Detecting Smooth Changes in Locally Stationary Processes

Michael Vogt
University of Cambridge

Holger Dette
Ruhr-Universität Bochum

October 3, 2013

Abstract

In a wide range of applications, the stochastic properties of the observed time series change over time. It is often realistic to assume that the properties are approximately the same over short time periods and then gradually start to vary. This behaviour is well modelled by locally stationary processes. In this paper, we investigate the question how to estimate time spans where the stochastic features of a locally stationary time series are the same. We set up a general method which allows to deal with a wide variety of features including the mean, covariances, higher moments and the distribution of the time series under consideration. In the theoretical part of the paper, we derive the asymptotic properties of our estimation method. In addition, we examine its finite sample performance by means of a simulation study and illustrate the methodology by an application to financial data.

Key words: Local stationarity; empirical processes; measure of time-variation.
AMS 2010 subject classifications: 62G05, 62G20, 62M10.

1 Introduction

In many applications, the stochastic properties of the observed time series such as the mean, the variance or the distribution change over time. Examples can be found in a wide range of application areas: In climatology, temperature data frequently exhibit a trending behaviour, i.e., their mean varies over time. In neuroscience, EEG and MEG signals change their characteristics depending on the state of the patient. Finally, financial time series are often characterized by a time-varying volatility level.

One way to model the time-varying features of a time series is to use change point methods. In this approach, the time series is split into segments. Its stochastic properties are assumed to be the same within segments but are allowed to vary across them. Estimating the change point between the segments is a well studied problem which has been analyzed in a variety of settings; see for example Hinkley [14], Worsley [29] and

Carlstein [3] among many others. The change point framework is apt to model abrupt changes over time. However, in many applications, the observed time series changes its behaviour gradually rather than abruptly. In such cases, it is not very appropriate to partition it into segments where the stochastic properties are the same. It is more realistic to assume that the stochastic features of interest are (approximately) stable within a certain time span and then gradually start to vary. This behaviour is well captured by locally stationary processes as introduced in Dahlhaus [5, 6].

An important issue in a locally stationary framework is to identify time periods where the stochastic features of interest are (approximately) the same. More specifically, suppose we observe a sample $\{X_{t,T} : t = 1, \dots, T\}$ from a locally stationary process and are interested in its behaviour around the time point t^*, or equivalently, around the rescaled time point $u^* = t^*/T$. Moreover, assume that the stochastic properties of interest are the same within the time interval $[u_0, u_1]$ around the point u^* but gradually start to vary outside it. Knowledge of the interval $[u_0, u_1]$ is crucial in many situations. As an example, suppose we want to forecast a specific feature of the process $\{X_{t,T}\}$ such as its volatility. In this case, $u^* = 1$ and our interest focuses on the interval $[u_0, 1]$. If we knew the interval $[u_0, 1]$, we could estimate the feature of interest from the data in this interval and base our forecasts on the resulting estimate. Hence, identifying the time span $[u_0, 1]$ is an important first step when performing forecasts.

The main goal of this paper is to develop a statistical procedure to estimate the interval $[u_0, u_1]$ around the time point u^*. We tackle this problem within a locally stationary framework which is formally introduced in Sections 2 and 3. Rather than restricting attention to a specific stochastic property, we set up a general procedure which allows to deal with a wide variety of features including the mean, covariances, higher moments and the distribution of the process under consideration. To keep the exposition as simple as possible, we stick to the case $u^* = 1$ throughout the paper, i.e., we focus on estimating the lower end point u_0 of the interval $[u_0, 1]$. Our methods and results can however be easily extended to the case of an arbitrary point $u^* \in [0, 1]$ at the cost of a more involved notation. Remark 5.3 in Section 5 outlines how to achieve this.

The problem of estimating the point u_0 can be approached in different ways. In some settings, it is possible to make use of change point methods. Consider for example the time-varying mean model $X_{t,T} = \mu(\frac{t}{T}) + \varepsilon_t$, where the mean function μ takes a constant value on $[u_0, 1]$ and smoothly deviates from it prior to u_0. Under certain conditions, the function μ has a break point in the k-th derivative. When k is known, u_0 can be estimated by standard methods to detect a break in the k-th derivative; see e.g. Müller [21] and Wu & Chu [30]. In the vast majority of applications, however, the order k is unknown. Moreover, an approach based on derivative estimation only seems reasonable in fairly simple model settings. When concerned with more complicated

stochastic features like the distribution function of the process, we would ideally like to do without estimating intricate higher-order derivatives.

Another possible way is to work with testing ideas. In recent years, a variety of procedures have been proposed to test whether the covariance structure of a locally stationary process is stable over time. Most of these tests for second-order stationarity are based on comparing a local spectral density with a global version; see Paparoditis [22, 23], Dette et al. [8] and Preuß et al. [24] among others. A Portmanteau-type test has been constructed in Dwivedi & Subba Rao [10]; a Wavelet-based test can be found in von Sachs & Neumann [28]. To estimate the time point u_0, one may try to sequentially apply such testing methods.

A sequential procedure roughly works as follows: Given a suitable test statistic, we first perform the test on a small interval $[u, 1]$ with u being close to 1 and then successively shift the point u further and further away from 1 until the test rejects. Such a sequential method has for example been applied to a simple time-varying volatility model in Chen et al. [4]. Importantly, the critical values are different in each step of such a procedure. In particular, we cannot just use the critical values from the static version of the test; we rather have to calculate a whole sequence of critical values. This may become quite cumbersome and involved when the test statistic has a complicated form. Even in the static testing case, computing the critical values is often an issue and bootstrap procedures are required to achieve a reasonable approximation.

In this paper, we introduce an alternative method to estimate the point u_0 which avoids the disadvantages and problems outlined above. Our approach is based on a function $\mathcal{D} : [0, 1] \to \mathbb{R}_{\geq 0}$, where $\mathcal{D}(u)$ measures the amount of time-variation in the stochastic features of interest within the interval $[u, 1]$. By construction, $\mathcal{D}(u) = 0$ if there is no time-variation on $[u, 1]$ and $\mathcal{D}(u) > 0$ if there is some time-variation involved. Since $\mathcal{D}$ is not observed, we replace it by an estimator $\hat{\mathcal{D}}_T$. Section 4 gives a detailed account of how to construct the measure of time-variation $\mathcal{D}$ and its estimator $\hat{\mathcal{D}}_T$. The time point u_0 can now be characterized as the point where the measure $\mathcal{D}$ starts to deviate from zero. Since $\mathcal{D}$ generally deviates from zero in a smooth fashion, we transform it together with its estimator to behave approximately like a function with a jump at u_0. This transformed measure in turn is used to set up a criterion function which is approximately minimized at u_0. The minimizer $\hat{u}_0$ serves as our estimator of u_0. It is worthwhile mentioning that the proposed procedure does not require the choice of any smoothing parameter for curve estimation. Section 5 describes in detail how the measure $\mathcal{D}$ and its estimator $\hat{\mathcal{D}}_T$ are used to construct $\hat{u}_0$.

In Section 6, we discuss the asymptotic properties of our estimation method. In particular, we derive the convergence rate of the new estimator $\hat{u}_0$. The smoother the stochastic properties of interest vary around u_0, the slower the convergence rate of the estimator turns out to be. This reflects the intuition that the smoother the

time-variation, the harder it is to detect the point u_0. Section 7 discusses how to implement our procedure in practice to achieve a good finite sample behaviour. In Section 8, we perform a simulation study which confirms that the procedure works well in small samples. In addition, we illustrate the method by a real data example in Section 9 where it is applied to a sample of financial return and volatility data. Finally, all proofs and technical details are deferred to an appendix.

2 Local Stationarity

Throughout the paper, we assume that the sample of observations $\{X_{t,T} : t = 1, \dots, T\}$ comes from a locally stationary process of d-dimensional variables $X_{t,T}$. Specifically, we work with the following concept of local stationarity, which was introduced in Vogt [27].

Definition 2.1. *The array $\{X_{t,T} : t = 1, \dots, T\}_{T=1}^{\infty}$ is called a locally stationary process if for each rescaled time point $u \in [0,1]$, there exists a strictly stationary process $\{X_t(u) : t \in \mathbb{Z}\}$ with the property that*

$$\left\| X_{t,T} - X_t(u) \right\| \leq \left(\left| \frac{t}{T} - u \right| + \frac{1}{T} \right) U_{t,T}(u) \quad a.s.$$

Here, $\|\cdot\|$ denotes a norm on $\mathbb{R}^d$ and $\{U_{t,T}(u) : t = 1, \dots, T\}_{T=1}^{\infty}$ is an array of positive random variables whose ρ-th moment is uniformly bounded for some $\rho > 0$, that is, $\mathbb{E}[U_{t,T}^{\rho}(u)] \leq C < \infty$ for some fixed constant C.

Our definition of local stationarity is similar to those in Dahlhaus & Subba Rao [7] and Koo & Linton [15] for example. The intuitive idea behind these definitions is that a process is locally stationary if it behaves approximately stationary locally in time, i.e., over short time periods. This idea is turned into a rigorous concept by requiring that locally around each rescaled time point u, the process $\{X_{t,T}\}$ can be approximated by a stationary process $\{X_t(u)\}$ in a stochastic sense.

There is a wide range of time series processes which are locally stationary in the sense of Definition 2.1. In particular, many processes with time-varying parameters can be locally approximated by a stationary process provided that the parameters are smoothly changing over time. This is fairly straightforward to show for linear models like time-varying MA or AR processes. However, it may also be verified for more complicated models like time-varying ARCH and GARCH processes; see for example Dahlhaus & Subba Rao [7] and Subba Rao [25]. It is also possible to derive local stationarity for nonparametric models. Vogt [27] for example has shown that nonparametric AR processes with a time-varying regression function are locally stationary under appropriate conditions.

3 Model Setting

Let $\lambda_{t,T}$ be some time-varying feature of the locally stationary process $\{X_{t,T}\}$ such as the mean $\mathbb{E}[X_{t,T}]$, the variance $\mathrm{Var}(X_{t,T})$ or the distribution function $F_{t,T}(\cdot) = \mathbb{P}(X_{t,T} \leq \cdot\ )$, where for vectors the inequality sign is understood componentwise. Generally speaking, we allow for any feature $\lambda_{t,T}$ which fulfills the following property:

(P_λ) $\lambda_{t,T}$ is uniquely determined by the set of moments $\{\mathbb{E}[f(X_{t,T})] : f \in \mathcal{F}\}$, where $\mathcal{F}$ is a family of measurable functions $f : \mathbb{R}^d \to \mathbb{R}$.

We illustrate the property (P_λ) by some examples:

Example I. Let $\lambda_{t,T}$ be the mean $\mu_{t,T} = \mathbb{E}[X_{t,T}]$ of a univariate locally stationary process $\{X_{t,T}\}$. Then the corresponding family of functions is simply $\mathcal{F} = \{\mathrm{id}\}$, since the mean $\mu_{t,T}$ can be written as $\mathbb{E}[\mathrm{id}(X_{t,T})]$.

Example II. Let $\lambda_{t,T}$ be the vector of the first p autocovariances of a univariate locally stationary process $\{Y_{t,T}\}$ whose elements $Y_{t,T}$ are centred for simplicity. Specifically, define $\gamma_{\ell,t,T} = \mathrm{Cov}(Y_{t,T}, Y_{t-\ell,T})$ to be the ℓ-th order autocovariance and set $\lambda_{t,T} = (\gamma_{0,t,T}, \ldots, \gamma_{p,t,T})^\intercal$. To handle this case, we regard the data as coming from the $(p+1)$-dimensional process $\{X_{t,T}\}$ with $X_{t,T} = (Y_{t,T}, Y_{t-1,T}, \ldots, Y_{t-p,T})^\intercal$. We now define functions $f_\ell : \mathbb{R}^{p+1} \to \mathbb{R}$ for $0 \leq \ell \leq p$ by $f_\ell(x) = x_0 x_\ell$, where $x = (x_0, \ldots, x_p)^\intercal$. As $\mathbb{E}[f_\ell(X_{t,T})] = \mathbb{E}[Y_{t,T}Y_{t-\ell,T}] = \gamma_{\ell,t,T}$, we obtain that $\mathcal{F} = \{f_0, \ldots, f_p\}$ in this setting.

Example III. As in the previous example, let $\{Y_{t,T}\}$ be a real-valued locally stationary process and write $X_{t,T} = (Y_{t,T}, Y_{t-1,T}, \ldots, Y_{t-p,T})^\intercal$. We now set $\lambda_{t,T}$ to be the distribution function $F_{t,T}$ of the variable $X_{t,T}$, or put differently, the joint distribution function of the variables $(Y_{t,T}, Y_{t-1,T}, \ldots, Y_{t-p,T})$. Define $I(y \leq x) = \prod_{\ell=0}^{p} 1(y_\ell \leq x_\ell)$ for vectors $x = (x_0, \ldots, x_p)^\intercal$ and $y = (y_0, \ldots, y_p)^\intercal$, where $1(\cdot)$ denotes the indicator function. Noting that $\mathbb{E}[I(X_{t,T} \leq x)] = F_{t,T}(x)$, we obtain that $\mathcal{F} = \{I(\ \cdot \leq x) : x \in \mathbb{R}^{p+1}\}$.

Generally speaking, (P_λ) is a fairly weak condition which is satisfied by a wide range of stochastic features. Indeed, it essentially allows us to deal with any feature that can be expressed in terms of a set of moments.
Let us now define λ_u to be the stochastic feature of the approximating process $\{X_t(u)\}$ which corresponds to $\lambda_{t,T}$. This means that λ_u is fully characterized by the set of moments $\{\mathbb{E}[f(X_t(u))] : f \in \mathcal{F}\}$. Throughout the paper, we assume that

$$\sup_{f \in \mathcal{F}} \big|\mathbb{E}[f(X_{t,T})] - \mathbb{E}[f(X_t(u))]\big| \leq C\Big(\Big|\frac{t}{T} - u\Big| + \frac{1}{T}\Big), \tag{3.1}$$

which is implied by the high-order condition (C4) in Subsection 6.1. In a wide range of cases, the inequality (3.1) boils down to mild moment conditions on the random variables $X_{t,T}$, $X_t(u)$ and $U_{t,T}(u)$. This in particular holds true in the settings from

Examples I–III as shown in Subsection 6.4. The inequality (3.1) essentially says that $\lambda_{t,T}$ and λ_u are close to each other locally in time. In the time-varying mean setting from Example I, it can be expressed as

$$\big|\mu_{t,T} - \mu(u)\big| \leq C\Big(\Big|\frac{t}{T} - u\Big| + \frac{1}{T}\Big)$$

with $\mu(u)$ being the mean of $X_t(u)$. Similarly, in Example II, it is equivalent to the statement

$$\big\|(\gamma_{0,t,T}, \ldots, \gamma_{p,t,T})^\intercal - (\gamma_0(u), \ldots, \gamma_p(u))^\intercal\big\| \leq C\Big(\Big|\frac{t}{T} - u\Big| + \frac{1}{T}\Big),$$

where $\gamma_\ell(u) = \mathrm{Cov}(Y_t(u), Y_{t-\ell}(u))$ and $\|\cdot\|$ is some norm on $\mathbb{R}^{p+1}$. Finally, in Example III, it says that

$$\sup_{x\in\mathbb{R}^d}\big|F_{t,T}(x) - F(u,x)\big| \leq C\Big(\Big|\frac{t}{T} - u\Big| + \frac{1}{T}\Big),$$

where $F(u,\cdot)$ denotes the distribution function of the variables $X_t(u)$. Hence, if (3.1) holds true, then the feature $\lambda_{t,T}$ converges to λ_u locally in time. In particular, time-variation in $\lambda_{t,T}$ is asymptotically equivalent to time-variation in λ_u. To detect whether the stochastic feature $\lambda_{t,T}$ of interest changes over time, we may thus check for variations in the approximating quantity λ_u.

Our estimation problem can now be formulated as follows: Assume that λ_u does not vary on the rescaled time interval $[u_0, 1]$ but is time-varying prior to u_0. Our aim is to estimate the time point u_0 where λ_u starts to change over time.

4 A Measure of Time-Variation

In this section, we construct a function $\mathcal{D} : [0,1] \to \mathbb{R}_{\geq 0}$ which captures time-variations in the stochastic feature λ_w of interest and explain how to estimate it. The function $\mathcal{D}$ is assumed to have the property

$$(P_{\mathcal{D}}) \qquad \mathcal{D}(u)\begin{cases} = 0 & \text{if } \lambda_w \text{ does not vary on } [u,1] \\ > 0 & \text{if } \lambda_w \text{ varies on } [u,1] \end{cases}$$

and is called a measure of time-variation. In what follows, we describe how to set up such a measure for a generic stochastic feature that satisfies (P_λ) and then reconsider the features from Examples I–III.

Our construction is based on the following idea: By the property (P_λ), the feature λ_w is fully characterized by the values $\mathbb{E}[f(X_t(w))]$ with f running over all functions in the family $\mathcal{F}$. This implies that time-variation in λ_w is equivalent to time-variation in the moments $\mathbb{E}[f(X_t(w))]$ for some $f \in \mathcal{F}$. To detect changes in λ_w over time, we may

thus set up a function which captures time-variations in the quantities $\mathbb{E}[f(X_t(w))]$ for any $f \in \mathcal{F}$. This idea underlies the following definition:

$$\mathcal{D}(u) = \sup_{f\in\mathcal{F}, v\in[u,1]} \big|D(u,v,f)\big|, \tag{4.1}$$

where

$$D(u,v,f) = \int_v^1 \mathbb{E}[f(X_t(w))]dw - \Big(\frac{1-v}{1-u}\Big)\int_u^1 \mathbb{E}[f(X_t(w))]dw. \tag{4.2}$$

The function $\mathcal{D}$ has the following property: If the moment function $\mathbb{E}[f(X_t(\cdot))]$ is constant on the interval $[u,1]$, then the average $\int_v^1 \mathbb{E}[f(X_t(w))]dw/(1-v)$ takes the same value at all time points $v \in [u,1]$. From this, it immediately follows that $\mathcal{D}(u,v,f) = 0$ for any $v \in [u,1]$. Hence, if the function $\mathbb{E}[f(X_t(\cdot))]$ is constant on $[u,1]$ for any $f \in \mathcal{F}$, then $\mathcal{D}(u) = 0$. If $\mathbb{E}[f(X_t(\cdot))]$ varies on $[u,1]$ for some f in contrast, then the average $\int_v^1 \mathbb{E}[f(X_t(w))]dw/(1-v)$ varies on this time span as well. This is ensured by the fact that $\mathbb{E}[f(X_t(\cdot))]$ is a Lipschitz continuous function of rescaled time, i.e., $|\mathbb{E}[f(X_t(u))] - \mathbb{E}[f(X_t(v))]| \leq C|u-v|$ for any $u,v \in [0,1]$, which is an immediate consequence of (3.1). We thus obtain that $D(u,v,f) > 0$ for some $v \in [u,1]$, which in turn yields that $\mathcal{D}(u) > 0$. As a result, the function $\mathcal{D}$ satisfies $(P_{\mathcal{D}})$.

