\mu}] \right. \\
- \left. \frac{k}{T} \sum_{t=k+1}^{T} [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}]^T R_T(X_t) [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}] I_{\{X_t \in \mathcal{G}_T\}} \right]^2 \right\} \\
\leq \frac{8 \cdot 2q}{T} \mathbb{E} \left\{ \left( \sum_{j=1}^{p} \left( \frac{1}{\hat{\sigma}} - 1 \right) \left( 1 - \frac{k}{T} \right) \sum_{t=1}^{k} \left[ \Psi_j^{(i)'}(X_t) X_t^{(j)} I_{\{X_t \in \mathcal{G}_T\}} \right] \right. \\
- \left. \mathbb{E} \left( \Psi_j^{(i)'}(X_1) X_1^{(j)} I_{\{X_1 \in \mathcal{G}_T\}} \right) \right. \\
- \left. \frac{k}{T} \sum_{t=k+1}^{T} \left[ \Psi_j^{(i)'}(X_t) X_t^{(j)} I_{\{X_t \in \mathcal{G}_T\}} - \mathbb{E} \left( \Psi_j^{(i)'}(X_1) I_{\{X_1 \in \mathcal{G}_T\}} \right) \right. \\
- \left. \mathbb{E} \left( \Psi_j^{(i)'}(X_1) I_{\{X_1 \in \mathcal{D}_T\}} \right) \right. \\
+ \left. \frac{8 \cdot 2q}{T} \mathbb{E} \left\{ \left( \sum_{j=1}^{p} \frac{1}{\hat{\sigma}} \hat{\mu}^{(j)} \left( \frac{k}{T} \sum_{t=1}^{k} \left[ \Psi_j^{(i)'}(X_t) I_{\{X_t \in \mathcal{G}_T\}} - \mathbb{E} \left( \Psi_j^{(i)'}(X_1) I_{\{X_1 \in \mathcal{D}_T\}} \right) \right. \\
- \left. \mathbb{E} \left( \Psi_j^{(i)'}(X_1) I_{\{X_1 \in \mathcal{G}_T\}} \right) \right. \\
+ \left. \frac{8 \cdot 2q}{T} \mathbb{E} \left\{ \left( \sum_{t=1}^{k} [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}]^T R_T(X_t) [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}] I_{\{X_t \in \mathcal{G}_T\}} \right)^2 \right\} \\
+ \left. \frac{8 \cdot 2q}{T} \mathbb{E} \left\{ \left( \sum_{t=k+1}^{T} [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}]^T R_T(X_t) [(D_\sigma - I_p) X_t - D_\sigma \hat{\mu}] I_{\{X_t \in \mathcal{G}_T\}} \right)^2 \right\} \right) \\
= G_1 + G_2 + G_3 + G_4. \]
For $G_1$ we get

$$G_1 \leq \frac{8 \cdot 2^q}{T} 2^{p+1}(1 - \frac{k}{T})^2 \sum_{j=1}^{p} E \left( \left\{ \left( \frac{1}{\sigma} - 1 \right) \sum_{t=1}^{k} \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] - E \left( \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] \right) \right\}^2 \right)$$

$$+ \frac{8 \cdot 2^q}{T} 2^{p+1}(1 - \frac{k}{T})^2 \sum_{j=1}^{p} E \left( \left\{ \left( \frac{1}{\sigma} - 1 \right) \sum_{t=k+1}^{T} \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] - E \left( \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] \right) \right\}^2 \right) = G_{1,1} + G_{1,2}. $$

In the following we look only at one summand $j \in \{1, \ldots, p\}$ :

$$E \left( \left\{ \left( \frac{1}{\sigma} - 1 \right) \sum_{t=1}^{k} \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] - E \left( \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] \right) \right\}^2 \right)$$

$$\leq \frac{1}{T^{1-2\epsilon}} E \left( \sum_{t=1}^{k} \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] - E \left( \left[ \mu_j^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] \right) \right)^2 \right)$$

$$\leq \frac{1}{T^{1-2\epsilon}} C T. $$

Together with the factor $\frac{32}{T} 2^{p+1}(1 - \frac{k}{T})^2$ this converges to 0. We get the same result for the remaining summands and $G_{1,2}$. For $G_3$ and $G_4$ we use the boundedness-condition:

$$G_3 = \frac{8 \cdot 2^q}{T} E \left( \left\{ 1 - \frac{k}{T} \right\} \sum_{t=1}^{k} \left[ (D \sigma - I_p) X_t - D \sigma \hat{\mu} \right]^T R_T(X_t) \right)$$

$$\left[ (D \sigma - I_p) X_t - D \sigma \hat{\mu} \right] I_{\{X_t \in G_T\}} \right)^2 \right)$$

$$\leq \frac{8 \cdot 2^q}{T} \left( \sum_{t=1}^{k} C T^{-1+2\epsilon} \right) \to 0. $$