Since the feature λ_w is constant on $[u_0,1]$ but varies before u_0, the property $(P_{\mathcal{D}})$ immediately implies that

$$\mathcal{D}(u)\begin{cases} = 0 & \text{for } u \geq u_0 \\ > 0 & \text{for } u < u_0. \end{cases}$$

The point u_0 is thus characterized as the time point where the measure of time-variation starts to deviate from zero. Importantly, the measure $\mathcal{D}$ does not have a jump at u_0 in general, but smoothly deviates from zero at this point. Its degree of smoothness depends on how smoothly the moments $\mathbb{E}[f(X_t(w))]$ vary over time, or put differently, on how smoothly the feature λ_w varies over time. In particular, the smoother the time-variation in λ_w, the smoother the function $\mathcal{D}$.

Since our measure of time-variation depends on the unobserved moment functions $\mathbb{E}[f(X_t(\cdot))]$, we cannot work with it directly but have to replace it by an estimator. This can be achieved as follows: The integral $\int_v^1 \mathbb{E}[f(X_t(w))]dw$ can be regarded as an average of the moments $\mathbb{E}[f(X_t(w))]$, where all time points from v to 1 are taken into account. This suggests to estimate it by a sample average of the form $T^{-1}\sum_{t=\lceil vT+1\rceil}^{T} f(X_{t,T})$. Following this idea, an estimator of $\mathcal{D}(u)$ is given by

$$\hat{\mathcal{D}}_T(u) = \sup_{f\in\mathcal{F}, v\in[u,1]} \big|\hat{D}_T(u,v,f)\big|,$$

where we set

$$\hat{D}_T(u,v,f) = \frac{1}{T}\sum_{t=\lceil vT+1\rceil}^{T} f(X_{t,T}) - \Big(\frac{1-v}{1-u}\Big)\frac{1}{T}\sum_{t=\lceil uT+1\rceil}^{T} f(X_{t,T}).$$

It is important to note that the statistic $\hat{\mathcal{D}}_T$ is completely free of bandwidth parameters, even though we have not imposed any parametric restrictions on the time-varying feature λ_w. This is possible for the following reason: To estimate the moments $\mathbb{E}[f(X_t(w))]$ at a fixed time point w, we would require some bandwidth to localize in time. However, we are not interested in the moments at a fixed point w but rather want to estimate averages of them over some time spans $[u,1]$ and $[v,1]$. This can be achieved by using partial sum processes, i.e., by simply forming sample averages of the observations that lie in the time spans $[u,1]$ and $[v,1]$, respectively.
We now apply the general definitions from above to the settings from Examples I–III.

Example I. In the time-varying mean setting, the function family $\mathcal{F}$ only consists of the identity function. Our measure of time-variation in the mean is thus given by

$$\mathcal{D}_\mu(u) = \sup_{v\in[u,1]} \big|D_\mu(u,v)\big|$$

together with

$$D_\mu(u,v) = \int_v^1 \mu(w)dw - \Big(\frac{1-v}{1-u}\Big)\int_u^1 \mu(w)dw,$$

where $\mu(w) = \mathbb{E}[X_t(w)]$. This quantity can be estimated by

$$\hat{\mathcal{D}}_{\mu,T}(u) = \sup_{v\in[u,1]} \big|\hat{D}_{\mu,T}(u,v)\big|,$$

where we set

$$\hat{D}_{\mu,T}(u,v) = \frac{1}{T}\sum_{t=\lceil vT+1\rceil}^{T} X_{t,T} - \Big(\frac{1-v}{1-u}\Big)\,\frac{1}{T}\sum_{t=\lceil uT+1\rceil}^{T} X_{t,T}.$$

Example II. Let the feature of interest be the vector of the first p autocovariances of the process at hand. In this case, the family $\mathcal{F}$ consists of the $(p+1)$ functions $f_1,\ldots,f_p$. Our measure of time-variation may thus be written as

$$\mathcal{D}_\gamma(u) = \max_{0\le\ell\le p}\sup_{v\in[u,1]} \big|D_\gamma(u,v,\ell)\big|$$

together with

$$D_\gamma(u,v,\ell) = \int_v^1 \gamma_\ell(w)dw - \Big(\frac{1-v}{1-u}\Big)\int_u^1 \gamma_\ell(w)dw,$$

where $\gamma_\ell(w) = \mathrm{Cov}(Y_t(w), Y_{t-\ell}(w))$. The overall measure $\mathcal{D}_\gamma(u)$ can be regarded as aggregating the individual measures $\sup_{v\in[u,1]}|D_\gamma(u,v,\ell)|$ each of which captures time-variations in the autocovariance function γ_ℓ of a different order ℓ. The estimator of $\mathcal{D}_\gamma(u)$ is given by

$$\hat{\mathcal{D}}_{\gamma,T}(u) = \max_{0\le\ell\le p}\sup_{v\in[u,1]} \big|\hat{D}_{\gamma,T}(u,v,\ell)\big|,$$

where we let

$$\hat{D}_{\gamma,T}(u,v,\ell) = \frac{1}{T}\sum_{t=\lceil vT+1\rceil}^{T} Y_{t,T}Y_{t-\ell,T} - \Big(\frac{1-v}{1-u}\Big)\,\frac{1}{T}\sum_{t=\lceil uT+1\rceil}^{T} Y_{t,T}Y_{t-\ell,T}.$$

Importantly, the statistic $\hat{\mathcal{D}}_{\gamma,T}$ can be used to detect time-variations in the parameters of a large class of linear locally stationary models. As an example, consider the time-varying AR(p) process

$$Y_{t,T} = \sum_{\ell=1}^{p} a_\ell\Big(\frac{t}{T}\Big)Y_{t-\ell,T} + \varepsilon_t \tag{4.3}$$

with i.i.d. residuals ε_t. Time-variation in the parameter functions $a_1,\ldots,a_p$ is equivalent to time-variation in the autocovariance structure of the process $\{Y_{t,T}\}$ up to the p-th order. If the underlying model has the autoregressive structure (4.3), we can thus use the above statistic to measure the amount of time-variation in the parameter functions.

Example III. We finally examine the case where the feature of interest is the distribution function of the process. In this situation, $\mathcal{F} = \{I(\cdot \le x) : x \in \mathbb{R}^{p+1}\}$ with $I(\cdot \le x)$ being a product of indicator functions as defined earlier on. As the family $\mathcal{F}$ can be identified with the set of points $x \in \mathbb{R}^{p+1}$, the measure of time-variation turns out to be

$$\mathcal{D}_F(u) = \sup_{x\in\mathbb{R}^{p+1},v\in[u,1]} \big|D_F(u,v,x)\big| \tag{4.4}$$

together with

$$D_F(u,v,x) = \int_v^1 F(w,x)dw - \Big(\frac{1-v}{1-u}\Big)\int_u^1 F(w,x)dw,$$

where $F(w,\cdot)$ is the distribution function of $X_t(w)$. The terms $\sup_{v\in[u,1]}|D_F(u,v,x)|$ measure the amount of time-variation in the distribution function evaluated at a fixed point x. The measure $\mathcal{D}_F(u)$ is obtained by aggregating these expressions, specifically by taking the supremum over all points x. To estimate $\mathcal{D}_F(u)$, we use the statistic

$$\hat{\mathcal{D}}_{F,T}(u) = \sup_{x\in\mathbb{R}^{p+1},v\in[u,1]} \big|\hat{D}_{F,T}(u,v,x)\big|,$$

where

$$\hat{D}_{F,T}(u,v,x) = \frac{1}{T}\sum_{t=\lceil vT+1\rceil}^{T} I(X_{t,T}\le x) - \Big(\frac{1-v}{1-u}\Big)\,\frac{1}{T}\sum_{t=\lceil uT+1\rceil}^{T} I(X_{t,T}\le x),$$

compares averages of the binary variables $I(X_{t,T}\le x)$.

5 Estimating the Smooth Change Point u_0

We now describe how to use our measure of time-variation to estimate the point u_0. Our estimation method is based on the observation that

$$\sqrt{T}\mathcal{D}(u)\begin{cases}=0 & \text{for } u \geq u_0\\ \to \infty & \text{for } u < u_0\end{cases}$$

as $T \to \infty$. As the statistic $\hat{\mathcal{D}}_T$ estimates the measure $\mathcal{D}$, its scaled version $\sqrt{T}\hat{\mathcal{D}}_T$ should exhibit a similar behaviour. Indeed, as we will see later on, it holds that

$$\sqrt{T}\hat{\mathcal{D}}_T(u)\begin{cases}=O_p(1) & \text{for } u \geq u_0\\ \xrightarrow{P} \infty & \text{for } u < u_0.\end{cases}$$

The main idea is to exploit this dichotomous behaviour of the process $\sqrt{T}\hat{\mathcal{D}}_T$.
To set up the estimation procedure, we proceed in two steps. First we transform the statistic $\sqrt{T}\hat{\mathcal{D}}_T$ to behave approximately like a function that has a jump at u_0. To achieve this, define $\Phi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ to be a strictly increasing function which is normalized to satisfy $\Phi(0) = 0$ and $\lim_{x\to\infty} \Phi(x) = 1$. Moreover, let $\{\rho_T\}$ be a sequence of positive constants which slowly converges to zero, in particular much slower than $O(T^{-1/2})$. Premultiplying $\sqrt{T}\hat{\mathcal{D}}_T(u)$ with the shrinkage factor ρ_T and then applying the function Φ yields the quantity $\hat{q}_T(u) = \Phi(\rho_T\sqrt{T}\hat{\mathcal{D}}_T(u))$ which has the property that

$$\hat{q}_T(u) \xrightarrow{P} \begin{cases}0 & \text{for } u \geq u_0\\ 1 & \text{for } u < u_0.\end{cases}$$

Hence, $\hat{q}_T(\cdot)$ behaves approximately like the step function $1(\cdot < u_0)$ which has a jump at the point u_0.
In the second step, we use the quantity $\hat{q}_T$ to construct a criterion function which is minimized approximately at u_0. To do so, define

$$\hat{Q}_T(u) = u + (1-u)\hat{q}_T(u).$$

The function value $\hat{Q}_T(u)$ measures the area below the curve $\zeta_u(w) = 1(w \leq u) + \hat{q}_T(u)1(w > u)$, which takes the value 1 at points $w \leq u$ and the value $\hat{q}_T(u)$ at points $w > u$. We can thus write $\hat{Q}_T(u) = \int_0^1 \zeta_u(w)dw$, which is graphically illustrated in Figure 1. Since $\hat{q}_T(\cdot)$ approximately behaves like the indicator function $1(\cdot < u_0)$, the area $\hat{Q}_T(u)$ should be minimized at a point close to u_0. Indeed, the asymptotic counterpart $Q(u) = u + (1-u)1(u < u_0)$ of $\hat{Q}_T(u)$ is easily seen to take its minimum exactly at u_0. These considerations suggest to estimate u_0 by

$$\hat{u}_0 := \underset{u\in[0,1]}{\operatorname{argmin}}\ \hat{Q}_T(u).$$

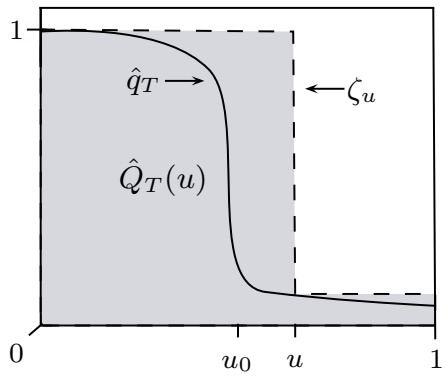

Figure 1: Graphical illustration of the criterion function $\hat{Q}_T$. Its value $\hat{Q}_T(u)$ at the time point u is equal to the grey shaded area.

The estimator $\hat{u}_0$ implicitly depends on the choice of the transformation function Φ and the shrinkage factor ρ_T. As demonstrated in Section 6, our asymptotic theory allows for a wide range of different choices. In particular, any sequence $\{\rho_T\}$ that converges more slowly to zero than $O(T^{-1/2})$ and any strictly increasing and Lipschitz continuous function Φ with $\Phi(0) = 0$ and $\lim_{x\to\infty} \Phi(x) = 1$ will do. Moreover, it goes without saying that the choice of Φ and ρ_T influences the finite sample behaviour of $\hat{u}_0$. In Section 7, we give some heuristic discussion of this issue and provide a natural data driven choice of Φ and ρ_T which yields a good finite sample performance of the estimator $\hat{u}_0$.

Remark 5.1. Our estimation procedure may alternatively be based on a criterion function of the form

$$\hat{Q}_{L_p,T}(u) = \int_0^u \big[1 - \hat{q}_T(w)\big]^p dw + \int_u^1 \hat{q}_T(w)^p dw$$

for some integer $p \geq 1$. This function measures the L_p-distance between $\hat{q}_T(\cdot)$ and the indicator $1(\cdot < u)$. Similarly as before, its asymptotic counterpart $Q_{L_p}(u) = \int_0^u [1 - 1(w \leq u_0)]^p dw + \int_u^1 1(w \leq u_0)^p dw$ is minimized exactly at u_0, suggesting that $\hat{Q}_{L_p,T}(u)$ should take its minimum close to u_0 as well. The minimizer of $\hat{Q}_{L_p,T}(u)$ may thus be used as an alternative estimator of u_0. It is worth noting that the estimators resulting from L_p-criteria with different p are essentially the same. This is due to the fact that asymptotically $Q_{L_p} \equiv Q_{L_1}$ for any $p \geq 1$, meaning that the criterion functions asymptotically coincide for all p.
Even though the criterion $\hat{Q}_{L_p,T}$ may appear to be an interesting alternative to $\hat{Q}_T$, it has an important drawback: Since the statistic $\hat{q}_T(\cdot)$ approximates the indicator function $1(\cdot < u_0)$, it sharply increases as soon as $u < u_0$. Hence, the point u_0 should be close to the point where $\hat{q}_T$ shots up towards a value of one. However, the L_p-criterion is not minimized at this point but at a smaller time point where $\hat{q}_T$ has already grown sufficiently large. This produces an estimator which has a strong

downward bias in small samples and thus notoriously underestimates the true value u_0. One way to obtain estimates with a smaller bias is to modify the L_p-criterion: Since $\int_u^1 \hat{q}_T(w)^p dw \approx (1-u)\hat{q}_T(u)^p$ for $u \geq u_0$, we may replace the integral by the term $(1-u)\hat{q}_T(u)^p$ which picks up an increase in the statistic $\hat{q}_T(u)$ much more strongly than the integral. A modified version of the L_p-criterion can thus be defined as

$$\hat{Q}_{L_p,T}(u) = \int_0^u \big[1 - \hat{q}_T(w)\big]^p dw + (1-u)\hat{q}_T(u)^p.$$

Simulations suggest that the minimizer of this modified criterion performs similarly well in small samples as the estimator $\hat{u}_0$. We thus recommend to work either with the criterion function $\hat{Q}_T$ or with the modified L_p-criterion. In the remainder of the paper, we restrict attention to $\hat{Q}_T$ to keep the exposition as clear as possible.

Remark 5.2. Our estimation method relies on a similar idea as the procedure proposed in Mallik et al. [19] and [18]. There, a p-value based method is suggested to estimate the point u_0 in the time-varying mean setting $X_{t,T} = \mu(\frac{t}{T}) + \varepsilon_t$ with μ being constant on $[u_0, 1]$. Whereas the approach of Mallik et al. is tailored to this simple mean setting, our method addresses the problem from a general perspective and allows to deal with a wide variety of stochastic features. For this reason, the technical arguments of our approach are very different from theirs which heavily draw on the structure of the time-varying mean setting. It is also worth noting that their procedure is based on a nonparametric kernel estimator of the function μ and thus requires to specify a bandwidth. The method proposed here in contrast is free of bandwidth parameters.

Remark 5.3. As noted in the Introduction, our estimation problem is a special case of the following issue: Let $u^* \in [0,1]$ and assume that λ_w is constant in the time region $[u_0, u_1]$ around u^* but varies outside it. We aim to estimate the interval $[u_0, u_1]$, or put differently, the two points u_0 and u_1.
To tackle this more general problem, we modify our estimation method in a straightforward way. First of all, we generalize our measure of time-variation. This has been designed to detect time-variations within intervals of the form $[u, 1]$. It can be easily extended to measure time-variations on a general interval $[\underline{u}, \overline{u}]$ with $0 \leq \underline{u} \leq \overline{u} \leq 1$. In particular, we may define

$$\mathcal{D}(\underline{u}, \overline{u}) = \sup_{f \in \mathcal{F}, v \in [\underline{u}, \overline{u}]} \big| D(\underline{u}, \overline{u}, v, f) \big|$$

along with

$$D(\underline{u}, \overline{u}, v, f) = \int_v^{\overline{u}} \mathbb{E}[f(X_t(w))] dw - \Big(\frac{\overline{u} - v}{\overline{u} - \underline{u}}\Big) \int_{\underline{u}}^{\overline{u}} \mathbb{E}[f(X_t(w))] dw.$$

This function has the property that $\mathcal{D}(\underline{u}, \overline{u}) = 0$ if λ_w does not vary on $[\underline{u}, \overline{u}]$ and $\mathcal{D}(\underline{u}, \overline{u}) > 0$ if λ_w varies within this time span. An estimator $\hat{\mathcal{D}}_T(\underline{u}, \overline{u})$ can be constructed in exactly the same way as described in Section 4.
We next apply the construction steps from above to the generalized statistic $\sqrt{T}\hat{\mathcal{D}}_T(\underline{u}, \overline{u})$. In particular, we define the transformed statistic

$$\hat{q}_T(\underline{u}, \overline{u}) = \Phi\big(\rho_T \sqrt{T} \hat{\mathcal{D}}_T(\underline{u}, \overline{u})\big)$$

and introduce the criterion function

$$\hat{Q}_T(\underline{u}, \overline{u}) = (1 - (\overline{u} - \underline{u})) + (\overline{u} - \underline{u})\hat{q}_T(\underline{u}, \overline{u}).$$

Analogous considerations as above suggest that $\hat{Q}_T(\underline{u}, \overline{u})$ should be minimized close to the point (u_0, u_1). We can thus use

$$(\hat{u}_0, \hat{u}_1) = \underset{(\underline{u}, \overline{u}):\ \underline{u} \leq u^* \leq \overline{u}}{\operatorname{argmin}} \hat{Q}_T(\underline{u}, \overline{u})$$

as an estimator of (u_0, u_1).

6 Asymptotic Properties

We now examine the asymptotic properties of the estimation method. We first investigate the weak convergence behaviour of the statistic $\hat{\mathcal{D}}_T$ and then derive the convergence rate of the estimator $\hat{u}_0$. Since the proofs are very technical and involved, they are deferred to the Appendix. To state the results, we use the symbol $\ell_\infty(S)$ for the space of bounded functions $f : S \to \mathbb{R}$ endowed with the supremum norm and let $\rightsquigarrow$ denote weak convergence.

6.1 Assumptions

Throughout the paper, we make the following assumptions:

(C1) The process $\{X_{t,T}\}$ is locally stationary in the sense of Definition 2.1.

(C2) The process $\{X_{t,T}\}$ is strongly mixing with mixing coefficients $\alpha(k)$ satisfying $\alpha(k) \leq Ca^k$ for some positive constants C and $a < 1$.

(C3) Let $p \geq 4$ be an even natural number and endow the set $\mathcal{F}$ with some semimetric $d_{\mathcal{F}}$. $(\mathcal{F}, d_{\mathcal{F}})$ is separable, totally bounded and not too complex in the sense that its covering number $\mathcal{N}(w, \mathcal{F}, d_{\mathcal{F}})$ satisfies the condition

$$\int_0^1 \mathcal{N}(w, \mathcal{F}, d_{\mathcal{F}})^{1/p} dw < \infty.$$

Moreover, the set $\mathcal{F}$ has an envelope F (i.e. $|f| \leq F$ for all $f \in \mathcal{F}$) which satisfies $\mathbb{E}[F(X_{t,T})^{(1+\delta)p}] \leq C < \infty$ for some small $\delta > 0$ and a fixed constant C. Finally, for any pair of functions $f, f' \in \mathcal{F}$,

$$\mathbb{E}\Big[\Big|\frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f, f')}\Big|^{(1+\delta)p}\Big] \leq C < \infty.$$

(C4) For $k = 1, 2$ and all $f \in \mathcal{F}$, it holds that $\mathbb{E}[|f(X_{t,T}) - f(X_t(u))|^k] \leq C(|\frac{t}{T} - u| + \frac{1}{T})$ for some fixed constant C.

Condition (C2) stipulates that the array $\{X_{t,T}\}$ is strongly mixing. A wide variety of locally stationary processes can be shown to be mixing under appropriate conditions; see for example Fryzlewicz & Subba Rao [12] and Vogt [27]. To keep the structure of the proofs as clear as possible, we have assumed the mixing rates to decay exponentially fast. Alternatively, we could work with slower polynomial rates at the cost of a more involved notation in the proofs. The conditions (C3) and (C4) are formulated in a very general way and depend on the family $\mathcal{F}$ under consideration. For many choices of $\mathcal{F}$, they boil down to simple moment conditions on the variables $X_{t,T}$, $X_t(u)$ and $U_{t,T}(u)$. This will be seen later on in Subsection 6.4, where we revisit Examples I–III and verify (C3) and (C4) in these settings.

6.2 Weak Convergence of the Measure of Time-Variation

To start with, we investigate the asymptotic properties of the expression

$$\hat{H}_T(u, v, f) = \sqrt{T}\big(\hat{D}_T(u, v, f) - D(u, v, f)\big).$$

To do so, let $\Delta = \{(u, v) \in [0, 1]^2 : v \geq u\}$ and equip the space $\Delta \times \mathcal{F}$ with the natural semimetric $|u - u'| + |v - v'| + d_{\mathcal{F}}(f, f')$. In what follows, we regard $\hat{H}_T$ as a process that takes values in $\ell_\infty(\Delta \times \mathcal{F})$ and show that it weakly converges to a Gaussian process H with covariance structure

$$\begin{aligned}\mathrm{Cov}(H(u, v, f), H(u', v', f')) = {} & \frac{(1-v)(1-v')}{(1-u)(1-u')} \sum_{l=-\infty}^{\infty} \int_{\max\{u,u'\}}^{1} c_l(w)dw \\ & - \frac{1-v'}{1-u'} \sum_{l=-\infty}^{\infty} \int_{\max\{v,u'\}}^{1} c_l(w)dw \\ & - \frac{1-v}{1-u} \sum_{l=-\infty}^{\infty} \int_{\max\{u,v'\}}^{1} c_l(w)dw \\ & + \sum_{l=-\infty}^{\infty} \int_{\max\{v,v'\}}^{1} c_l(w)dw, \end{aligned} \tag{6.1}$$

where $c_l(w) = \mathrm{Cov}(f(X_0(w)), f'(X_l(w)))$. The following theorem gives a precise description of the weak convergence of $\hat{H}_T$.

Theorem 6.1. *Let (C1)–(C4) be satisfied. Then*

$$\hat{H}_T = \sqrt{T}\big[\hat{D}_T - D\big] \rightsquigarrow H$$

as a process in $\ell_\infty(\Delta \times \mathcal{F})$, where $\hat{D}_T$ and D are defined in Section 4 and H is a Gaussian process on $\Delta \times \mathcal{F}$ with covariance kernel (6.1)*.*

This result can be used to characterize the weak convergence behaviour of the process $\hat{\mathcal{H}}_T$ defined by

$$\begin{aligned} \hat{\mathcal{H}}_T(u) &= \sup_{f\in\mathcal{F}, v\in[u,1]} \big|\hat{H}_T(u,v,f)\big| \\ &= \sqrt{T} \sup_{f\in\mathcal{F}, v\in[u,1]} \big|\hat{D}_T(u,v,f) - D(u,v,f)\big|. \end{aligned} \tag{6.2}$$

In particular, we can derive the following statement.

Theorem 6.2. *Let (C1)–(C4) be satisfied. Then*

$$\hat{\mathcal{H}}_T \rightsquigarrow \mathcal{H}$$

as a process in $\ell_\infty([0,1])$, where $\mathcal{H}(u) = \sup_{f\in\mathcal{F}, v\in[u,1]} |H(u,v,f)|$.

6.3 Convergence of the Estimator $\hat{u}_0$

Let us now turn to the asymptotic behaviour of the estimator $\hat{u}_0$. We assume throughout that the transformation function Φ is Lipschitz continuous. To capture the amount of smoothness of the measure $\mathcal{D}$ at the point u_0, we suppose that

$$\frac{\mathcal{D}(u)}{(u_0 - u)^k} \to c_k > 0 \qquad \text{as } u \nearrow u_0 \tag{6.3}$$

for some number $k > 0$ and a constant $c_k > 0$. The larger k, the more smoothly the measure $\mathcal{D}$ deviates from zero at the point u_0. The next theorem specifies the convergence rate at which the estimator $\hat{u}_0$ approaches the point u_0.

Theorem 6.3. *Let (C1)–(C4) be satisfied and assume that $u_0 \in [0,1)$. Moreover, let $\rho_T \searrow 0$ with $\rho_T\sqrt{T} \to \infty$. Then*

$$\hat{u}_0 - u_0 = O_p(\gamma_T),$$

where $\gamma_T = \max\{\rho_T, (\rho_T\sqrt{T})^{-1/k}\}$ and k is defined in (6.3)*.*

As can be seen from the theorem, the convergence rate of $\hat{u}_0$ depends on the degree of smoothness k of the measure $\mathcal{D}$ in the point u_0. In particular, the smoother $\mathcal{D}$, the slower the convergence rate. This reflects the intuition that it becomes harder to precisely localize the point u_0 when $\mathcal{D}$ varies more smoothly and gradually around this point. The convergence rate also depends on the shrinkage parameter ρ_T. The "optimal" rate is achieved by setting $\rho_T = (\rho_T\sqrt{T})^{-1/k}$, i.e., by setting the shrinkage factor ρ_T to equal $T^{-\frac{1}{2(k+1)}}$. This yields the rate $\gamma_T = T^{-\frac{1}{2(k+1)}}$.

6.4 Verification of Conditions

Theorems 6.1–6.3 are derived under the high-order conditions (C1)–(C4). Conditions (C1) and (C2) require the process $\{X_{t,T}\}$ to be locally stationary and strongly mixing. These properties are well understood and have been verified for a wide range of processes as we have pointed out earlier on. In what follows, we have a closer look at the conditions (C3) and (C4). In particular, we show that they boil down to simple moment conditions in Examples I–III. As we will see, these moment conditions are fairly weak, in particular much weaker than those usually imposed in the related literature on stationarity tests; cp. for example Dette et al. [8] and Preuss et al. [24] who assume all moments of the underlying process $\{X_{t,T}\}$ to exist.
We first consider the time-varying mean setting from Example I. Here, the variables $X_{t,T}$ are real-valued and the function family of interest is $\mathcal{F} = \{\mathrm{id}\}$. It is straightforward to show that in this scenario, (C3) and (C4) are satisfied if the model variables fulfill the following moment assumptions:

(A_μ) Either (a) $\mathbb{E}|X_{t,T}|^r \leq C$ for some $r > 4$ and $\mathbb{E}U_{t,T}^2(u) \leq C$ or (b) $\mathbb{E}|X_{t,T}|^r \leq C$, $\mathbb{E}|X_t(u)|^r \leq C$ and $\mathbb{E}U_{t,T}^{r/(r-1)}(u) \leq C$ for some $r > 4$ and a sufficiently large constant C that is independent of u, t and T.

A similar situation arises in the setup of Example II, where $X_{t,T} = (Y_{t,T}, \ldots, Y_{t-p,T})^\intercal$ takes values in $\mathbb{R}^{p+1}$ and the function family under consideration is $\mathcal{F} = \{f_0, \ldots, f_p\}$ with $\mathbb{E}[f_\ell(X_{t,T})] = \mathbb{E}[Y_{t,T}Y_{t-\ell,T}]$. As before, (C3) and (C4) are fulfilled under a set of moment conditions:

(A_γ) It holds that $\mathbb{E}\|X_{t,T}\|^r \leq C$, $\mathbb{E}\|X_t(u)\|^r \leq C$ and $\mathbb{E}U_{t,T}^q(u) \leq C$ for some $r > 8$ and $q = \frac{r}{3}/(\frac{r}{3} - 1)$, where C is a sufficiently large constant that is independent of u, t and T.

We finally examine the setting from Example III. First of all, note that in this setup, we can assume without loss of generality that the variables $X_{t,T} = (Y_{t,T}, \ldots, Y_{t-p,T})^\intercal$ along with their stationary approximations $X_t(u)$ have bounded support, say $[0,1]^{p+1}$. The reason for this is as follows: Take $\psi : \mathbb{R} \to [0,1]$ to be any function which is strictly increasing and Lipschitz continuous (i.e. $|\psi(y) - \psi(y')| \leq L|y - y'|$ for a fixed constant L and all $y, y' \in \mathbb{R}$). Now consider the variables $Z_{t,T} = (\psi(Y_{t,T}), \ldots, \psi(Y_{t-p,T}))^\intercal$ together with their approximations $Z_t(u)$ which are defined analogously. It is easily seen that $\{Z_{t,T}\}$ is locally stationary and strongly mixing with the same mixing rate as $\{X_{t,T}\}$. Moreover, it holds that

$$\begin{aligned} \forall u, v \in [0,1]: \quad X_t(u) \overset{\mathcal{L}}{=} X_t(v) \quad &\Longleftrightarrow \quad Z_t(u) \overset{\mathcal{L}}{=} Z_t(v) \\ \forall s, t \in \{1, \ldots, T\}: \quad X_{t,T} \overset{\mathcal{L}}{=} X_{s,T} \quad &\Longleftrightarrow \quad Z_{t,T} \overset{\mathcal{L}}{=} Z_{s,T}, \end{aligned}$$

where $V \overset{\mathcal{L}}{=} W$ means that V and W have the same distribution. Hence, the distribution of $\{X_{t,T}\}$ is time-varying on the interval $[u,1]$ if and only if the distribution of $\{Z_{t,T}\}$ varies within this time span. Consequently, we can replace the process $\{X_{t,T}\}$ by $\{Z_{t,T}\}$ and work with the latter which has support on $[0,1]^{p+1}$. In what follows, we simply assume that the variables $X_{t,T}$ have support on $[0,1]^{p+1}$ themselves.

The non-smooth nature of the indicator functions $I(\cdot \leq x)$ poses some technical problems when verifying (C3). To circumvent these issues, we replace $I(y \leq x)$ with a smoothed version defined by

$$\varphi(x,y) = \prod_{\ell=0}^{p} \Big(1 - 1(y_\ell > x_\ell)\frac{(y_\ell - x_\ell)}{\varepsilon}\Big)^+$$

for some small $\varepsilon > 0$. It is straightforward to see that φ is Lipschitz continuous in all arguments. Moreover, it holds that $\lim_{\varepsilon \to 0} \varphi(x,y) = I(y \leq x)$, which means that φ provides a smooth approximation of the indicator I. We now define

$$F_\varphi(w,x) = \int_{\mathbb{R}^{p+1}} \varphi(x,y) dF(w,y) dy,$$

which may be regarded as a smoothed version of the distribution function

$$F(w,x) = \int_{\mathbb{R}^{p+1}} I(y \leq x) dF(w,y) dy.$$

Whereas the function $F(w,\cdot)$ fully characterizes the distribution of the variables $X_t(w)$, this is in general not true for the function $F_\varphi(w,\cdot)$. Intuitively, by smoothing the indicator I, we slightly blur the structure of the distribution function $F(w,\cdot)$. As a result, we are not able to detect time-variations in all aspects of the distribution any more. Nevertheless, when ε is fairly small, the smoothed function $F_\varphi(w,\cdot)$ gives a good approximation of $F(w,\cdot)$ and thus provides a fairly accurate picture of the distribution of the variables $X_t(w)$. Thus, we should still be able to reliably detect time-variations in the distribution.

With these comments and definitions at hand, we now replace the measure of time-variation from Example III along with its estimator by the modified versions

$$\begin{aligned}
\mathcal{D}_{F_\varphi}(u) &= \sup_{x \in [0,1]^{p+1}, v \in [u,1]} \big| D_{F_\varphi}(u,v,x) \big| \\
\hat{\mathcal{D}}_{F_\varphi,T}(u) &= \sup_{x \in [0,1]^{p+1}, v \in [u,1]} \big| \hat{D}_{F_\varphi,T}(u,v,x) \big|,
\end{aligned}$$

where

$$\begin{aligned}
D_{F_\varphi}(u,v,x) &= \int_v^1 F_\varphi(w,x) dw - \Big(\frac{1-v}{1-u}\Big) \int_u^1 F_\varphi(w,x) dw \\
\hat{D}_{F_\varphi,T}(u,v,x) &= \frac{1}{T} \sum_{t=\lceil vT+1 \rceil}^{T} \varphi(x, X_{t,T}) - \Big(\frac{1-v}{1-u}\Big) \frac{1}{T} \sum_{t=\lceil uT+1 \rceil}^{T} \varphi(x, X_{t,T}).
\end{aligned}$$

The function family associated with these statistics is $\mathcal{F} = \{\varphi(x, \cdot) : x \in [0,1]^{p+1}\}$. Here, we can restrict attention to $x \in [0,1]^{p+1}$ as the model variables have support on the cube $[0,1]^{p+1}$.
Noting that the metric entropy of the function class $\mathcal{F}$ is the same as that of the unit cube $[0,1]^{p+1}$ and exploiting the Lipschitz continuity of φ, it is easily seen that (C3) is satisfied for the class $\mathcal{F}$. Moreover, (C4) is fulfilled under the following moment assumption:

(A_{F_φ}) It holds that $\mathbb{E}U_{t,T}^r(u) \leq C$ for some $r \geq 1$ and a sufficiently large constant C that is independent of u, t and T.

Hence, if we work with the family $\mathcal{F}$ of smoothed indicator functions φ, then (A_{F_φ}) is sufficient to ensure that (C3) and (C4) hold true.

7 Implementation

In this section, we discuss how to choose the transformation function Φ and the shrinkage factor ρ_T to achieve a good small sample performance of the estimator $\hat{u}_0$. Our estimation procedure is based on the idea that the transformed statistic $\hat{q}_T(\cdot)$ approximately behaves like the step function $1(\cdot \leq u_0)$. We thus aim to choose Φ and ρ_T in a way which ensures that $\hat{q}_T(\cdot)$ yields a reasonable approximation to $1(\cdot \leq u_0)$ even for moderate sample sizes. The heuristic idea to achieve this is as follows:

Step 1. As a preliminary step, we slightly rewrite the transformed statistic $\hat{q}_T(u)$. Since the function Φ is strictly increasing, we have

$$\begin{aligned}\hat{q}_T(u) &= \Phi\big(\rho_T\sqrt{T}\hat{\mathcal{D}}_T(u)\big) \\ &= \Phi\Big(\rho_T\sqrt{T}\sup_{f\in\mathcal{F}}\sup_{v\in[u,1]}|\hat{D}_T(u,v,f)|\Big) = \sup_{f\in\mathcal{F}}\hat{q}_{f,T}(u),\end{aligned} \tag{7.1}$$

where

$$\hat{q}_{f,T}(u) = \Phi\Big(\rho_T\sqrt{T}\sup_{v\in[u,1]}|\hat{D}_T(u,v,f)|\Big).$$

According to (7.1), $\hat{q}_T(\cdot)$ is close to the step function $1(\cdot \leq u_0)$ if $\hat{q}_{f,T}(\cdot)$ is close to $1(\cdot \leq u_0)$ for each function $f \in \mathcal{F}$. In what follows, we thus restrict attention to the statistic $\hat{q}_{f,T}(\cdot)$ for an arbitrary but fixed function f. In particular, we attempt to select Φ and ρ_T in a way which guarantees that $\hat{q}_{f,T}(\cdot)$ is close to $1(\cdot \leq u_0)$. To emphasize that f is fixed, we make use of the notation $\hat{D}_{f,T}(u,v) := \hat{D}_T(u,v,f)$ and define the expressions $D_f(u,v)$, $\hat{H}_{f,T}(u,v)$ and $H_f(u,v)$ analogously.