Now we look at $B$ :

$$C \leq \frac{2^{q+1}}{T} E \left( \sum_{t=1}^{k} \left[ Y_{t,T}^{(i)} - Y_{T}^{(i)} \right] I_{X_t \in B_T} \right)^2 \right) + \frac{2^{q+1}}{T} E \left( \sum_{t=k+1}^{T} \left[ Y_{t,T}^{(i)} - Y_{T}^{(i)} \right] I_{X_t \in B_T} \right)^2 \right)$$

$$\leq \frac{2^{q+1}}{T} \sum_{t=1}^{k} \sum_{s=1}^{T} \left( L^{\frac{1}{2+\epsilon}} L^{\frac{1}{2+\epsilon}} T^{\frac{1}{2+\epsilon}} \right) E(I_{X_t \in B_T} I_{X_t \in B_T}) \leq \frac{2^{q+2}}{T} T^{2-\frac{1}{2+\epsilon}} \to 0. $$

For $j \in \{1, \ldots, n\}$ arbitrary we look at

$$B_j \leq \frac{2^{q+1}}{T} E \left( \sum_{t=1}^{k} \left[ Y_{t,T}^{(i)} - Y_{T}^{(i)} \right] I_{X_t \in B_{j,T}} \right)^2 \right) + \frac{2^{q+1}}{T} E \left( \sum_{t=k+1}^{T} \left[ Y_{t,T}^{(i)} - Y_{T}^{(i)} \right] I_{X_t \in B_{j,T}} \right)^2 \right)$$

$$= B_{j,1} + B_{j,2} $$

25
and get for \( B_{j,1} \):

\[
B_{j,1} \leq \frac{2^{q+1}p}{T} \sum_{j=1}^{p} \mathbb{E} \left( \left[ \sum_{t=1}^{k} \left( \frac{\psi(i)}{\sigma(i)} \left( \frac{X_{t(j)}^{(i)} - \hat{\mu}^{(i)}}{\sigma(j)} - \frac{X_{\hat{t}(j)}^{(i)} - \hat{\mu}^{(i)}}{\sigma(j)} \right) \right) \right]^{2} \right)
\]

\[
\leq \frac{2^{q+1}p}{T} \cdot pk^{2}LT^{-\frac{1}{2}+\delta} \rightarrow 0.
\]

For \( j \in \{1, \ldots, n\} \) arbitrary we look at \( A_{j} \)

\[
A_{j} \leq \frac{2^{q+1}}{T} \mathbb{E} \left( \left[ \sum_{t=1}^{T} \{ Y_{t,T}^{(i)} - Y_{T}^{(i)} \} I_{X_{t}\in A_{j}} \right]^{2} \right)
+ \frac{2^{q+1}}{T} \mathbb{E} \left( \left[ \sum_{t=k+1}^{T} \{ Y_{t,T}^{(i)} - Y_{T}^{(i)} \} I_{X_{t}\in A_{j}} \right]^{2} \right) = A_{j,1} + A_{j,2}
\]

where for \( A_{j,1} \) we have

\[
\leq \frac{2^{q+1}}{T} \mathbb{E} \left( \left[ \sum_{t=1}^{k} \{ Y_{t,T}^{(i)} - Y_{T}^{(i)} \} I_{X_{t}\in A_{j}} \leq T^{-\frac{1}{2p}-\delta} \right]^{2} \right)
= \frac{2^{q+1}}{T} \mathbb{E} \left( \left[ \sum_{t=1}^{k} \{ Y_{t,T}^{(i)} - Y_{T}^{(i)} \} I_{T^{-\frac{1}{2p}-\delta} < |X_{t} - a_{j}| < \epsilon} \right]^{2} \right) = A_{j,1,1} + A_{j,1,2}
\]

For \( A_{j,1,1} \) we use that \( P(|X_{t} - a_{j}| \leq T^{-\frac{1}{2p}-\delta}) \leq C \cdot T^{-\frac{1}{2}-\delta} \) to show that \( A_{j,1,1} \to 0 \):

\[
\frac{2^{q+1}}{T} \mathbb{E} \left( \left[ \sum_{t=1}^{k} \{ Y_{t,T}^{(i)} - Y_{T}^{(i)} \} I_{X_{t}\in A_{j}} \leq T^{-\frac{1}{2p}-\delta} \right]^{2} \right)
\]

\[
\leq \frac{2^{q+1}}{T} \mathbb{E} \left( \left[ \sum_{t=1}^{k} 2K I_{X_{t}\in A_{j}} \leq T^{-\frac{1}{2p}-\delta} \right]^{2} \right)
\]

\[
\leq \frac{2^{q+2}K^{2}}{T} \mathbb{E} \left( \left[ \sum_{t=1}^{k} I_{X_{t}\in A_{j}} \leq T^{-\frac{1}{2p}-\delta} - P(|X_{1} - a_{j}| \leq T^{-\frac{1}{2p}-\delta}) \right]^{2} \right)
+ \frac{2^{q+2}K^{2}k^{2}}{T} P(|X_{1} - a_{j}| \leq T^{-\frac{1}{2p}-\delta})^{2}
\]

\[
\leq \frac{2^{q+2}}{T} C_{1} C_{2} \cdot T^{-1-\delta} + \frac{2^{q+2}K^{2}k^{2}}{T} \to 0
\]