Step 2. We now normalize the processes $\hat{D}_{f,T}$ and $\hat{H}_{f,T}$ in a suitable way. By Theorem 6.1, we know that

$$\hat{H}_{f,T}(u,\cdot) \rightsquigarrow H_f(u,\cdot)$$

for any time point u, where the asymptotic covariances are given in (6.1). Inspecting formula (6.1), the covariances are seen to depend on expressions of the form $(1-v)\sigma_f^2(v)$, where

$$\sigma_f^2(v) = \frac{\sum_{l=-\infty}^{\infty} \int_v^1 c_l(w)dw}{1-v}$$

and $c_l(w) = \mathrm{Cov}(f(X_0(w)), f(X_l(w)))$. In what follows, we take for granted that the functions $c_l(\cdot)$ are constant on the interval $[u_0, 1]$, which implies that $\sigma_f(\cdot)$ is constant on this time span as well. This assumption is not very restrictive, as it is satisfied in a wide range of settings. Consider for example the time-varying mean model $X_{t,T} = \mu(\frac{t}{T}) + \varepsilon_t$. As long as the error process $\{\varepsilon_t\}$ is stationary, $c_l(w)$ has the same value at all points $w \geq u_0$. A similar situation arises in a wide range of models with time-varying parameters such as the AR model in (4.3): Suppose we want to estimate the point u_0 where the model parameters start to vary over time. Since the parameters are constant on the interval $[u_0, 1]$, the process is stationary on the interval $[u_0, 1]$, implying that $c_l(\cdot)$ is a constant function on this time span. As a final example, consider the setting from Example III. There, $c_l(\cdot)$ is constant on $[u_0, 1]$ as well, since the processes $\{X_t(w)\}$ have the same distribution for all $w \geq u_0$.

We now use the expression $\sigma_f(u)$ to normalize the process $\hat{H}_{f,T}$. In particular, we define the scaled version $\hat{H}^{\mathrm{sc}}_{f,T}$ of the process by

$$\hat{H}^{\mathrm{sc}}_{f,T}(u,v) = \frac{\hat{H}_{f,T}(u,v)}{\sigma_f(u)\sqrt{1-u}}.$$

Analogously setting $H^{\mathrm{sc}}_f(u,v) = H_f(u,v)/\sigma_f(u)\sqrt{1-u}$, Theorem 6.1 implies that

$$\hat{H}^{\mathrm{sc}}_{f,T}(u,\cdot) \rightsquigarrow H^{\mathrm{sc}}_f(u,\cdot)$$

at each time point $u \in [0,1)$. Since $\sigma_f(u) = \sigma_f(v)$ for any pair of time points $u, v \geq u_0$, the covariance structure of the scaled limit process $H^{\mathrm{sc}}_f(u,\cdot)$ is given by

$$\mathrm{Cov}\big(H^{\mathrm{sc}}_f(u,v), H^{\mathrm{sc}}_f(u,v')\big) = \min\Big\{\Big(\frac{1-v}{1-u}\Big), \Big(\frac{1-v'}{1-u}\Big)\Big\} - \Big(\frac{1-v}{1-u}\Big)\Big(\frac{1-v'}{1-u}\Big)$$

whenever $u \geq u_0$. Noticing that $u \leq v, v' \leq 1$ and thus $0 \leq \frac{1-v}{1-u} \leq 1$ as well as $0 \leq \frac{1-v'}{1-u} \leq 1$, this turns out to be the covariance structure of a standard Brownian bridge on the unit interval.

Next define $\hat{D}^{\mathrm{sc}}_{f,T}(u,v) = \hat{D}_{f,T}(u,v)/\sigma_f(u)\sqrt{1-u}$. Since $\sqrt{T}\hat{D}^{\mathrm{sc}}_{f,T}(u,\cdot) = \hat{H}^{\mathrm{sc}}_{f,T}(u,\cdot)$ at any fixed time point $u \geq u_0$, the above considerations immediately imply that $\sqrt{T}\hat{D}^{\mathrm{sc}}_{f,T}(u,\cdot)$ weakly converges to a standard Brownian bridge B at any point $u \in [u_0, 1)$. Writing

$$\hat{\mathcal{D}}^{\mathrm{sc}}_{f,T}(u) = \sup_{v\in[u,1]} \big|\hat{D}^{\mathrm{sc}}_{f,T}(u,v)\big|,$$

we can further conclude that for any $f \in \mathcal{F}$,

$$\sqrt{T}\hat{\mathcal{D}}^{\mathrm{sc}}_{f,T}(u) \xrightarrow{d} \sup_{w\in[0,1]} |B(w)|$$

at any time point $u \in [u_0, 1)$, where the distribution function of $\sup_{w\in[0,1]} |B(w)|$ is given by

$$\Psi(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}\exp(-2k^2x^2). \tag{7.2}$$

Step 3. We next examine the behaviour of the expression

$$\hat{q}^{\mathrm{sc}}_{f,T}(u) = \Psi\Big(\overline{\rho}\sqrt{T}\hat{\mathcal{D}}^{\mathrm{sc}}_{f,T}(u)\Big), \tag{7.3}$$

where the distribution function Ψ defined in (7.2) plays the role of the transformation Φ and $\overline{\rho} = q_{0.5}/q_{0.99}$ with q_α being the α-quantile of Ψ. The constant $\overline{\rho}$ is closely linked to the shrinkage factor ρ_T as we will see shortly.
The main idea behind the construction of $\hat{q}^{\mathrm{sc}}_{f,T}(u)$ is to achieve a good approximation of the step function $1(u \le u_0)$:

(i) For time points $u \ge u_0$, the statistic $\sqrt{T}\hat{\mathcal{D}}^{\mathrm{sc}}_{f,T}(u)$ is approximately distributed according to Ψ. The factor $\overline{\rho}$ shrinks the statistic in a specific way: If Z is distributed according to Ψ, then $\mathbb{P}(\Psi(\overline{\rho}Z) \le 0.5) = 0.99$. Hence, $\hat{q}^{\mathrm{sc}}_{f,T}(u) \le 0.5$ with high probability, in particular with probability around 0.99.

(ii) At time points $u < u_0$, the statistic $\sqrt{T}\hat{\mathcal{D}}^{\mathrm{sc}}_{f,T}(u)$ mimics the behaviour of the diverging expression $\sqrt{T}\mathcal{D}^{\mathrm{sc}}_{f}(u)$. Consequently, $\sqrt{T}\hat{\mathcal{D}}^{\mathrm{sc}}_{f,T}(u)$ gets pushed into the extreme upper tail of Ψ, which means that $\hat{q}^{\mathrm{sc}}_{f,T}(u)$ should take values close to one.

Taken together, these considerations suggest that $\hat{q}^{\mathrm{sc}}_{f,T}(u)$ should give a reasonable approximation of $1(u \le u_0)$.
As already indicated above, the constant $\overline{\rho}$ plays the role of the shrinkage factor ρ_T in (7.3). It is designed to meet two opposite requirements: On the one hand, we should avoid shrinking the statistic $\sqrt{T}\hat{\mathcal{D}}^{\mathrm{sc}}_{f,T}(u)$ too strongly because otherwise its diverging behaviour at time points $u < u_0$ is dampened too much. On the other hand, we need to shrink it sufficiently strongly to guarantee that $\hat{q}^{\mathrm{sc}}_{f,T}(u)$ is small for $u \ge u_0$. In particular, we would like $\hat{q}^{\mathrm{sc}}_{f,T}(u)$ to be closer to zero than to one at the points $u \ge u_0$ for most of the time, i.e., we would like it to be smaller than 0.5 with high probability at these points. This suggests to use $\overline{\rho} = q_{0.5}/q_{0.99}$ as the shrinkage factor ρ_T.[1] However, as ρ_T must converge to zero from a theoretical perspective, we cannot take the constant $\overline{\rho}$ at face value but have to replace it by an appropriate sequence.

[1]Clearly, $q_{0.99}$ may be replaced by a slightly different quantile like $q_{0.95}$ or $q_{0.975}$.

Specifically, we may set $\rho_T = \overline{\rho} T^{-1/N}$, where N is a large natural number. The so-defined shrinkage factor ρ_T converges to zero but is very close to $\overline{\rho}$ for any reasonable sample size which appears in practice. In applications, we may ignore the difference between ρ_T and $\overline{\rho}$ and simply set $\rho_T = \overline{\rho}$.

Step 4. Since the normalization $\sigma_f(u)$ is not known in practice, we cannot work with the statistic $\hat{q}^{\text{sc}}_{f,T}(u)$ directly but have to replace $\sigma_f(u)$ by an estimate. As long as the model under consideration is not too complicated, this is a fairly straightforward task. Consider for example the time-varying mean model $X_{t,T} = \mu(\frac{t}{T}) + \varepsilon_t$ with i.i.d. residuals ε_t. Denoting the error variance by $\sigma^2 = \mathbb{E}[\varepsilon_t^2]$, it holds that $\sigma_f(u) = \sigma$ which is easily estimated by standard techniques.
To construct a general estimator of $\sigma_f(u)$, we define $Z_{t,T} = f(X_{t,T}) - \mathbb{E}[f(X_{t,T})]$ along with $Z_t(w) = f(X_t(w)) - \mathbb{E}[f(X_t(w))]$ and write

$$\sigma_f^2(u) = \sum_{l=-\infty}^{\infty} \nu_l(u) \quad \text{with} \quad \nu_l(u) := \frac{\int_u^1 c_l(w)dw}{1-u} = \frac{\int_u^1 \mathbb{E}[Z_0(w)Z_l(w)]dw}{1-u}.$$

This formula shows that $\sigma_f^2(u)$ essentially is the average long-run variance of the processes $\{Z_t(w)\}$ on the interval $[u,1]$. This suggests to estimate $\sigma_f^2(u)$ by

$$\hat{\sigma}_f^2(u) = \sum_{l=-L_T}^{L_T} \hat{\nu}_l(u)$$

with

$$\hat{\nu}_l(u) = \frac{1}{(1-u)T} \sum_{t=\lceil uT+1 \rceil}^{T} \hat{Z}_{t,T}\hat{Z}_{t+l,T},$$

where $\hat{Z}_{t,T} = f(X_{t,T}) - \hat{m}(\frac{t}{T})$ and L_T is a cutoff sequence that diverges to infinity at an appropriate rate. Here, $\hat{m}(\frac{t}{T})$ is a standard Nadaraya-Watson estimator of $\mathbb{E}[f(X_{t,T})]$. Specifically, $\hat{m}(w) = T^{-1}\sum_{t=1}^{T} K_h(w - \frac{t}{T}) f(X_{t,T})$ with K being a kernel function and $K_h(x) = h^{-1}K(x/h)$. Alternatively, a local linear or more generally a local polynomial estimator may be employed.
It is worth noting that it is not essential for our method to have an extremely precise estimate of $\sigma_f(u)$ at one's disposal. The estimate should just be precise enough to make sure that the distribution function of the statistic $\sqrt{T}\hat{\mathcal{D}}^{\text{sc}}_{f,T}(u)$ is not too far away from Ψ for time points $u \geq u_0$. Hence, in most cases a fairly rough estimator based on a crude choice of the truncation parameter L_T and the bandwidth h will do in practice.

Taken together, the above considerations suggest to implement our procedure as follows in practice: To start with, we normalize our measure of time-variation by the term $\hat{\sigma}_f(u)\sqrt{1-u}$, thus yielding

$$\hat{\mathcal{D}}^{\text{sc}}_{f,T}(u) = \sup_{v \in [u,1]} \Big| \frac{\hat{D}_T(u,v,f)}{\hat{\sigma}_f(u)\sqrt{1-u}} \Big|$$

with $\hat{\sigma}_f(u)$ defined in Step 4. Choosing Ψ as the transformation function and setting $\rho_T = \overline{\rho} T^{-1/N}$ as described in Step 3, we further define

$$\hat{q}^{\text{sc}}_{f,T}(u) = \Psi\Big(\rho_T \sqrt{T} \hat{\mathcal{D}}^{\text{sc}}_{f,T}(u)\Big)$$

together with $\hat{q}^{\text{sc}}_T(u) = \sup_{f \in \mathcal{F}} \hat{q}^{\text{sc}}_{f,T}(u)$, which leads to the criterion function $\hat{Q}^{\text{sc}}_T(u) = u + (1-u)\hat{q}^{\text{sc}}_T(u)$. The estimator that results from minimizing this criterion function is denoted by $\hat{u}^{\text{sc}}_0$. The following corollary summarizes its asymptotic properties.

Corollary 7.1. *Let the conditions of Theorem 6.3 be satisfied. Moreover, assume that there exist constants $\underline{\sigma}$ and $\overline{\sigma}$ such that $0 < \underline{\sigma} \leq \hat{\sigma}_f(u) \leq \overline{\sigma} < \infty$ for all $u \in [0,1]$ and $f \in \mathcal{F}$ with probability approaching one. Then*

$$\hat{u}^{\text{sc}}_0 - u_0 = O_p(\gamma_T)$$

with γ_T as defined in Theorem 6.3.

The additional condition that $\underline{\sigma} \leq \hat{\sigma}_f(u) \leq \overline{\sigma}$ for all u and f with probability approaching one can be shown to be satisfied in a wide range of cases. Rather than going into the technical details, we briefly describe a way to get rid of this condition: We simply replace the estimate $\hat{\sigma}_f(u)$ with the truncated version

$$\hat{\sigma}^{\text{tr}}_f(u) = \underline{\sigma} \cdot 1(\hat{\sigma}_f(u) < \underline{\sigma}) + \hat{\sigma}_f(u) \cdot 1(\underline{\sigma} \leq \hat{\sigma}_f(u) \leq \overline{\sigma}) + \overline{\sigma} \cdot 1(\overline{\sigma} < \hat{\sigma}_f(u)),$$

which is bounded by the constants $\underline{\sigma}$ and $\overline{\sigma}$ by construction. When $\underline{\sigma}$ and $\overline{\sigma}$ are chosen sufficiently small and large, respectively, then $\hat{\sigma}_f(u)$ and $\hat{\sigma}^{\text{tr}}_f(u)$ differ only slightly and there is no difference in using $\hat{\sigma}_f(u)$ or $\hat{\sigma}^{\text{tr}}_f(u)$ from an applied point of view.

8 Simulations

We now examine the finite sample performance of our estimator $\hat{u}^{\text{sc}}_0$ in a Monte-Carlo experiment. To do so, we consider a variety of different time series processes which are stationary on the rescaled time interval $[u_0, 1]$ but deviate from stationarity on any interval $[u, 1]$ with $u < u_0$. In all settings, u_0 is equal to 0.5. For each model, we generate $N = 5000$ samples of length $T \in \{500, 1000\}$ and apply our procedure to estimate u_0. We thus obtain $N = 5000$ estimates of u_0 for each model specification. The results are presented by histograms that show the empirical distribution of the estimates for each specification. In particular, the bars in the plots give the number of simulations (out of a total of 5000) in which a certain value $\hat{u}^{\text{sc}}_0$ is obtained.

8.1 Time-Varying Mean Models

To start with, we consider the model

$$X_{t,T} = \mu\Big(\frac{t}{T}\Big) + \varepsilon_t \tag{8.1}$$

with different mean functions $\mu = \mu_k$ $(k = 1, 2, 3)$. The residuals ε_t are assumed to follow the AR(1) process $\varepsilon_t = 0.25\varepsilon_{t-1} + \eta_t$, where the innovations η_t are i.i.d. normal with zero mean and standard deviation 0.5. The mean functions are all piecewise linear and equal to zero for time points larger than $u_0 = 0.5$. Specifically,

$$\begin{aligned}
\mu_1(u) &= 1(u \leq 0.4) + [1 - 10(u - 0.4)] \cdot 1(0.4 < u < 0.5) \\
\mu_2(u) &= 1(u \leq 0.25) + [1 - 4(u - 0.25)] \cdot 1(0.25 < u < 0.5) \\
\mu_3(u) &= 4u \cdot 1(u \leq 0.25) + [1 - 4(u - 0.25)] \cdot 1(0.25 < u < 0.5).
\end{aligned}$$

The functions μ_1 and μ_2 start to linearly deviate from zero at the point 0.5 until they reach a value of one and then constantly remain at this value. The function μ_3, in contrast, is tent-shaped on the interval $[0, 0.5]$.

To estimate the point u_0, we base our estimation method on the statistic $\hat{D}_{\mu,T}$ from Example I and use the implementation strategy outlined in Section 7. As suggested there, the shrinkage parameter is set equal to $\overline{\rho} = q_{0.5}/q_{0.99}$. In the present setting, the scaling factor $\sigma_f^2(u) = \sigma_{\text{id}}^2(u)$ is equal to the long-run variance $\sum_{l=-\infty}^{\infty} \mathbb{E}[\varepsilon_0 \varepsilon_l]$ of the error terms. To estimate it, we proceed as described in Step 4 of Section 7. Specifically, we pick the bandwidth h of the Nadaraya-Watson estimator to equal 0.1 and truncate the infinite sum at 5 (and -5), i.e., we take into account autocovariances up to the fifth order. As a robustness check, we have varied the truncation points between 2 and 10. This yields very similar results, underpinning our claim from Section 7 that it is not essential to work with a very precise estimator of the scaling factor.