For \( A_{j,1,2} \) we use a Taylor-decomposition and get analogously to the decomposition of \( D, A_{1,2,1}, A_{1,2,2}, A_{1,2,3} \) and \( A_{1,2,4} \) where we split \( A_{1,2,1} \) into \( A_{1,2,1,1} \) and \( A_{1,2,1,2} \) similar to \( D_{1} \) which is split into \( D_{1,1} \)
and $D_1$. So $A_{1,2,1,1}$ has the following form:

$$A_{1,2,1,1} = \frac{2^{q+1}}{T} 2^{p+1} (1 - \frac{k}{T})^2 \sum_{m=1}^{p} \mathbb{E} \left( \left\{ \left( \frac{1}{\sigma} - 1 \right) \sum_{t=1}^{k} \left[ \Psi_{m}^{(i)} (X_t) X_t^{(m)} I \{ T - \frac{1}{T^2} - \delta_1 < |X_t - a_j| < \epsilon \} \right] \right\}^2 \right)$$

For every summand $m$ we get

$$\mathbb{E} \left( \left\{ \frac{1}{\sigma} - 1 \right\} \sum_{t=1}^{k} \left[ \Psi_{m}^{(i)} (X_t) X_t^{(m)} I \{ T - \frac{1}{T^2} - \delta_1 < |X_t - a_j| < \epsilon \} \right] \right)^2 \leq \frac{1}{T^{1-2\delta_2}} \mathbb{E} \left( \left\{ \sum_{t=1}^{k} \left[ \Psi_{m}^{(i)} (X_t) X_t^{(m)} I \{ T - \frac{1}{T^2} - \delta_1 < |X_t - a_j| < \epsilon \} \right] \right\}^2 \right)$$

$$\leq \sqrt{ \frac{1}{T^{1-2\delta_2}} CT \mathbb{E} \left( \left[ \Psi_{m}^{(i)} (X_t) X_t^{(m)} I \{ T - \frac{1}{T^2} - \delta_1 < |X_t| < \epsilon \} \right]^2 \right) \right} \leq \sqrt{ \frac{1}{T^{1-2\delta_2}} CT^{2-\delta} }.$$

Together with the remaining factor $\frac{2^{q+1}}{T} 2^{p+1} (1 - \frac{k}{T})^2$ this converges to 0. Analogous calculations shows convergence for the other summands.

Finally we look at $E_1, \ldots, E_r$. To shorten notation we assume $E_1 = \{ x \in \mathbb{R}^p : x_1 = 0 \}$. Define
by $N_s = \{ x \in \mathbb{R} : \frac{s-1}{T^2} \leq x < \frac{s}{T^2} \}$ for $s = -T, \ldots, T$, then we decompose $E_1$ further:

\[
\frac{2^q}{T} \mathbb{E} \left( \left[ 1 - \frac{k}{T} \right] \sum_{t=1}^{k} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{Y_t \in E_1,T} - \frac{k}{T} \sum_{t=k+1}^{T} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{Y_t \in E_1,T} \right)^2
\]

\[
= \frac{2^q}{T} \mathbb{E} \left( \left[ 1 - \frac{k}{T} \right] \sum_{s=-T}^{T} I_{\mu(s) \in N_s} \left( \sum_{t=1}^{k} \sum_{v=-T}^{T} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{X_v \in N_v} \right) \right)^2
\]

\[
- \frac{k}{T} \sum_{t=k+1}^{T} \sum_{s=-T}^{T} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{X_t \in N_v} \right)^2
\]

\[
= \frac{2^q}{T} \mathbb{E} \left( \left[ 1 - \frac{k}{T} \right] \sum_{s=-T}^{T} I_{\mu(s) \in N_s} \left( \sum_{t=1}^{k} \sum_{v=-T}^{T} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{X_v \in N_v} \right) \right)^2
\]

\[- \mathbb{E} \left( [Y_{1,T}^{(i)} - Y_1^{(i)}] I_{X_1 \in N_v} \right)^2
\]

\[
\leq \frac{2^q}{T} \mathbb{E} \left( \sum_{s=-T}^{T} I_{\mu(s) \in N_s} \left( \sum_{t=1}^{k} \sum_{v=-T}^{T} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{X_v \in N_v} - \mathbb{E} \left( [Y_{1,T}^{(i)} - Y_1^{(i)}] I_{X_1 \in N_v} \right) \right)^2
\]

\[
+ \frac{2^q}{T} \mathbb{E} \left( \sum_{s=-T}^{T} I_{\mu(s) \in N_s} \left( \sum_{t=k+1}^{T} \sum_{v=-T}^{T} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{X_v \in N_v} - \mathbb{E} \left( [Y_{1,T}^{(i)} - Y_1^{(i)}] I_{X_1 \in N_v} \right) \right)^2
\]

\[= C_1 + C_2
\]