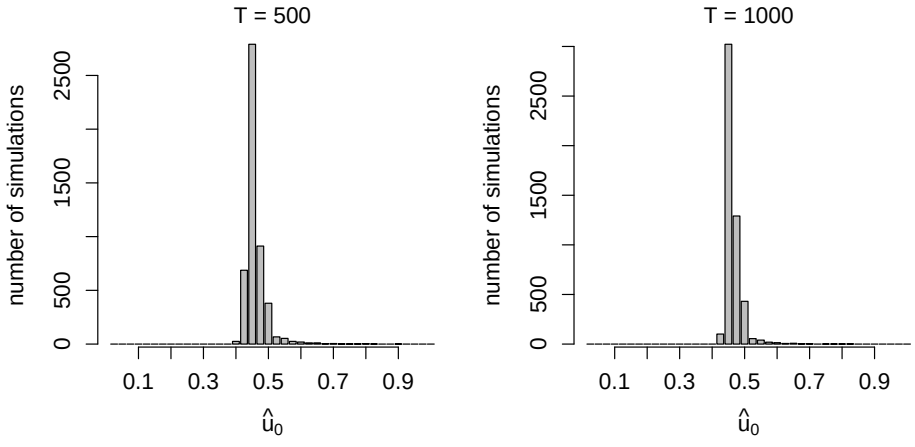

Figure 2: Simulation results for model (8.1) with the mean function μ_1.

The simulation results for the design with μ_1 are presented in Figure 2, the left-hand panel corresponding to a sample size of $T = 500$ and the right-hand one to $T = 1000$. Since μ_1 deviates from zero fairly quickly, our procedure is able to localize the point $u_0 = 0.5$ quite precisely. Indeed, the histograms show that the estimates are not very dispersed but cluster tightly around u_0. The plots also make visible a downward bias of the estimates. This bias reflects the difficulty of the change point problem under consideration. In fact it is very hard to detect smooth time-variations on an interval $[u_0 - \xi, 1]$ if ξ is very small. As can be seen, the bias becomes less pronounced when moving to the larger sample size $T = 1000$.

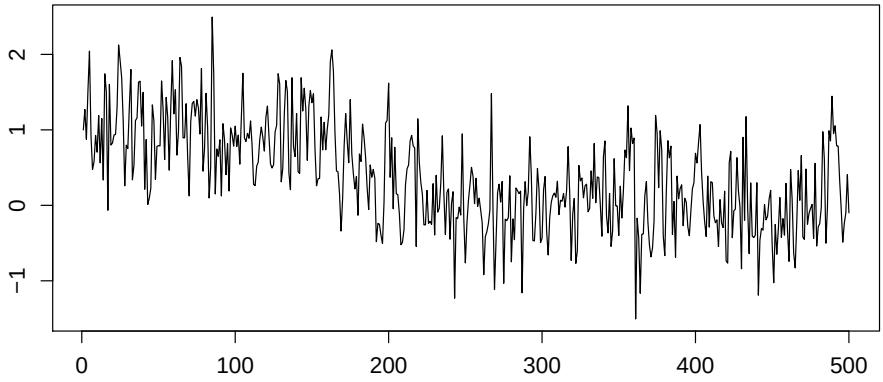

Figure 3: A typical sample of length 500 simulated from model (8.1) with the function μ_2.

We next turn to the design with μ_2. Since μ_2 deviates from zero much more slowly than μ_1, it is harder for our method to localize the point u_0. This is illustrated by Figure 3 which depicts a typical sample of length 500 drawn from this design. As can be seen, the deviation of μ_2 from zero is clearly visible only at time points u much smaller than u_0. When getting closer to u_0, the signal of the time-variation becomes fairly weak and is more and more dominated by the noise of the error term.

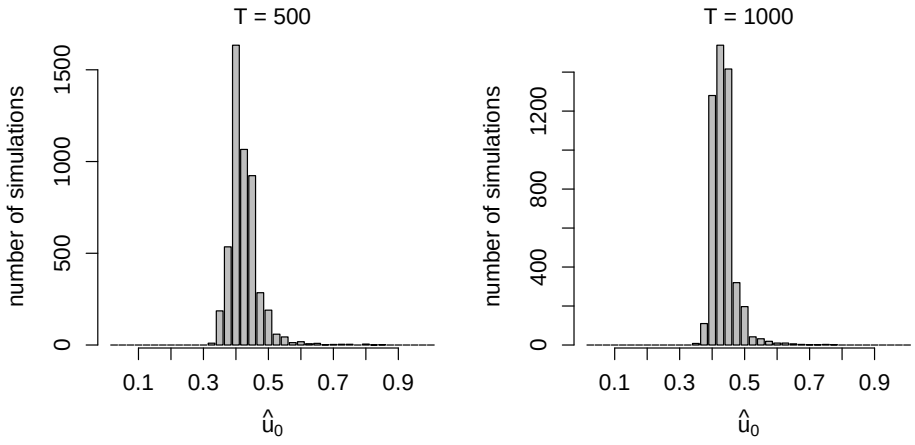

Figure 4: Simulation results for model (8.1) with the mean function μ_2.

The simulation results for the design with μ_2 are shown in Figure 4. As can be seen there, the distribution of the estimates is more dispersed than in the design with μ_1, reflecting the fact that it is harder to detect the point u_0 in this setting. Nevertheless, the great bulk of estimates takes values in the region between 0.4 and 0.5, thus providing us with a reasonable approximation of u_0.

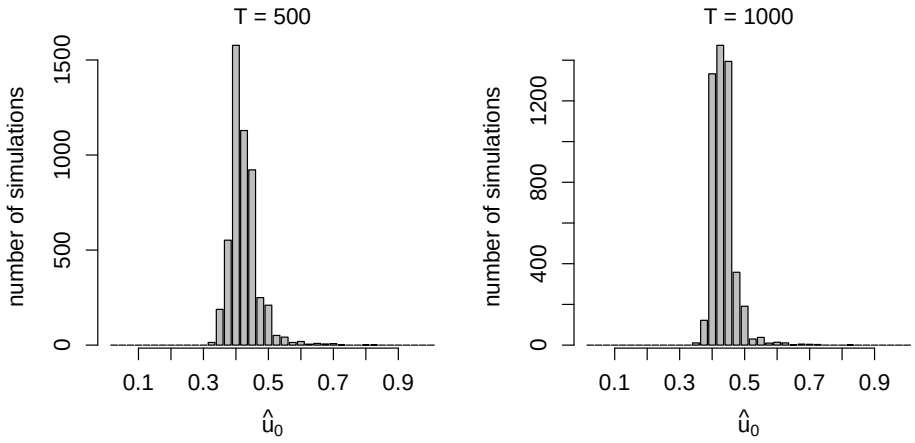

Figure 5: Simulation results for model (8.1) with the mean function μ_3.

We finally turn to the results for the setting with μ_3. Since the function μ_3 deviates from zero in the same way as μ_2 on the segment $[0.25, 0.5]$, we may expect our procedure to perform similarly as in the previous setting. This is confirmed by the histograms in Figure 5 which strongly resemble those in Figure 4.

8.2 Time-Varying Autoregressive Models

We next investigate a couple of time-varying AR models. In particular, we consider the AR(1) setting

$$Y_{t,T} = a\Big(\frac{t}{T}\Big)Y_{t-1,T} + \varepsilon_t \tag{8.2}$$

with two different parameter functions $a = a_k$ $(k = 1, 2)$ and i.i.d. residuals ε_t that are normally distributed with zero mean and unit variance. The coefficient functions are given by

$$a_1(u) = 0.5 \cdot 1(u < 0.5) - 0.5 \cdot 1(u \geq 0.5)$$
$$a_2(u) = 0.5 \cdot 1(u \leq 0.4) + [0.5 - 10(u - 0.4)] \cdot 1(0.4 < u < 0.5) - 0.5 \cdot 1(u \geq 0.5).$$

The function a_1 has a break at the point $u_0 = 0.5$, where it jumps from its baseline value -0.5 to a value of 0.5. The function a_2 in contrast linearly deviates from its baseline -0.5 until it reaches the value 0.5. To estimate the point u_0, we employ the statistic $\hat{\mathcal{D}}_{\gamma,T}$ from Example II with $p = 1$, i.e., we take into account covariances up to the first order and implement our method along the lines of Section 7. To estimate the

scaling factor, we set $h = 0.1$ and truncate the infinite sum occurring in its definition as before at 5. As a robustness check, we have again varied the truncation point but found that it does not affect the procedure in any notable way.
We first report the simulation results for the AR design with a_1. In this setting, the deviation from stationarity occurs instantaneously. Our method is thus able to detect the point $u_0 = 0.5$ quite precisely as can be seen from the histograms in Figure 6. Again the estimates are downward biased in small samples, the bias being much less pronounced for the larger sample size $T = 1000$.

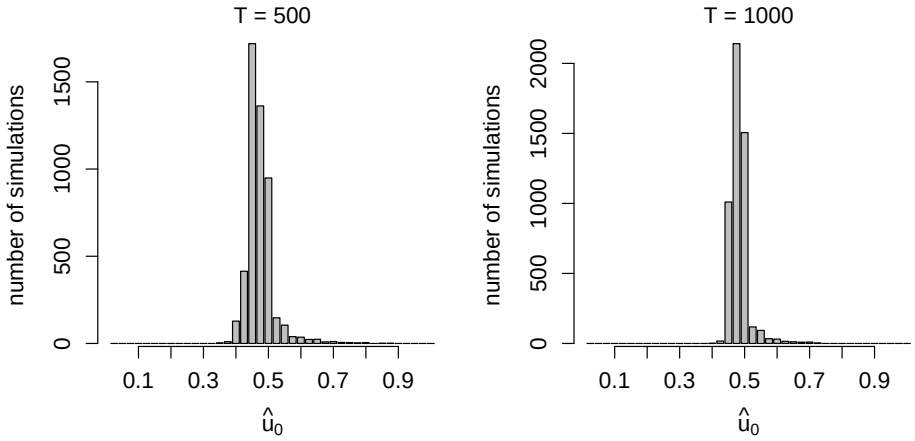

Figure 6: Simulation results for the AR model (8.2) with the parameter function a_1.

Moving to the second design with the function a_2, the deviation from stationarity is not as clear-cut as in the previous setting but occurs gradually. As is illustrated by Figure 7, the visual appearance of the time series in the transition region $[0.4, 0.5]$ is fairly similar to that within the time span $[0.5, 1]$. Hence, it is quite difficult for our method to localize the time point $u_0 = 0.5$.

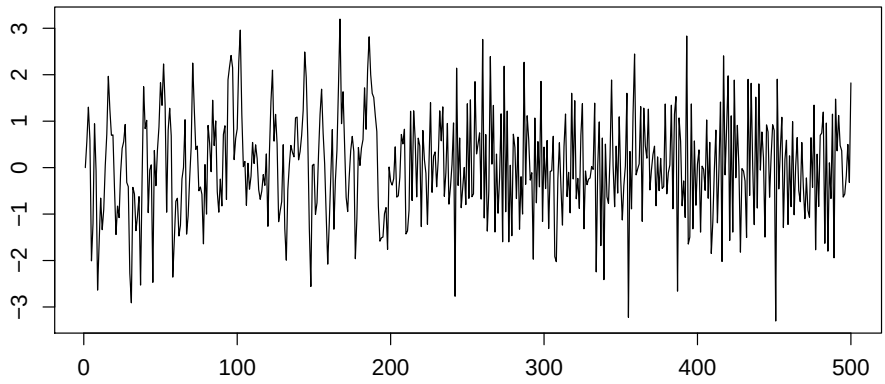

Figure 7: A typical sample of length 500 simulated from model (8.2) with the parameter function a_2.

Figure 8 displays the simulation results for the design with a_2. As expected, the precision of the estimator is lower than in the jump design with a_1. Nevertheless, most of the estimates take values between 0.4 and 0.5, thus picking up the structural change in the parameter function quite quickly and accurately.

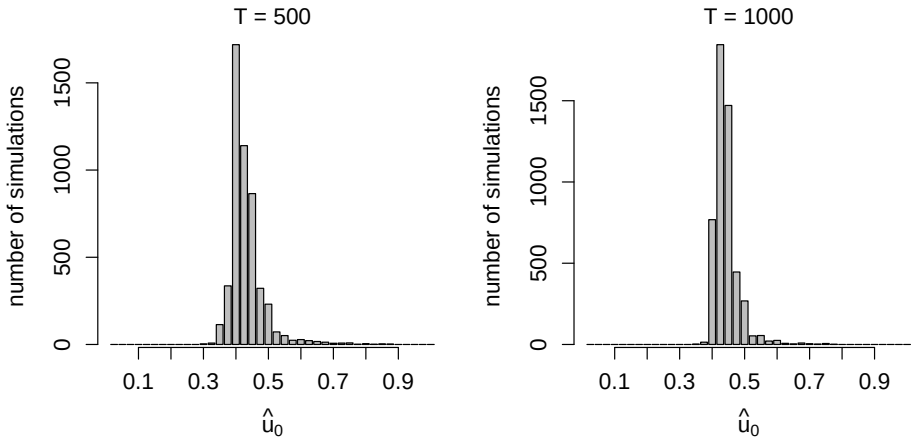

Figure 8: Simulation results for the AR model (8.2) with the parameter function a_2.

8.3 A Time-Varying Volatility Model

We finally consider the time-varying volatility model

$$Y_{t,T} = \sigma\Big(\frac{t}{T}\Big)\varepsilon_t \tag{8.3}$$

with two different volatility functions $\sigma = \sigma_k$ $(k = 1, 2)$ and i.i.d. residuals ε_t that are normally distributed with zero mean and unit variance. The volatility functions are given by

$$\begin{aligned} \sigma_1(u) &= 2 \cdot 1(u < 0.5) + 1 \cdot 1(u \geq 0.5) \\ \sigma_2(u) &= 2 \cdot 1(u < 0.4) + [2 - 10(u - 0.4)] \cdot 1(0.4 < u < 0.5) + 1 \cdot 1(u \geq 0.5). \end{aligned}$$

As in the AR setting, the function σ_1 has a jump at $u_0 = 0.5$, whereas σ_2 smoothly deviates from its baseline value 1. We base our method on the statistic $\hat{\mathfrak{D}}_{F,T}$ from Example III with $p = 0$, thus restricting attention to time-variations in the marginal distribution of the variables $Y_{t,T}$. As before, we follow the strategy from Section 7 to implement our method.

The simulation results for the design with σ_1 are summarized in Figure 9. Since the time-variation is exactly localized at the point $u_0 = 0.5$, our method is able to pick it up rather quickly. The distribution of the estimates is thus not very dispersed, but is again downward biased in small samples.

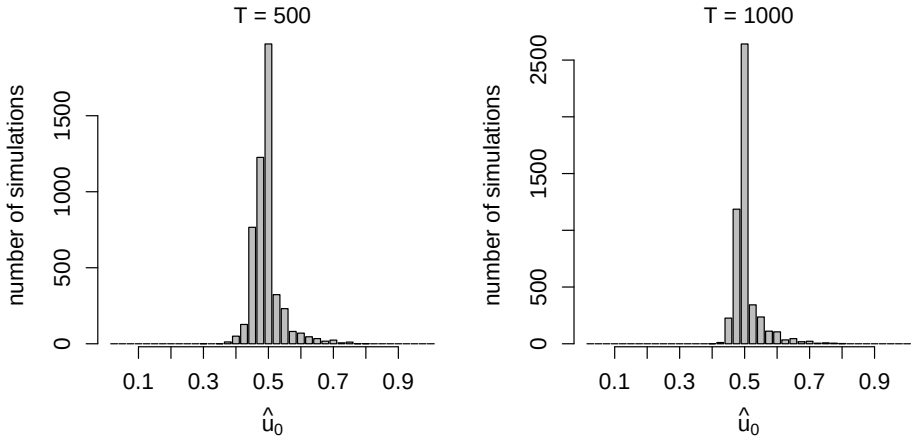

Figure 9: Simulation results for model (8.3) with the volatility function σ_1.

We finally turn to the results for the design with σ_2 which are displayed in Figure 10. As expected, the histograms are more dispersed in this setting, reflecting the fact that the time-variation is smooth and gradual in this case.

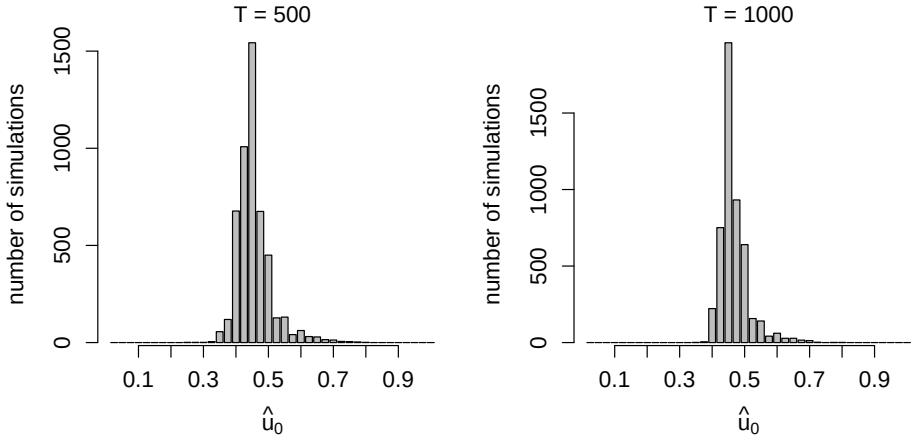

Figure 10: Simulation results for model (8.3) with the volatility function σ_2.

9 Application

To illustrate our estimation procedure, we apply it to a sample of financial return and volatility data. Specifically, we consider a sample of daily returns and realized volatilities for the S&P 500 index which are depicted in Figure 11.[2] The data span the period from the beginning of 2011 to the beginning of 2013, leaving us with a sample of approximately 500 data points. In what follows, we apply our estimation method

[2]The data are taken from Oxford-Man Institute's "realized library" [13]. In particular, our volatility data are calculated by taking the square root of the library's realized variance time series that is based on a realized kernel estimator.

both to the return and to the volatility time series. For each of the two series, we estimate the time point u_0 where the data start to severely deviate from stationarity.

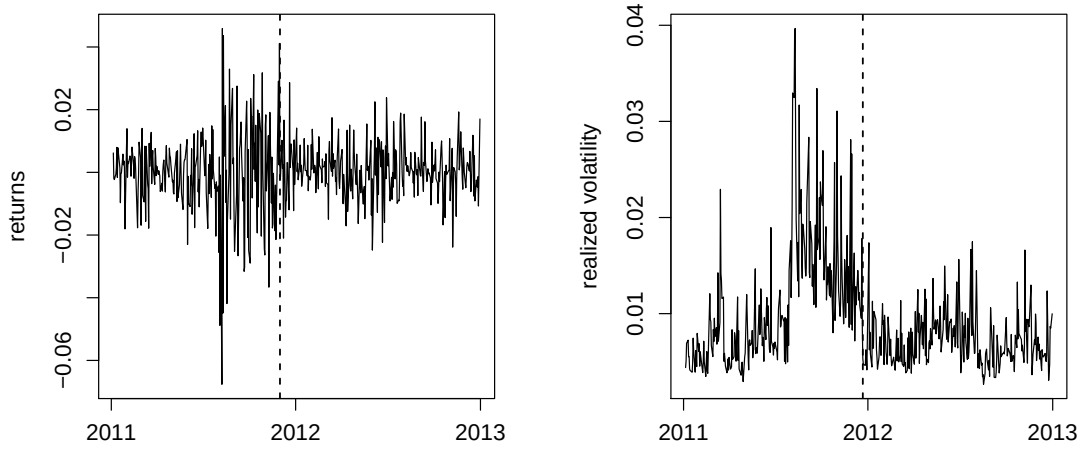

Figure 11: The left-hand panel shows the time series of daily returns, the right-hand panel the time series of realized volatilities. The estimates of u_0 are indicated by the dashed vertical lines.

We first examine the time series of daily returns. A simple locally stationary model for financial returns is given by the equation

$$r_{t,T} = \sigma\Big(\frac{t}{T}\Big)\varepsilon_t, \tag{8.4}$$

where $r_{t,T}$ denotes the daily return, σ is a time-varying volatility function and ε_t are i.i.d. residuals. Model (8.4) has been studied in a variety of papers; see Drees & Stărică [9] and Fryzlewicz et al. [11] among others. It suggests to estimate u_0 by means of a statistic that measures time-variations in the unconditional variance level $\sigma^2(\frac{t}{T})$. Such a statistic is given by $\hat{\mathcal{D}}_T(u) = \sup_{v\in[u,1]} |\hat{D}_T(u,v)|$, where

$$\hat{D}_T(u,v) = \frac{1}{T}\sum_{t=\lceil vT+1\rceil}^{T} r_{t,T}^2 - \Big(\frac{1-v}{1-u}\Big)\,\frac{1}{T}\sum_{t=\lceil uT+1\rceil}^{T} r_{t,T}^2.$$

Alternatively, we may base our estimation method on a statistic which is able to detect time-variations in a wider range of distributional features. In particular, as in the simulations, we may work with the statistic $\hat{\mathcal{D}}_T(u) = \sup_{x\in\mathbb{R},v\in[u,1]} |\hat{D}_T(u,v,x)|$, where

$$\hat{D}_T(u,v,x) = \frac{1}{T}\sum_{t=\lceil vT+1\rceil}^{T} I(r_{t,T}\le x) - \Big(\frac{1-v}{1-u}\Big)\,\frac{1}{T}\sum_{t=\lceil uT+1\rceil}^{T} I(r_{t,T}\le x).$$

As turns out, both variants of our method yield the same estimate $\hat{u}_0^{\mathrm{sc}}$ which is depicted by the dashed vertical line in the left-hand panel of Figure 11. Having a closer look at

the plot of the return data, $\hat{u}_0^{\text{sc}}$ appears to be a reasonable estimate of the point where the time series starts to deviate from stationarity. Indeed, visual inspection suggests that the returns become much more volatile in the time period before $\hat{u}_0^{\text{sc}}$.
We next turn to the time series of volatilities. Daily realized volatility is commonly modelled by means of autoregressive processes. Since it is characterized by slowly decaying sample autocorrelations, long-memory models like ARFIMA have been suggested quite frequently; see Andersen et al. [1] among others. However, as pointed out for example in Mikosch & Stărică [20], the long-memory behaviour may be spuriously generated by nonstationarities in the volatility process. This has led several authors to use autoregressive processes with time-varying parameters for modelling volatility; see e.g. Chen et al. [4]. Following this line of thought, a simple model of daily volatility is given by

$$v_{t,T} = a_0\Big(\frac{t}{T}\Big) + a_1\Big(\frac{t}{T}\Big)v_{t-1,T} + \varepsilon_t, \tag{8.5}$$

where $v_{t,T}$ denotes realized volatility and ε_t are i.i.d. innovations. Model (8.5) suggests to base our estimation method on a statistic which is able to capture time-variations in the parameter functions a_0 and a_1. This is for example achieved by the statistic $\hat{\mathcal{D}}_T(u) = \max_{1\le k\le 3}\sup_{v\in[u,1]}\big|\hat{D}_T(u,v,k)\big|$, where

$$\hat{D}_T(u,v,k) = \frac{1}{T}\sum_{t=\lceil vT+1\rceil}^{T} w_{t,T}^{(k)} - \Big(\frac{1-v}{1-u}\Big)\,\frac{1}{T}\sum_{t=\lceil uT+1\rceil}^{T} w_{t,T}^{(k)}$$

with $w_{t,T}^{(1)} = v_{t,T}$, $w_{t,T}^{(2)} = v_{t,T}^2$ and $w_{t,T}^{(3)} = v_{t,T}v_{t-1,T}$. This statistic can be regarded as combining the measures from Examples I and II. By construction, it is able to detect time-variations in the mean, the variance as well as the first autocovariance of the volatility process. The estimate of u_0 obtained from using this statistic is plotted as the vertical dashed line in the right-hand panel of Figure 11. Again, visual inspection of the volatility data suggests that our method gives a reasonable approximation of the time point where the volatility process starts to deviate from stationarity.

Acknowledgements

This work has been supported in part by the Collaborative Research Center "Statistical modeling of nonlinear dynamic processes" (SFB 823, Teilprojekt A1, C1) of the German Research Foundation (DFG). We would like to thank Rainer Dahlhaus, Wolfgang Polonik and Stanislav Volgushev for helpful discussions and comments on an earlier version of this manuscript. We are also grateful to Alina Dette and Martina Stein, who typed parts of this paper with considerable techincal expertise.

Appendix

In this appendix, we prove the main theoretical results of the paper. Throughout the appendix, the symbol C denotes a generic constant which may take a different value on each occurrence. Moreover, the expression $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$ is used to denote the L_p-norm of a real-valued random variable X.

Auxiliary Results

Before we turn to the proofs of the main theorems, we derive some technical lemmas which are needed later on. To formulate them, we introduce some additional notation. To start with, partition the observations $\{X_{t,T}, t = 1, \dots, T\}$ into blocks of size q, where the r-th block spans the observations from time point $(r-1)q+1$ to rq and we set $q = CT^b$ for some small $b > 0$ (in particular $b < \frac{1}{4}$). Now define

$$W_T(k, k') = \sup_{f \in \mathcal{F}} \Big| \sum_{r=k}^{k'} Q_{r,T}(f) \Big|$$

along with

$$Q_{r,T}(f) = \frac{1}{\sqrt{(k'-k+1)q}} \sum_{t=(2r-2)q+1}^{(2r-1)q \wedge T} \big(f(X_{t,T}) - \mathbb{E} f(X_{t,T}) \big).$$

The terms $Q_{r,T}(f)$ are scaled sums of the variables $f(X_{t,T}) - \mathbb{E} f(X_{t,T})$, the summation running over the observations of the $(2r-1)$-th block. The expression $W_T(k, k')$ sums up the terms $Q_{k,T}(f), \dots, Q_{k',T}(f)$ which correspond to the odd blocks $(2k-1), (2k+1), (2k+3), \dots, (2k'-1)$. The next two lemmas provide a bound on the L_p-norm of $W_T(k, k')$.

Lemma A.1. *Let (C1) and (C2) be satisfied and let $f_0 \in \mathcal{F}$ have the property that $\mathbb{E}|f_0(X_{t,T})|^{(1+\delta)p} \leq C$ for some even $p \in \mathbb{N}$ and a small $\delta > 0$. Then*

$$\Big\| \sum_{r=k}^{k'} Q_{r,T}(f_0) \Big\|_p \leq C$$

for some sufficiently large constant C.

Proof. To shorten notation, write $w_{t,T} = f_0(X_{t,T}) - \mathbb{E} f_0(X_{t,T})$ and consider the term

$$\begin{aligned} V_T = V_T(k, k') &= \mathbb{E}\Big[\Big(\sum_{r=k}^{k'} Q_{r,T}(f_0) \Big)^p \Big] \\ &\leq \frac{1}{((k'-k+1)q)^{p/2}} \sum_{r_1, \dots, r_p = k}^{k'} \sum_{t_1=(2r_1-2)q+1}^{(2r_1-1)q \wedge T} \cdots \sum_{t_p=(2r_p-2)q+1}^{(2r_p-1)q \wedge T} \big| \mathbb{E}[w_{t_1,T} \dots w_{t_p,T}] \big| \end{aligned}$$

$$\leq \frac{p!}{((k'-k+1)q)^{p/2}} \sum_{\substack{t_1,\ldots,t_p=(2k-2)q+1 \\ t_1\leq\ldots\leq t_p}}^{(2k'-1)q\wedge T} \big|\mathbb{E}[w_{t_1,T}\ldots w_{t_p,T}]\big|.$$

Let $(t_1,\ldots,t_p)$ be a tuple of ordered indices, that is, $t_1 \leq \ldots \leq t_p$. We say that the index t_i has a neighbour if $|t_i - t_{i-1}| \leq C^* \log T$ or $|t_i - t_{i+1}| \leq C^* \log T$ for some large constant C^* to be specified later on. Moreover, t_i is said to have exactly one neighbour if either $|t_i - t_{i-1}| \leq C^* \log T$ and $|t_i - t_{i+1}| > C^* \log T$ or vice versa. Finally, we call (t_{i-1}, t_i) a pair of neighbours if $|t_i - t_{i-1}| \leq C^* \log T$. Now let $S_\leq$ denote the set of ordered tuples $(t_1,\ldots,t_p) \in \{(2k-2)q+1,\ldots,(2k'-1)q\wedge T\}^p$ such that each index t_i has a neighbour. In addition, let $S_>$ be the set of tuples such that at least one index does not have a neighbour. With this notation at hand, we can write

$$V_T = V_T^{\leq} + V_T^{>},$$

where for $\ell \in \{\leq, >\}$,

$$V_T^\ell = \frac{p!}{((k'-k+1)q)^{p/2}} \sum_{(t_1,\ldots,t_p)\in S_\ell} \big|\mathbb{E}[w_{t_1,T}\ldots w_{t_p,T}]\big|.$$

We now analyze the two terms $V_T^{\leq}$ and $V_T^{>}$ separately. For the investigation of $V_T^{\leq}$, define

$$S_{\leq,a} = \big\{(t_1,\ldots,t_p) \in S_\leq \mid \text{each index } t_i \text{ has exactly one neighbour}\big\}$$

together with

$$S_{\leq,b} = S_\leq \setminus S_{\leq,a}.$$

First suppose that $(t_1,\ldots,t_p) \in S_{\leq,a}$. In this case, there are exactly p pairs (t_{2i-1}, t_{2i}) of neighbours (recalling that p is even by assumption). Using Davydov's inequality (see e.g. Corollary 1.1 in Bosq [2]) to bound the covariances of the mixing variables $w_{t,T}$, we obtain that

$$\begin{aligned}
\big|\mathbb{E}[w_{t_1,T}\ldots w_{t_p,T}]\big| &\leq \big|\mathbb{E}[w_{t_1,T}w_{t_2,T}]\mathbb{E}[w_{t_3,T}\ldots w_{t_p,T}]\big| + \big|\mathrm{Cov}(w_{t_1,T}w_{t_2,T}, w_{t_3,T}\ldots w_{t_p,T})\big| \\
&= \big|\mathbb{E}[w_{t_1,T}w_{t_2,T}]\mathbb{E}[w_{t_3,T}\ldots w_{t_p,T}]\big| + O\big(\alpha(C^*\log T)\big) \\
&= \big|\mathrm{Cov}(w_{t_1,T}, w_{t_2,T})\mathbb{E}[w_{t_3,T}\ldots w_{t_p,T}]\big| + O\big(\alpha(C^*\log T)\big) \\
&\;\;\vdots \\
&\leq \Big|\prod_{i=1}^{p/2} \mathrm{Cov}(w_{t_{2i-1},T}, w_{t_{2i},T})\Big| + O(T^{-\nu}),
\end{aligned}$$

where we have used the fact that the mixing coefficients are decaying exponentially fast and the constant $\nu > 0$ can be made arbitrarily large (by choosing the constant

C^* sufficiently large). This implies that

$$\begin{aligned}
V_T^{\leq,a} &= \frac{p!}{((k'-k+1)q)^{p/2}} \sum_{(t_1,\dots,t_p)\in S_{\leq,a}} \big|\mathbb{E}[w_{t_1,T}\dots w_{t_p,T}]\big| \\
&\leq \frac{p!}{((k'-k+1)q)^{p/2}} \sum_{(t_1,\dots,t_p)\in S_{\leq,a}} \Big|\prod_{i=1}^{p/2} \mathrm{Cov}(w_{t_{2i-1},T}, w_{t_{2i},T})\Big| + o(1) \\
&\leq \frac{p!}{((k'-k+1)q)^{p/2}} \prod_{i=1}^{p/2}\Big(\sum_{\ell=0}^{\lceil C^*\log T\rceil} \sum_{t_{2i-1}=(2k-2)q+1}^{(2k'-1)q\wedge T} \big|\mathrm{Cov}(w_{t_{2i-1},T}, w_{t_{2i-1}+\ell,T})\big|\Big) + o(1) \\
&\leq C\frac{p!}{((k'-k+1)q)^{p/2}} ((k'-k+1)q)^{p/2}\Big(\sum_{\ell=0}^{\lceil C^*\log T\rceil} \alpha(\ell)\Big)^{p/2} + o(1) \leq C
\end{aligned}$$

for some sufficiently large constant C, where the last line again uses Davydov's inequality to bound the covariance expressions in the formula.
Next consider the sum $V_T^{\leq,b}$ corresponding to indices in the set $S_{\leq,b}$. The cardinality of this set is bounded by

$$C\big((k'-k+1)q\big)^{\frac{p}{2}-1}\big(\log T\big)^{\frac{p}{2}+1},$$

which implies

$$\begin{aligned}
V_T^{\leq,b} &= \frac{p!}{((k'-k+1)q)^{p/2}} \sum_{(t_1,\dots,t_p)\in S_{\leq,b}} \big|\mathbb{E}[w_{t_1,T}\dots w_{t_p,T}]\big| \\
&\leq C\frac{(\log T)^{p/2+1}}{(k'-k+1)q} = o(1)
\end{aligned}$$

(noting that $q = T^b$). This shows that the term $V_T^{\leq}$ is bounded.
Finally, we examine the term $V_T^{>}$ corresponding to the index set $S_{>}$. By definition, the tuples contained in this set have at least one element, say t_i, without a neighbour, that is, $|t_i - t_{i+1}| > C^*\log T$ and $|t_i - t_{i-1}| > C^*\log T$. Exploiting the mixing conditions on the model variables in a similar way as above, we obtain that

$$\begin{aligned}
&\mathbb{E}[w_{t_1,T}\dots w_{t_p,T}] \\
&= \mathbb{E}[w_{t_1,T}\dots w_{t_{i-1},T}]\mathbb{E}[w_{t_i,T}\dots w_{t_p,T}] + \mathrm{Cov}(w_{t_1,T}\dots w_{t_{i-1},T}, w_{t_i,T}\dots w_{t_p,T}) \\
&= \mathbb{E}[w_{t_1,T}\dots w_{t_{i-1},T}]\mathrm{Cov}(w_{t_i,T}, w_{t_{i+1},T}\dots w_{t_p,T}) + O(T^{-\nu}) \\
&= O(T^{-\nu}),
\end{aligned}$$

where ν can be chosen arbitrarily large (if C^* is chosen large enough). Recalling the definition of $V_T^{>}$, this yields that $V_T^{>} = o(1)$. Putting everything together, the quantity V_T is seen to be bounded. This completes the proof. □

Lemma A.2. *Let (C1) and (C2) be satisfied. Moreover, assume that for some even $p \in \mathbb{N}$ and some small $\delta > 0$,*

$$\mathbb{E}\Big[\Big|\frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f,f')}\Big|^{(1+\delta)p}\Big] \leq C$$

for all functions $f, f' \in \mathcal{F}$. Then for any $f_0 \in \mathcal{F}$,

$$\big\|W_T(k,k')\big\|_p \leq C\Big(\Big\|\sum_{r=k}^{k'} Q_{r,T}(f_0)\Big\|_p + \int_0^{\operatorname{diam}(\mathcal{F})} \mathcal{N}(w/2, \mathcal{F}, d_{\mathcal{F}})^{1/p} dw\Big),$$

where $\mathcal{N}(w, \mathcal{F}, d_{\mathcal{F}})$ is the covering number of $(\mathcal{F}, d_{\mathcal{F}})$ and $\operatorname{diam}(\mathcal{F}) = \sup_{f,f' \in \mathcal{F}} d_{\mathcal{F}}(f,f')$ denotes the diameter of $\mathcal{F}$.