Depending on which sector $\mu$ and $X_t$ lie most summands are 0. So for example $C_1$ simplifies to

\[
\frac{2^{q+1}}{T} \mathbb{E} \left( \sum_{s=-T}^{T} I_{\mu(s) \in N_s} \left( \sum_{t=1}^{k} \sum_{v=-T}^{T} [Y_{t,T}^{(i)} - Y_t^{(i)}] I_{X_v \in N_v} - \mathbb{E} \left( [Y_{1,T}^{(i)} - Y_1^{(i)}] I_{X_1 \in N_v} \right) \right)^2
\]

Now we want to show that $(\mathbb{R}_T(x))_{x \in [0,1]}$ respectively every component of it, is tight. We use the moment criteria by Billingsley (1968) with $\gamma = 4$ and $\alpha = 2$ so we have to show that

\[
\mathbb{E}(|R_n(y^{(i)} - R_n(x^{(i)})|^4) \leq C_5 |y - x|^2, \ \forall \ x, y \in [0,1], \ T \in \mathbb{N}.
\]

We denote $k = |T x| > j = |T y|$ and follow the same ideas as before, namely splitting the sums by the $c_s$ inequality into $A_1 + \ldots + A_n + B + B_1 + \ldots + B_l + E_1 + \ldots + E_r + G$ and treating
them as before. The only difference now is that we have to look at 4-th moments instead of 2nd moments.

\[
\mathbb{E}[R_T(y^{(i)} - R_T(x^{(i)})]^4
\]
\[
= \frac{1}{T^2} \mathbb{E} \left[ \left(1 - \frac{k}{T} \right) \sum_{t=1}^{k} (Y_{t,T}^{(i)} - Y_t^{(i)}) - \frac{k}{T} \sum_{t=k+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) \right]^4
\]
\[
- \left(1 - \frac{j}{T} \right) \sum_{t=1}^{j} (Y_{t,T}^{(i)} - Y_t^{(i)}) + \frac{j}{T} \sum_{t=j+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) \right|^4
\]
\[
\leq \frac{8^q}{T^2} \sum_{j=1}^{n} \mathbb{E} \left[ \left(1 - \frac{k}{T} \right) \sum_{t=1}^{k} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in A_T\}} - \frac{k}{T} \sum_{t=k+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in A_T\}} \right]^4
\]
\[
- \left(1 - \frac{j}{T} \right) \sum_{t=1}^{j} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in B_T\}} + \frac{j}{T} \sum_{t=j+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in B_T\}} \right|^4
\]
\[
\leq \frac{8^q}{T^2} \sum_{j=1}^{l} \mathbb{E} \left[ \left(1 - \frac{k}{T} \right) \sum_{t=1}^{k} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in E_T\}} - \frac{k}{T} \sum_{t=k+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in E_T\}} \right]^4
\]
\[
- \left(1 - \frac{j}{T} \right) \sum_{t=1}^{j} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in E_T\}} + \frac{j}{T} \sum_{t=j+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in E_T\}} \right|^4
\]
\[
\leq \frac{8^q}{T^2} \sum_{j=1}^{l} \mathbb{E} \left[ \left(1 - \frac{k}{T} \right) \sum_{t=1}^{k} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in G_T\}} - \frac{k}{T} \sum_{t=k+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in G_T\}} \right]^4
\]
\[
- \left(1 - \frac{j}{T} \right) \sum_{t=1}^{j} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in G_T\}} + \frac{j}{T} \sum_{t=j+1}^{T} (Y_{t,T}^{(i)} - Y_t^{(i)}) I_{\{X_t \in G_T\}} \right|^4
\]
\[
= \frac{8^q}{T^2} (A_1 + \ldots + A_n + B + B_1 + \ldots + B_l + E_1 + \ldots + E_r + G).
\]
We look exemplarily at $G$: 