Proof. The claim immediately follows from Theorem 2.2.4 and Corollary 2.2.5 in van der Vaart & Wellner [26] (see their remark on p.100 before Subsection 2.2.1). It thus suffices to verify the conditions of Theorem 2.2.4. In particular, we have to show that

$$\mathbb{E}\Big[\Big|\sum_{r=k}^{k'} Q_{r,T}(f) - \sum_{r=k}^{k'} Q_{r,T}(f')\Big|^p\Big] \leq C d_{\mathcal{F}}(f,f')^p$$

for some sufficiently large constant C. To prove this, we introduce the notation

$$w_{t,T} = \frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f,f')} - \mathbb{E}\Big[\frac{f(X_{t,T}) - f'(X_{t,T})}{d_{\mathcal{F}}(f,f')}\Big]$$

and consider

$$\begin{aligned}
V_T = V_T(k,k') &= \mathbb{E}\Big[\Big|\sum_{r=k}^{k'} \frac{Q_{r,T}(f) - Q_{r,T}(f')}{d_{\mathcal{F}}(f,f')}\Big|^p\Big] \\
&\leq \frac{1}{((k'-k+1)q)^{p/2}} \sum_{r_1,\ldots,r_p=k}^{k'} \sum_{t_1=(2r_1-2)q+1}^{(2r_1-1)q\wedge T} \cdots \sum_{t_p=(2r_p-2)q+1}^{(2r_p-1)q\wedge T} \big|\mathbb{E}[w_{t_1,T} \ldots w_{t_p,T}]\big| \\
&\leq \frac{p!}{((k'-k+1)q)^{p/2}} \sum_{\substack{t_1,\ldots,t_p=(2k-2)q+1 \\ t_1 \leq \cdots \leq t_p}}^{(2k'-1)q\wedge T} \big|\mathbb{E}[w_{t_1,T} \ldots w_{t_p,T}]\big|.
\end{aligned}$$

Repeating the arguments from Lemma A.1, we can show that V_T is bounded, thus completing the proof. □

Proof of Theorem 6.1

To show that $\hat{H}_T = \sqrt{T}[\hat{D}_T - D]$ weakly converges to H, it suffices to prove that

$$\hat{H}_T^c := \sqrt{T}\big[\hat{D}_T - \mathbb{E}\hat{D}_T\big] \rightsquigarrow H \tag{A.1}$$

together with

$$\sqrt{T} \sup_{(u,v,f) \in \Delta \times \mathcal{F}} |\mathbb{E}\hat{D}_T - D| = o(1), \tag{A.2}$$

where $\hat{H}_T^c$ is the centred version of $\hat{H}_T$. We start with the proof of (A.2). Making use of condition (C4), we obtain that

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=\lceil uT+1\rceil}^{T}\mathbb{E}\big[f(X_{t,T})\big] &= \frac{1}{\sqrt{T}}\sum_{t=\lceil uT+1\rceil}^{T}\mathbb{E}\Big[f\Big(X_t\Big(\frac{t}{T}\Big)\Big)\Big] + o(1)\\ &= \sqrt{T}\sum_{t=\lceil uT+1\rceil}^{T-1}\int_{\frac{t}{T}}^{\frac{t+1}{T}}\mathbb{E}\big[f(X_t(w))\big]dw + o(1)\\ &= \sqrt{T}\int_u^1\mathbb{E}\big[f(X_t(w))\big]dw + o(1)\end{aligned}$$

uniformly with respect to $u \in [0,1]$ and $f \in \mathcal{F}$. From this, (A.2) immediately follows. To verify (A.1), we show weak convergence of the finite dimensional distributions of $\hat{H}_T^c$ as well as stochastic equicontinuity of $\hat{H}_T^c$. In particular, we derive the following two results.

Proposition A.1. *For any finite number of points (u_i, v_i, f_i) with $1 \le i \le n$, it holds that*

$$\begin{pmatrix}\hat{H}_T^c(u_1,v_1,f_1)\\ \vdots\\ \hat{H}_T^c(u_n,v_n,f_n)\end{pmatrix} \xrightarrow{d} N(0,\Sigma)$$

where $\Sigma = (\Sigma_{ij})_{1\le i,j\le n}$ and $\Sigma_{ij} = \mathrm{Cov}(H(u_i,v_i,f_i), H(u_j,v_j,f_j))$.

Proposition A.2. *The sequence of processes $\hat{H}_T^c$ is asymptotically stochastically equicontinuous, that is, for any $\varepsilon > 0$,*

$$\lim_{\delta\searrow 0}\limsup_{T\to\infty}\mathbb{P}\Big(\sup_{\substack{|u-u'|+|v-v'|\\ +d_{\mathcal{F}}(f,f')\le\delta}}\big|\hat{H}_T^c(u,v,f) - \hat{H}_T^c(u',v',f')\big| > \varepsilon\Big) = 0.$$

Combining the above two propositions, (A.1) now follows from a standard functional central limit theorem (see van der Vaart & Wellner [26]).

Proof of Proposition A.1. We start by calculating the asymptotic expectation and covariances of the process $\hat{H}_T^c$. As the process is centered, it holds that $\mathbb{E}[\hat{H}_T^c(u,v,f)] = 0$. To calculate the asymptotic covariances, we introduce the notation

$$\hat{H}_T^c(u,v,f) = \hat{G}_T(v,f) - \Big(\frac{1-v}{1-u}\Big)\,\hat{G}_T(u,f) \tag{A.3}$$

together with

$$\hat{G}_T(u,f) = \frac{1}{\sqrt{T}}\sum_{t=\lceil uT+1\rceil}^{T}\big(f(X_{t,T}) - \mathbb{E}f(X_{t,T})\big). \tag{A.4}$$

With this, we can write

$$\begin{aligned}\mathrm{Cov}\big(\hat{H}_T^c(u_1,v_1,f_1),\hat{H}_T^c(u_2,v_2,f_2)\big) &= \frac{(1-v_1)(1-v_2)}{(1-u_1)(1-u_2)}\,\mathbb{E}\big[\hat{G}_T(u_1,f_1)\hat{G}_T(u_2,f_2)\big]\\ &\quad-\frac{1-v_2}{1-u_2}\,\mathbb{E}\big[\hat{G}_T(v_1,f_1)\hat{G}_T(u_2,f_2)\big]\\ &\quad-\frac{1-v_1}{1-u_1}\,\mathbb{E}\big[\hat{G}_T(u_1,f_1)\hat{G}_T(v_2,f_2)\big]\\ &\quad+\mathbb{E}\big[\hat{G}_T(v_1,f_1)\hat{G}_T(v_2,f_2)\big]. \qquad (A.5)\end{aligned}$$

In what follows, we show that

$$\mathbb{E}\big[\hat{G}_T(u_1,f_1)\hat{G}_T(u_2,f_2)\big] = \sum_{\ell=-\infty}^{\infty}\int_{\max\{u_1,u_2\}}^{1} c_\ell(w)dw + o(1) \qquad (A.6)$$

with $c_\ell(w) = \mathrm{Cov}(f_1(X_0(w)), f_2(X_\ell(w)))$. Plugging (A.6) into (A.5) yields

$$\mathrm{Cov}\big(\hat{H}_T^c(u_1,v_1,f_1),\hat{H}_T^c(u_2,v_2,f_2)\big) = \mathrm{Cov}\big(H(u_1,v_1,f_1),H(u_2,v_2,f_2)\big) + o(1).$$

Hence, the covariances of $\hat{H}_T^c$ converge to those of the Gaussian process H.
To show (A.6), we assume without loss of generality that $u_1 \geq u_2$. Exploiting the mixing conditions of (C2) by means of Davydov's inequality, it can be seen that $\mathrm{Cov}\big(f_1(X_{t,T}), f_2(X_{s,T})\big) \leq C\alpha(|s-t|) \leq Ca^{|s-t|}$ for some $a < 1$ and a sufficiently large constant C. We thus obtain that

$$\begin{aligned}&\mathbb{E}\big[\hat{G}_T(u_1,f_1)\hat{G}_T(u_2,f_2)\big]\\ &= \frac{1}{T}\sum_{t=\lceil u_1T+1\rceil}^{T}\sum_{s=\lceil u_2T+1\rceil}^{T}\mathrm{Cov}\big(f_1(X_{t,T}),f_2(X_{s,T})\big)\\ &= \frac{1}{T}\sum_{t=\lceil u_1T+1\rceil}^{T}\sum_{s=\lceil u_2T+1\rceil}^{T} I\{|s-t|\leq C^*\log T\}\mathrm{Cov}\big(f_1(X_{t,T}),f_2(X_{s,T})\big) + o(1)\\ &=: Q_T^{(1)} + Q_T^{(2)} + Q_T^{(3)} + o(1)\end{aligned}$$

for some sufficiently large constant C^*, where the random variables $Q_T^{(j)}$ $(j = 1, 2, 3)$ are defined by

$$Q_T^{(1)} = \frac{1}{T}\sum_{\ell=1}^{\lceil C^*\log T\rceil}\sum_{t=1}^{T-\ell} I\big\{t\geq\lceil u_1T+1\rceil, t+\ell\geq\lceil u_2T+1\rceil\big\}\mathrm{Cov}\big(f_1(X_{t,T}),f_2(X_{t+\ell,T})\big)$$

$$Q_T^{(2)} = \frac{1}{T}\sum_{t=\lceil u_1T+1\rceil}^{T}\mathrm{Cov}\big(f_1(X_{t,T}),f_2(X_{t,T})\big)$$

$$Q_T^{(3)} = \frac{1}{T}\sum_{\ell=1}^{\lceil C^*\log T\rceil}\sum_{t=\ell+1}^{T} I\big\{t\geq\lceil u_1T+1\rceil, t-\ell\geq\lceil u_2T+1\rceil\big\}\mathrm{Cov}\big(f_1(X_{t,T}),f_2(X_{t-\ell,T})\big).$$

By assumption (C4), it follows for $\ell \leq \lceil C^* \log T \rceil$ and any w with $|w - \frac{t}{T}| \leq \frac{1}{T}$ that

$$\begin{aligned} c_{t,T,\ell} &:= \mathrm{Cov}\big(f_1(X_{t,T}), f_2(X_{t+\ell,T})\big) \\ &= \mathrm{Cov}\Big(f_1\Big(X_t\Big(\frac{t}{T}\Big)\Big), f_2\Big(X_{t+\ell}\Big(\frac{t+\ell}{T}\Big)\Big)\Big) + O\Big(\frac{\log T}{T}\Big) \\ &= \mathrm{Cov}\Big(f_1\Big(X_t\Big(\frac{t}{T}\Big)\Big), f_2\Big(X_{t+\ell}\Big(\frac{t}{T}\Big)\Big)\Big) + O\Big(\frac{\log T}{T}\Big) \\ &= \mathrm{Cov}\big(f_1(X_0(w)), f_2(X_\ell(w))\big) + O\Big(\frac{\log T}{T}\Big) \\ &=: c_\ell(w) + O\Big(\frac{\log T}{T}\Big), \end{aligned}$$

the last line defining $c_\ell(w)$ in an obvious manner. From this, it is easy to see that

$$\begin{aligned} \frac{1}{T} \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{T-\ell} |c_{t,T,\ell}| &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{T-\ell} \int_{\frac{t-1}{T}}^{\frac{t}{T}} \Big|c_\ell\Big(\frac{t}{T}\Big)\Big| dw + O\Big(\frac{(\log T)^2}{T}\Big) \\ &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=1}^{T-\ell} \int_{\frac{t-1}{T}}^{\frac{t}{T}} |c_\ell(w)| dw + O\Big(\frac{(\log T)^2}{T}\Big) \\ &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \int_0^1 |c_\ell(w)| dw + O\Big(\frac{(\log T)^2}{T}\Big). \end{aligned}$$

Because of the mixing assumption (C2), the left-hand side of this equation is bounded as $T \to \infty$ and consequently $\sum_{\ell=1}^{\infty} \int_0^1 c_\ell(w) dw$ is absolutely convergent. Therefore we obtain for the term $Q_T^{(1)}$ as $T \to \infty$ (recall that $u_1 \geq u_2$)

$$\begin{aligned} Q_T^{(1)} &= \sum_{\ell=1}^{\lceil C^* \log T \rceil} \sum_{t=\lceil u_1 T + 1 \rceil}^{T-\ell} \int_{\frac{t-1}{T}}^{\frac{t}{T}} c_\ell(w) dw + O\Big(\frac{(\log T)^2}{T}\Big) \\ &= \sum_{\ell=1}^{\infty} \int_{u_1}^1 c_\ell(w) dw + O\Big(\frac{(\log T)^2}{T}\Big) \end{aligned}$$

and similarly

$$Q_T^{(2)} = \int_{u_1}^1 c_0(w) dw + O\Big(\frac{(\log T)^2}{T}\Big), \quad Q_T^{(3)} = \sum_{\ell=1}^{\infty} \int_{u_1}^1 c_{-\ell}(w) dw + O\Big(\frac{(\log T)^2}{T}\Big).$$

Putting everything together, we arrive at (A.6).

Having calculated the asymptotic covariance structure of $\hat{H}_T^c$, we now apply a central limit theorem for mixing arrays of random variables (see e.g. Liebscher [16]) together with the Cramér-Wold device to obtain weak convergence of the finite dimensional distributions. □

Proof of Proposition A.2. As for the proof of Proposition A.1, we make use of the notation (A.3) and (A.4). Straightforward calculations show that

$$\begin{aligned}\sup_{\substack{|u-u'|+|v-v'|\\+d_{\mathcal{F}}(f,f')\le\delta}} \big|\hat{H}_T^c(u,v,f)-\hat{H}_T^c(u',v',f')\big| &\le 2\sup_{\substack{|u-u'|\le\delta\\f\in\mathcal{F}}}\big|\hat{G}_T(u,f)-\hat{G}_T(u',f)\big|\\ &\quad+2\sup_{\substack{d_{\mathcal{F}}(f,f')\le\delta\\u\in[0,1]}}\big|\hat{G}_T(u,f)-\hat{G}_T(u,f')\big|\\ &\quad+2\sqrt{\delta}\sup_{\substack{u\in[0,1]\\f\in\mathcal{F}}}\Big|\frac{1}{\sqrt{1-u}}\hat{G}_T(u,f)\Big|.\end{aligned}$$

Therefore, stochastic equicontinuity follows from the statements

$$\lim_{\delta\searrow 0}\limsup_{T\to\infty}\mathbb{P}\Big(\sup_{\substack{|u-u'|\le\delta\\f\in\mathcal{F}}}\Big|\hat{G}_T(u,f)-\hat{G}_T(u',f)\Big|>\varepsilon\Big)=0 \tag{A.7}$$

$$\lim_{\delta\searrow 0}\limsup_{T\to\infty}\mathbb{P}\Big(\sup_{\substack{d_{\mathcal{F}}(f,f')\le\delta\\u\in[0,1]}}\Big|\hat{G}_T(u,f)-\hat{G}_T(u,f')\Big|>\varepsilon\Big)=0 \tag{A.8}$$

$$\lim_{\delta\searrow 0}\limsup_{T\to\infty}\mathbb{P}\Big(\sqrt{\delta}\sup_{\substack{u\in[0,1]\\f\in\mathcal{F}}}\Big|\frac{1}{\sqrt{1-u}}\hat{G}_T(u,f)\Big|>\varepsilon\Big)=0. \tag{A.9}$$

As the three statements can be shown by similar arguments, we restrict ourselves to the proof of (A.7).

First of all, observe that for any function $g:[0,1]\to\mathbb{R}$, the inequality

$$\begin{aligned}\sup_{\substack{|u-u'|\le\delta\\u,u'\in[0,1]}}|g(u)-g(u')| &\le \max_{j=1,\dots,\lceil 1/\delta\rceil}\sup_{u\in[u_{j-1},u_j]}|g(u)-g(u_j)|\\ &\quad+\max_{j=1,\dots,\lceil 1/\delta\rceil}\sup_{u'\in[u_{j-2},u_{j+1}]}|g(u')-g(u_j)|\end{aligned}$$

holds, where $u_{-1}=u_0=0$, $u_j=j\delta$ $(j=1,\dots,\lceil 1/\delta\rceil-1)$ and $u_{\lceil 1/\delta\rceil}=u_{\lceil 1/\delta\rceil+1}=1$. This implies that (A.7) is a consequence of

$$\lim_{\delta\searrow 0}\limsup_{T\to\infty}\mathbb{P}\Big(\max_{j=1,\dots,\lceil 1/\delta\rceil}\sup_{u\in[u_{j-1},u_j]}\sup_{f\in\mathcal{F}}\Big|\hat{G}_T(u,f)-\hat{G}_T(j\delta,f)\Big|>\varepsilon\Big)=0. \tag{A.10}$$

In the sequel, we derive a suitable bound for the probability

$$P_T(\delta,\varepsilon)=\mathbb{P}\Big(\max_{j=1,\dots,\lceil 1/\delta\rceil}\sup_{u\in[u_{j-1},u_j]}\sup_{f\in\mathcal{F}}\Big|\hat{G}_T(u,f)-\hat{G}_T(j\delta,f)\Big|>\varepsilon\Big)$$

in (A.10). To start with, we crudely bound this probability by

$$P_T(\delta,\varepsilon)\le\sum_{j=1}^{\lceil 1/\delta\rceil}P_{T,j}(\delta,\varepsilon),$$

where

$$\begin{aligned}P_{T,j}(\delta,\varepsilon)&=\mathbb{P}\Big(\sup_{u\in[u_{j-1},u_j]}\sup_{f\in\mathcal{F}}\Big|\hat{G}_T(u,f)-\hat{G}_T(j\delta,f)\Big|>\varepsilon\Big)\\ &=\mathbb{P}\Big(\max_{(j-1)\delta T\le\ell\le j\delta T}\sup_{f\in\mathcal{F}}\Big|\hat{G}_T\Big(\frac{\ell}{T},f\Big)-\hat{G}_T(j\delta,f)\Big|>\varepsilon\Big).\end{aligned}$$

To bound the probabilities $P_{T,j}(\delta,\varepsilon)$, we write

$$\hat{G}_T\Big(\frac{\ell}{T},f\Big)-\hat{G}_T(j\delta,f)=B_T^{\ell+}(f)+\sum_{r=\lceil\frac{\ell}{q}\rceil+1}^{\lfloor\frac{j\delta T}{q}\rfloor}B_{r,T}(f)+B_T^{j-}(f).$$

Here, $B_{r,T}(f)$ are blocks of length q given by

$$B_{r,T}(f)=\frac{1}{\sqrt{T}}\sum_{t=(r-1)q+1}^{rq}\big(f(X_{t,T})-\mathbb{E}f(X_{t,T})\big),$$

where as in the subsection on auxiliary results, we set $q=CT^b$ for some small $b>0$ (specifically, $b<\frac{1}{4}$). In addition,

$$B_T^{\ell+}(f)=\frac{1}{\sqrt{T}}\sum_{t=\ell+1}^{\lceil\frac{\ell}{q}\rceil q}\big(f(X_{t,T})-\mathbb{E}f(X_{t,T})\big)$$
$$B_T^{j-}(f)=\frac{1}{\sqrt{T}}\sum_{t=\lfloor\frac{j\delta T}{q}\rfloor q+1}^{\lceil j\delta T\rceil}\big(f(X_{t,T})-\mathbb{E}f(X_{t,T})\big)$$

denote the first and the last block, respectively. With this notation at hand, we obtain

$$\begin{aligned}P_{T,j}(\delta,6\varepsilon)&\le\mathbb{P}\Big(\max_{(j-1)\delta T\le\ell\le j\delta T}\sup_{f\in\mathcal{F}}\Big|\sum_{r=\lceil\frac{\ell}{q}\rceil+1}^{\lfloor\frac{j\delta T}{q}\rfloor}B_{r,T}(f)\Big|>4\varepsilon\Big)\\&\quad+\mathbb{P}\Big(\max_{(j-1)\delta T\le\ell\le j\delta T}\sup_{f\in\mathcal{F}}|B_T^{\ell+}(f)|>\varepsilon\Big)+\mathbb{P}\Big(\sup_{f\in\mathcal{F}}|B_T^{j-}(f)|>\varepsilon\Big)\\&=:P_{T,j,1}(\delta,4\varepsilon)+P_{T,j,2}(\delta,\varepsilon)+P_{T,j,3}(\delta,\varepsilon).\end{aligned}$$