\[
G \leq 8^{12} \left[ \frac{j-k}{T} \right]^4 \mathbb{E} \left( \left\{ \sum_{j=1}^{p} \left( \frac{1}{\hat{\sigma}(j)} - 1 \right) \sum_{t=1}^{k} \left[ \Psi_{j}^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] - \mathbb{E} \left( \Psi_{j}^{(i)}(X_1) X_1^{(j)} I_{\{X_1 \in D_T\}} \right) \right\}^4 \right) \\
+ 8^{12} \left[ \frac{T-j}{T} \right]^4 \mathbb{E} \left( \left\{ \sum_{j=1}^{p} \left( \frac{1}{\hat{\sigma}(j)} - 1 \right) \sum_{t=k+1}^{j} \left[ \Psi_{j}^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] - \mathbb{E} \left( \Psi_{j}^{(i)}(X_1) X_1^{(j)} I_{\{X_1 \in G_T\}} \right) \right\}^4 \right) \\
+ 8^{12} \left[ \frac{j-k}{T} \right]^4 \mathbb{E} \left( \left\{ \sum_{j=1}^{p} \left( \frac{1}{\hat{\sigma}(j)} - 1 \right) \sum_{t=k+1}^{j} \left[ \Psi_{j}^{(i)}(X_t) X_t^{(j)} I_{\{X_t \in G_T\}} \right] - \mathbb{E} \left( \Psi_{j}^{(i)}(X_1) X_1^{(j)} I_{\{X_1 \in G_T\}} \right) \right\}^4 \right)
\]
+ 8^{12} \left( \frac{T-j}{T} \right)^{4} \mathbb{E} \left( \left\{ \sum_{l=k+1}^{j} \left( (D_{\sigma} - I_p) X_{l} - D_{\sigma} \hat{\mu} \right)^{T} R_{T}(X_{l}) \right\} \left( (D_{\sigma} - I_p) X_{l} - D_{\sigma} \hat{\mu} \right) I_{\{X_{l} \in G_{T}\}} \right) \\
+ 8^{12} \left( \frac{j-k}{T} \right)^{4} \mathbb{E} \left( \left\{ \sum_{l=j+1}^{T} \left( (D_{\sigma} - I_p) X_{l} - D_{\sigma} \hat{\mu} \right)^{T} R_{T}(X_{l}) \right\} \left( (D_{\sigma} - I_p) X_{l} - D_{\sigma} \hat{\mu} \right) I_{\{X_{l} \in G_{T}\}} \right) \\
+ 8^{12} \left( \frac{k}{T} \right)^{4} \mathbb{E} \left( \left\{ \sum_{l=k+1}^{j} \left( (D_{\sigma} - I_p) X_{l} - D_{\sigma} \hat{\mu} \right)^{T} R_{T}(X_{l}) \right\} \left( (D_{\sigma} - I_p) X_{l} - D_{\sigma} \hat{\mu} \right) I_{\{X_{l} \in G_{T}\}} \right) \\
= 8^{12} \sum_{i=1}^{12} T A_{i}

For TA_{1}, \ldots, TA_{12} one applies the 4th-moment inequality for strongly mixing sequences ([Meyerleve and Peligrad, 2000]). In case of TA_{1} this yields:

\begin{equation}
TA_{1} \leq (x - y)^{2} q^{2p} \sqrt{T^{-1/2 + r}} + k^{2} D
\end{equation}

and together with the prefactor 1/T^2 we get the desired bound containing (x - y)^2. For the remainder terms of the Taylor series we use the boundedness and get for example for TA_{12}:

\begin{equation}
TA_{12} \leq (j - k)^{2} T^{2} D (T^{1-2r})^{4}
\end{equation}

which also results together with the prefactor 1/T^2 an upper bound containing (x - y)^2. \hfill \Box

Now we show consistency of the long run variance estimator \( \hat{U} \):

**Proposition 3.** Under Assumptions 1-5, \( \hat{U} \) is weakly consistent for U.

**Proof.** Consistency of the long run variance estimator under known \( \mu \) and \( \sigma \) is known under very general conditions, see [Andrews, 1991] and [Jong and Davidson, 2000]. If r-th moments exists, it is required that mixing coefficients \((a_{k})_{k \in \mathbb{N}} \) fulfill \( a_{k} = O(k^{r/(r-2)}) \) (Theorem 2.1 of Jong and Davidson, 2000) which is covered by Assumption 1. In Jong and Davidson (2000) even the case of estimated standardization is covered as long as \( \Psi \) is differentiable and some technical conditions are fulfilled. Since we also want work with discontinuous \( \Psi \) we proof the asymptotic negligibility of the estimation of \( \mu \) and \( \sigma \). Therefore we use the following decomposition

\begin{align*}
&\frac{1}{T} \sum_{s, t} \left( Y_{s, T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(i)} \right) \left( Y_{t, T}^{(j)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(j)} \right) k \left( \frac{s-t}{b_{T}} \right) \\
&- \frac{1}{T} \sum_{s, t} \left( Y_{s, T}^{(i)} - \mathbb{E}[Y_{1}^{(i)}] \right) \left( Y_{s, T}^{(j)} - \mathbb{E}[Y_{1}^{(j)}] \right) k \left( \frac{s-t}{b_{T}} \right) \\
&= \frac{1}{T} \sum_{s, t} \left( Y_{s, T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(i)} \right) \left( Y_{t, T}^{(j)} - Y_{t}^{(j)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(j)} \right) k \left( \frac{s-t}{b_{T}} \right) \\
&+ \frac{1}{T} \sum_{s, t} \left( Y_{t, T}^{(j)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(j)} \right) \left( Y_{s, T}^{(i)} - Y_{s}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(i)} \right) k \left( \frac{s-t}{b_{T}} \right) \\
&+ \frac{1}{T} \sum_{s, t} \left( Y_{s, T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(i)} \right) \left( \mathbb{E}[Y_{1}^{(j)}] - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(j)} \right) k \left( \frac{s-t}{b_{T}} \right) \\
&+ \frac{1}{T} \sum_{s, t} \left( Y_{t, T}^{(j)} - \mathbb{E}[Y_{1}^{(j)}] \right) \left( \mathbb{E}[Y_{1}^{(i)}] - \frac{1}{T} \sum_{u=1}^{T} Y_{u, T}^{(i)} \right) k \left( \frac{s-t}{b_{T}} \right) \\
&= K_{1} + K_{2} + K_{3} + K_{4}.
\end{align*}
Every summand has to converge against 0 in probability. Like in the proof of Theorem 1 we expand them by indicator functions. Since $P(A_{T}^{C}) → 0$ we assume in the following the case of $A_{T}$. An expansion of $K_{1}$ yields

$$K_{1} = \frac{1}{T} \sum_{s,t} \left( Y_{s,T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u,T}^{(i)} \right) I_{X_{s},X_{u} \in A_{r}(a_{1})} \left( Y_{t,T}^{(j)} - Y_{t}^{(j)} - \frac{1}{T} \sum_{v=1}^{T} Y_{v,T}^{(j)} - Y_{u}^{(j)} \right)$$

$$I_{X_{s},X_{u} \in A_{r}(a_{1})}k \left( \frac{s - t}{bT} \right)$$

$$+ \frac{1}{T} \sum_{s,t} \left( Y_{s,T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u,T}^{(i)} \right) I_{X_{s},X_{u} \in A_{r}(a_{1})} \left( Y_{t,T}^{(j)} - Y_{t}^{(j)} - \frac{1}{T} \sum_{v=1}^{T} Y_{v,T}^{(j)} - Y_{u}^{(j)} \right)$$

$$I_{X_{s},X_{u} \in A_{r}(a_{2})}k \left( \frac{s - t}{bT} \right)$$

$$:$$

$$+ \frac{1}{T} \sum_{s,t} \left( Y_{s,T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u,T}^{(i)} \right) I_{X_{s},X_{u} \in G_{T}} \left( Y_{t,T}^{(j)} - Y_{t}^{(j)} - \frac{1}{T} \sum_{v=1}^{T} Y_{v,T}^{(j)} - Y_{u}^{(j)} \right)$$

$$I_{X_{s},X_{u} \in G_{T}}k \left( \frac{s - t}{bT} \right) = K_{1A_{1}A_{1}} + K_{1A_{1}A_{2}} + K_{1A_{1}A_{3}} + \ldots + K_{1GG}$$

Let us look exemplarily at $K_{1GG}$ where we use the Taylor expansion \[(X,3)\]. We omit the indicator functions in the following to improve readability:

$$K_{1GG} = \frac{1}{T} \sum_{m=1}^{p} \left( 1 - \frac{1}{\sigma(m)} \right) \sum_{s,t=1}^{T} \left[ Y_{s,T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u}^{(i)} \right]$$

$$\left[ \Psi_{m}^{(j)}(X_{t})X_{t}^{(m)} - \frac{1}{T} \sum_{u=1}^{T} \Psi_{m}^{(j)}(X_{u})X_{u}^{(m)} \right] k \left( \frac{s - t}{bT} \right)$$

$$+ \frac{1}{T} \sum_{m=1}^{p} \frac{\hat{\mu}^{(j)}}{\sigma^{(j)}} \sum_{s,t=1}^{T} \left[ Y_{s,T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u}^{(i)} \right]$$

$$\left[ \Psi_{m}^{(j)}(X_{t}) - \frac{1}{T} \sum_{u=1}^{T} \Psi_{m}^{(j)}(X_{u}) \right] k \left( \frac{s - t}{bT} \right)$$

$$+ \frac{1}{T} \sum_{s,t=1}^{T} \left[ Y_{s,T}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_{u}^{(i)} \right] \left\{ \left[ (D_{\sigma} - I_{p})X_{t} - D_{\sigma}\hat{\mu} \right]^{T} R_{T}(X_{t}) \left[ (D_{\sigma} - I_{p})X_{t} - D_{\sigma}\hat{\mu} \right] 

- \frac{1}{T} \sum_{u=1}^{T} \left[ (D_{\sigma} - I_{p})X_{u} - D_{\sigma}\hat{\mu} \right]^{T} R_{T}(X_{u}) \left[ (D_{\sigma} - I_{p})X_{u} - D_{\sigma}\hat{\mu} \right] \right\} k \left( \frac{s - t}{bT} \right)$$

$$+ \frac{1}{T} \sum_{n,m=1}^{p} \left( 1 - \frac{1}{\sigma(m)} \right) \left( 1 - \frac{1}{\sigma(n)} \right) \sum_{s,t=1}^{T} \left[ \Psi_{m}^{(j)}(X_{s})X_{s}^{(i)} - \frac{1}{T} \sum_{u=1}^{T} \Psi_{m}^{(j)}(X_{u})X_{u}^{(i)} \right]$$

32
\[
\left[ \Psi_n^{(j)}(X_t)X_t^{(n)} - \frac{1}{T} \sum_{u=1}^{T} \Psi_n^{(j)}(X_u)X_u^{(j)} \right] k \left( \frac{s-t}{b_T} \right) \\
+ \frac{1}{T} \sum_{n,m=1}^{p} \left( 1 - \frac{1}{\hat{\sigma}(m)} \right) \sum_{s,t=1}^{T} \left[ \Psi_m^{(i)}(X_s)X_s^{(i)} - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(i)}(X_u)X_u^{(i)} \right] \\
\left[ \Psi_n^{(j)}(X_t) - \frac{1}{T} \sum_{u=1}^{T} \Psi_n^{(j)}(X_u) \right] k \left( \frac{s-t}{b_T} \right) \\
+ \frac{1}{T} \sum_{n,m=1}^{p} \left( 1 - \frac{1}{\hat{\sigma}(n)} \right) \sum_{s,t=1}^{T} \left[ \Psi_m^{(i)}(X_s) - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(i)}(X_u) \right] \\
\left\{ \left[ (D_\sigma - I_p)X_t - D_\sigma \hat{\mu} \right] R_T(X_t) \right\} k \left( \frac{s-t}{b_T} \right) \\
- \frac{1}{T} \sum_{u=1}^{T} \left\{ \left[ (D_\sigma - I_p)X_u - D_\sigma \hat{\mu} \right] R_T(X_u) \right\} k \left( \frac{s-t}{b_T} \right) \\
+ \frac{1}{T} \sum_{n,m=1}^{p} \left( \frac{1}{\hat{\sigma}(m)} \right) \sum_{s,t=1}^{T} \left[ \Psi_m^{(i)}(X_s) - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(i)}(X_u) \right] \\
\left[ \Psi_n^{(j)}(X_t) - \frac{1}{T} \sum_{u=1}^{T} \Psi_n^{(j)}(X_u) \right] k \left( \frac{s-t}{b_T} \right) \\
+ \frac{1}{T} \sum_{n,m=1}^{p} \left( \frac{1}{\hat{\sigma}(n)} \right) \sum_{s,t=1}^{T} \left[ \Psi_m^{(i)}(X_s) - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(i)}(X_u) \right] \\
\left\{ \left[ (D_\sigma - I_p)X_t - D_\sigma \hat{\mu} \right] R_T(X_t) \right\} k \left( \frac{s-t}{b_T} \right) \\
- \frac{1}{T} \sum_{u=1}^{T} \left\{ \left[ (D_\sigma - I_p)X_u - D_\sigma \hat{\mu} \right] R_T(X_u) \right\} k \left( \frac{s-t}{b_T} \right) \\
\left[ \Psi_m^{(i)}(X_t)X_t^{(m)} - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(i)}(X_u)X_u^{(m)} \right] k \left( \frac{s-t}{b_T} \right) \\
+ \frac{1}{T} \sum_{m=1}^{p} \left( \frac{1}{\hat{\sigma}(m)} \right) \sum_{s,t=1}^{T} \left[ \Psi_m^{(i)}(X_s) - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(i)}(X_u) \right] \\
\left\{ \left[ (D_\sigma - I_p)X_t - D_\sigma \hat{\mu} \right] R_T(X_s) \right\} k \left( \frac{s-t}{b_T} \right)
\]
To shorten Notation, we assume that

\[ -\frac{1}{T} \sum_{u=1}^{T} [(D_{\sigma} - I_p)X_u - D_{\sigma} \bar{\mu}]^T R_T(X_u) [(D_{\sigma} - I_p)X_u - D_{\sigma} \bar{\mu}] k \left( \frac{s-t}{bT} \right) \]

\[ \left[ \Psi_m^{(j)}(X_t) - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(j)}(X_u) \right] k \left( \frac{s-t}{bT} \right) \]

+ \frac{1}{T} \sum_{s,t=1}^{T} \left\{ [(D_{\sigma} - I_p)X_t - D_{\sigma} \bar{\mu}]^T R_T(X_s) [(D_{\sigma} - I_p)X_t - D_{\sigma} \bar{\mu}] \right\}

− \frac{1}{T} \sum_{u=1}^{T} [(D_{\sigma} - I_p)X_u - D_{\sigma} \bar{\mu}]^T R_T(X_u) [(D_{\sigma} - I_p)X_u - D_{\sigma} \bar{\mu}] \}

\left\{ [(D_{\sigma} - I_p)X_t - D_{\sigma} \bar{\mu}]^T R_T(X_t) [(D_{\sigma} - I_p)X_t - D_{\sigma} \bar{\mu}] \right\}

− \frac{1}{T} \sum_{u=1}^{T} [(D_{\sigma} - I_p)X_u - D_{\sigma} \bar{\mu}]^T R_T(X_u) [(D_{\sigma} - I_p)X_u - D_{\sigma} \bar{\mu}] \}

k \left( \frac{s-t}{bT} \right)

The first three summands are the ones converging most slowly to 0. We examplarily look at the first summand for an arbitrary \( m \in \{1, \ldots, p\} \).

\[ \frac{1}{T} \left( 1 - \frac{1}{\sigma(m)} \right) \sum_{s,t=1}^{T} \left[ Y_s^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_u^{(i)} \right] \]

\[ \left[ \Psi_m^{(j)}(X_t)Y_t^{(m)} - \frac{1}{T} \sum_{u=1}^{T} \Psi_m^{(j)}(X_u)Y_u^{(m)} \right] k \left( \frac{s-t}{bT} \right) \]

= \frac{1}{T} \left( 1 - \frac{1}{\sigma(m)} \right) \sum_{s,t=1}^{T} \left[ Y_s^{(i)} - \mathbb{E}(Y_1^{(i)}) - \frac{1}{T} \sum_{u=1}^{T} Y_u^{(i)} - \mathbb{E}(Y_1^{(i)}) \right] \]

\[ \left[ \Psi_m^{(j)}(X_t)Y_t^{(m)} - \mathbb{E} \left\{ \Psi_m^{(j)}(X_1)Y_1^{(m)} \right\} \right] \]

− \frac{1}{T} \sum_{u=1}^{T} \left[ \Psi_m^{(j)}(X_u)Y_u^{(m)} - \mathbb{E} \left\{ \Psi_m^{(j)}(X_1)Y_1^{(m)} \right\} \right] k \left( \frac{s-t}{bT} \right)

To shorten Notation, we assume that \( \mathbb{E}(Y_1^{(i)}) = 0, \mathbb{E}\{\psi_m^{(j)}(X_t)Y_t^{(m)}\} = 0 \). A further expansion
yields

\[
\frac{1}{T} \left( 1 - \frac{1}{\hat{\sigma}^2(m)} \right) \sum_{s,t=1}^{T} \left[ Y_s^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_u^{(i)} \right] \left[ \psi_m^{(j)}(X_t)X_t^{(m)} - \frac{1}{T} \sum_{u=1}^{T} \psi_m^{(j)}(X_u)X_u^{(m)} \right] k \left( \frac{s-t}{b_T} \right) \\
= \left( 1 - \frac{1}{\hat{\sigma}^2(m)} \right) \frac{1}{T} \sum_{s,t} Y_s^{(i)} \psi_m^{(j)}(X_t)X_t^{(m)} k \left( \frac{s-t}{b_T} \right) \\
- \left( 1 - \frac{1}{\hat{\sigma}^2(m)} \right) \frac{1}{T} \sum_{s,t} \psi_m^{(j)}(X_s)X_s^{(m)} k \left( \frac{s-t}{b_T} \right) \frac{1}{T} \sum_{u=1}^{T} X_u^{(i)} \\
+ \left( 1 - \frac{1}{\hat{\sigma}^2(m)} \right) \frac{1}{T} \sum_{s,t} k \left( \frac{s-t}{b_T} \right) \frac{1}{T} \sum_{u=1}^{T} \psi_m^{(j)}(X_u)X_u^{(m)} \frac{T}{T} \sum_{v=1}^{T} X_v^{(j)} = I + II + III + IV
\]

The summands \( II, III \) and \( IV \) contain at least one arithmetic mean converging against 0. Therefore we only have to look at \( I \):

\[
E(|I|) = E \left\{ \left( 1 - \frac{1}{\hat{\sigma}^2(m)} \right) \sum_{s,t=1}^{T} \left[ Y_s^{(i)} - \frac{1}{T} \sum_{u=1}^{T} Y_u^{(i)} \right] \left[ \psi_m^{(j)}(X_t)X_t^{(m)} - \frac{1}{T} \sum_{u=1}^{T} \psi_m^{(j)}(X_u)X_u^{(m)} \right] k \left( \frac{s-t}{b_T} \right) \right\} \\
\leq \frac{1}{T^{\frac{1}{2}-\epsilon}} \frac{1}{T} \left\{ \sum_{s,t=1}^{T} E \left| Y_s^{(i)} \psi_m^{(j)}(X_t)X_t^{(m)} \right| \right\} \\
\leq \frac{1}{T^{\frac{1}{2}-\epsilon}} 2 \sum_{k=0}^{\infty} a_k \to 0
\]

which completes the proof. \(\square\)