The terms $P_{T,j,2}$ and $P_{T,j,3}$ can be bounded by fairly straightforward arguments: Applying a maximal inequality (see e.g. Section 2.1.3 in van der Vaart & Wellner [26]), we get that

$$\Big\|\max_{(j-1)\delta T\le\ell\le j\delta T}\sup_{f\in\mathcal{F}}|B_T^{\ell+}(f)|\Big\|_p\le C(\delta T)^{1/p}\max_{(j-1)\delta T\le\ell\le j\delta T}\big\|\sup_{f\in\mathcal{F}}|B_T^{\ell+}(f)|\big\|_p.$$

Moreover,

$$\sup_{f\in\mathcal{F}}|B_T^{\ell+}(f)|\le\frac{2}{\sqrt{T}}\sum_{t=\ell+1}^{\lceil\frac{\ell}{q}\rceil q}F(X_{t,T})$$

and by the moment conditions on the envelope F in (C3), $\|\sup_{f\in\mathcal{F}}|B_T^{\ell+}(f)|\|_p\le Cq/\sqrt{T}$. Hence by Markov's inequality,

$$P_{T,j,2}(\delta,\varepsilon)\le\varepsilon^{-p}\Big\|\max_{(j-1)\delta T\le\ell\le j\delta T}\sup_{f\in\mathcal{F}}|B_T^{\ell+}(f)|\Big\|_p^p\le C\delta T\Big(\frac{q}{\varepsilon\sqrt{T}}\Big)^p=o(1)$$

for $T \to \infty$ given that $q = T^b$ with $b < \frac{1}{4}$. By similar considerations, $P_{T,j,3}(\delta, \varepsilon)$ is seen to converge to zero as well. To deal with $P_{T,j,1}$, we split it up into two parts:

$$P_{T,j,1}(\delta, 4\varepsilon) \leq \Delta_T^{(0)} + \Delta_T^{(1)}$$

with

$$\Delta_T^{(0)} = \mathbb{P}\Big(\max_{\lfloor \frac{(j-1)\delta T}{2q} \rfloor \leq k \leq \lceil \frac{j\delta T}{2q} \rceil} \sup_{f \in \mathcal{F}} \Big| \sum_{r=k}^{\lfloor \frac{j\delta T}{2q} \rfloor} B_{2r,T}(f) \Big| > 2\varepsilon \Big)$$

$$\Delta_T^{(1)} = \mathbb{P}\Big(\max_{\lfloor \frac{(j-1)\delta T}{2q} \rfloor \leq k \leq \lceil \frac{j\delta T}{2q} \rceil} \sup_{f \in \mathcal{F}} \Big| \sum_{r=k}^{\lceil \frac{j\delta T}{2q} \rceil} B_{2r-1,T}(f) \Big| > 2\varepsilon \Big).$$

As the two terms can be treated in the same way, we restrict ourselves to $\Delta_T^{(1)}$. Applying a version of Ottaviani's inequality for α-mixing processes (which has the form stated in Chapter 10.2 of Lin & Bai [17] and can be proven by the arguments therein), we obtain that

$$\Delta_T^{(1)} \leq \frac{\mathbb{P}\Big(\sup_{f \in \mathcal{F}} \Big| \sum_{r=\lfloor \frac{(j-1)\delta T}{2q} \rfloor}^{\lceil \frac{j\delta T}{2q} \rceil} B_{2r-1,T}(f) \Big| > \varepsilon \Big) + \frac{\delta T}{2q} \alpha(q)}{1 - \max_{\lfloor \frac{(j-1)\delta T}{2q} \rfloor \leq k \leq \lceil \frac{j\delta T}{2q} \rceil} \mathbb{P}\Big(\sup_{f \in \mathcal{F}} \Big| \sum_{r=\lfloor \frac{(j-1)\delta T}{2q} \rfloor}^{k} B_{2r-1,T}(f) \Big| > \varepsilon \Big)}. \tag{A.11}$$

In order to bound the right-hand side of (A.11), we make use of the random variables

$$Q_{r,T}(f) = \frac{1}{\sqrt{(k'-k+1)q}} \sum_{t=(2r-2)q+1}^{(2r-1)q \wedge T} \big(f(X_{t,T}) - \mathbb{E} f(X_{t,T}) \big)$$

and

$$W_T(k, k') = \sup_{f \in \mathcal{F}} \Big| \sum_{r=k}^{k'} Q_{r,T}(f) \Big|,$$

which have been introduced at the beginning of the appendix. Combining Lemmas A.1 and A.2 and noting that the integral $\int_0^{\mathrm{diam}(\mathcal{F})} \mathcal{N}(w/2, \mathcal{F}, d)^{1/p} dw$ is finite by assumption (C3), we get that

$$\mathbb{E}\big[|W_T(k, k')|^p \big] \leq C < \infty \tag{A.12}$$

for some sufficiently large constant C. This implies that

$$\begin{aligned} \mathbb{P}\Big(\sup_{f \in \mathcal{F}} \Big| \sum_{r=k}^{k'} B_{2r-1,T}(f) \Big| > \varepsilon \Big) &= \mathbb{P}\Big(W_T(k, k') > \frac{\varepsilon \sqrt{T}}{\sqrt{(k'-k+1)q}} \Big) \\ &\leq \mathbb{E}\big[|W_T(k, k')|^p \big] \Big(\frac{(k'-k+1)q}{\varepsilon^2 T} \Big)^{p/2} \\ &\leq C \Big(\frac{(k'-k+1)q}{\varepsilon^2 T} \Big)^{p/2}. \end{aligned}$$

Specifically, whenever $(k - k' + 1)q \leq \delta T$,

$$\mathbb{P}\Big(\sup_{f\in\mathcal{F}}\Big|\sum_{r=k}^{k'} B_{2r-1,T}(f)\Big| > \varepsilon\Big) \leq C\frac{\delta^{p/2}}{\varepsilon^p}. \tag{A.13}$$

With (A.13), it is easy to see that the denominator in (A.11) is bounded away from zero as $T \to \infty$ and to infer that

$$\Delta_T^{(1)} \leq C\Big(\frac{\delta^{p/2}}{\varepsilon^p} + \frac{\delta T}{2q}\alpha(q)\Big).$$

Using an analogous bound for the term $\Delta_T^{(0)}$, it follows that

$$P_T(\delta,\varepsilon) \leq \sum_{j=1}^{\lceil 1/\delta\rceil} P_{T,j}(\delta,\varepsilon) \leq C\Big\lceil\frac{1}{\delta}\Big\rceil\Big(\frac{\delta^{p/2}}{\varepsilon^p} + \frac{\delta T}{2q}\alpha(q)\Big).$$

This yields that

$$\lim_{\delta\searrow 0}\limsup_{T\to\infty} P_T(\delta,\varepsilon) = 0$$

and the assertion (A.10) follows. By the discussion at the beginning of this proof we obtain (A.7), which implies stochastic equicontinuity. □

Proof of Theorem 6.2

Note that the operator $\mathcal{L} : \ell_\infty(\Delta\times\mathcal{F}) \to \ell_\infty([0,1])$ defined by

$$\mathcal{L}(g)(u) = \sup_{f\in\mathcal{F}, v\in[u,1]} |g(u,v,f)|$$

is continuous. Applying the continuous mapping theorem thus yields the result. □

Proof of Theorem 6.3

To start with, we introduce some notation. Recall that the criterion function $\hat{Q}_T(u)$ has the form $\hat{Q}_T(u) = u + (1-u)\hat{q}_T(u)$ with $\hat{q}_T(u) = \Phi(\rho_T\sqrt{T}\hat{\mathcal{D}}_T(u))$. Additionally, we define

$$Q_T(u) = u + (1-u)q_T(u)$$

with $q_T(u) = \Phi(\rho_T\sqrt{T}\mathcal{D}(u))$ and let

$$u_{0,T} = \operatorname*{argmin}_{u\in[0,1]} Q_T(u).$$

In the sequel, we prove that

$$|u_{0,T} - u_0| = O\big(\rho_T\sqrt{T}\big)^{-1/k} \tag{A.14}$$

$$|\hat{u}_0 - u_{0,T}| = O_p(\gamma_T). \tag{A.15}$$

Combining (A.14) and (A.15) completes the proof.

Proof of (A.14). It suffices to verify that for large sample sizes T,

$$(1 - C_1\nu_T)u_0 \leq u_{0,T} \leq u_0, \tag{A.16}$$

where $\nu_T = (\rho_T\sqrt{T})^{-1/k}$ and C_1 is a large positive constant. Since $Q_T(u) = u > u_0 = Q_T(u_0)$ for any $u > u_0$, it directly follows that $u_{0,T} \leq u_0$. To prove that $(1 - C_1\nu_T)u_0 \leq u_{0,T}$, we verify that $Q_T(u) > Q_T(u_0)$ for any $u < (1 - C_1\nu_T)u_0$, which is equivalent to

$$(1 - u)q_T(u) > u_0 - u \quad \text{for } u < (1 - C_1\nu_T)u_0. \tag{A.17}$$

(A.17) can be seen as follows: To start with, notice that

$$\min_{u\in[0,(1-C_1\nu_T)u_0]} \mathcal{D}(u) \geq \frac{c_k}{2}(C_1\nu_T u_0)^k$$

for sufficiently large T, which easily follows upon inspection of (6.3). This implies that

$$q_T(u) \geq \Phi\Big(\rho_T\sqrt{T} \min_{u\in[0,(1-C_1\nu_T)u_0]} \mathcal{D}(u)\Big) \geq \Phi\Big(\frac{c_k C_1^k u_0^k}{2}\Big)$$

for any $u \leq (1 - C_1\nu_T)u_0$. Choosing C_1 sufficiently large, we further obtain that $\Phi(c_k C_1^k u_0^k/2) > (1 - \delta)$ for an arbitrarily small $\delta > 0$. As a result, $(1 - u)q_T(u) > (1 - u)(1 - \delta) > (1 - \delta) - u > u_0 - u$ given that $u_0 < 1$ and δ is sufficiently small. This yields (A.17). □

Proof of (A.15). As a first step, we verify the following fact: There exists a (small) positive constant κ such that

$$|u - u_{0,T}| > M\gamma_T \;\Rightarrow\; |Q_T(u) - Q_T(u_{0,T})| > \kappa M\gamma_T \tag{A.18}$$

for sufficiently large constants M and samples sizes T. To show this, first suppose that $|u - u_{0,T}| > M\gamma_T$ and $u < u_{0,T}$. Repeating the arguments from the proof of (A.14), we get that $q_T(u) > (1 - \delta)$ for an arbitrarily small $\delta > 0$ provided that M is sufficiently large. Hence,

$$\begin{aligned} Q_T(u) - Q_T(u_{0,T}) &\geq Q_T(u) - Q_T(u_0) \\ &= (u - u_0) + (1 - u)q_T(u) \\ &\geq (u - u_0) + (1 - u)(1 - \delta) \\ &\geq (1 - \delta) - u_0. \end{aligned}$$

Recalling that $u_0 < 1$ and setting $\kappa = \frac{(1-\delta)-u_0}{u_0}$, we now arrive at

$$Q_T(u) - Q_T(u_{0,T}) \geq \kappa u_0 \geq \kappa|u - u_0| \geq \kappa|u - u_{0,T}| > \kappa M\gamma_T.$$

We next turn to the case with $u > u_{0,T}$. From (A.14), we know that $u_0 - u_{0,T} \leq C_1\nu_T \leq \frac{M}{2}\gamma_T$ provided that M is chosen sufficiently large, in particular $M > 2C_1$. Since $u - u_{0,T} > M\gamma_T$, we can infer that $u > u_0$ and thus

$$\begin{aligned} Q_T(u) - Q_T(u_{0,T}) \geq Q_T(u) - Q_T(u_0) &= (u - u_{0,T}) + (u_{0,T} - u_0) \\ &> M\gamma_T - \frac{M}{2}\gamma_T = \frac{M}{2}\gamma_T. \end{aligned}$$

This completes the proof of (A.18).
In the next step, we apply (A.18) to get

$$\mathbb{P}\Big(|\hat{u}_0 - u_{0,T}| > M\gamma_T\Big) \leq \mathbb{P}\Big(|Q_T(\hat{u}_0) - Q_T(u_{0,T})| > \kappa M\gamma_T\Big).$$

Since

$$\begin{aligned} |Q_T(\hat{u}_0) - Q_T(u_{0,T})| &= Q_T(\hat{u}_0) - Q_T(u_{0,T}) \\ &= \big[Q_T(\hat{u}_0) - \hat{Q}_T(\hat{u}_0)\big] + \big[\hat{Q}_T(\hat{u}_0) - Q_T(u_{0,T})\big] \\ &\leq \big[Q_T(\hat{u}_0) - \hat{Q}_T(\hat{u}_0)\big] + \big[\hat{Q}_T(u_{0,T}) - Q_T(u_{0,T})\big] \\ &\leq 2 \sup_{u\in[0,1]} |\hat{Q}_T(u) - Q_T(u)|, \end{aligned}$$

we can further conclude that

$$\mathbb{P}\Big(|\hat{u}_0 - u_{0,T}| > M\gamma_T\Big) \leq \mathbb{P}\Big(\sup_{u\in[0,1]} |\hat{Q}_T(u) - Q_T(u)| > \frac{\kappa M\gamma_T}{2}\Big).$$

To complete the proof, we show that for any given $\varepsilon > 0$, we can choose the constant M large enough to obtain

$$\mathbb{P}\Big(\sup_{u\in[0,1]} |\hat{Q}_T(u) - Q_T(u)| > \frac{\kappa M\gamma_T}{2}\Big) \leq \varepsilon. \tag{A.19}$$

To see this, note that

$$\sqrt{T}\mathcal{D}(u) - \hat{\mathcal{H}}_T(u) \leq \sqrt{T}\hat{\mathcal{D}}_T(u) \leq \sqrt{T}\mathcal{D}(u) + \hat{\mathcal{H}}_T(u).$$

Since Φ is Lipschitz, this implies that

$$\begin{aligned} \big|\hat{Q}_T(u) - Q_T(u)\big| &\leq (1-u)\big|\hat{q}_T(u) - q_T(u)\big| \\ &\leq C\rho_T\big|\sqrt{T}\hat{\mathcal{D}}_T(u) - \sqrt{T}\mathcal{D}(u)\big| \\ &\leq C\rho_T\hat{\mathcal{H}}_T(u) \leq C\rho_T \sup_{u\in[0,1]} |\hat{\mathcal{H}}_T(u)|. \end{aligned}$$

As $\hat{\mathcal{H}}_T$ weakly converges in $\ell_\infty([0,1])$, we know that for any $\varepsilon > 0$, we can find a constant $C_\varepsilon > 0$ with

$$\limsup_{T\to\infty} \mathbb{P}\Big(\sup_{u\in[0,1]} |\hat{\mathcal{H}}_T(u)| > C_\varepsilon\Big) \leq \varepsilon.$$

Hence, for any given $\varepsilon > 0$,

$$\mathbb{P}\Big(\sup_{u\in[0,1]} \big|\hat{Q}_T(u) - Q_T(u)\big| > \frac{\kappa M \gamma_T}{2}\Big) \leq \mathbb{P}\Big(C\rho_T \sup_{u\in[0,1]} |\hat{\mathcal{H}}_T(u)| > \frac{\kappa M \gamma_T}{2}\Big)$$
$$\leq \mathbb{P}\Big(C \sup_{u\in[0,1]} |\hat{\mathcal{H}}_T(u)| > \frac{\kappa M}{2}\Big) \leq \varepsilon,$$

provided that M is chosen sufficiently large. □

Proof of Corollary 7.1

The proof follows by slightly modifying the arguments for Theorem 6.3. To start with, define

$$\hat{\mathcal{D}}_T^{\text{sc}}(u) = \sup_{f\in\mathcal{F}, v\in[u,1]} \Big| \frac{\hat{D}_T(u,v,f)}{\hat{\sigma}_f(u)\sqrt{1-u}} \Big|$$
$$\mathcal{D}_T^{\text{sc}}(u) = \sup_{f\in\mathcal{F}, v\in[u,1]} \Big| \frac{D(u,v,f)}{\hat{\sigma}_f(u)\sqrt{1-u}} \Big|$$

together with

$$\hat{Q}_T^{\text{sc}}(u) = u + (1-u)\hat{q}_T^{\text{sc}}(u)$$
$$Q_T^{\text{sc}}(u) = u + (1-u)q_T^{\text{sc}}(u),$$

where $\hat{q}_T^{\text{sc}}(u) = \Psi(\rho_T\sqrt{T}\hat{\mathcal{D}}_T^{\text{sc}}(u))$ and $q_T^{\text{sc}}(u) = \Psi(\rho_T\sqrt{T}\mathcal{D}_T^{\text{sc}}(u))$. Moreover, let $\hat{u}_0^{\text{sc}}$ and $u_{0,T}^{\text{sc}}$ be the minimizers of the two criterion functions $\hat{Q}_T^{\text{sc}}$ and Q_T^{sc}, respectively. Analogously to the proof of Theorem 6.3, we show that

$$|u_{0,T}^{\text{sc}} - u_0| = O_p\big(\rho_T\sqrt{T}\big)^{-1/k} \tag{A.20}$$
$$|\hat{u}_0^{\text{sc}} - u_{0,T}^{\text{sc}}| = O_p(\gamma_T). \tag{A.21}$$

The following remarks are helpful for verifying (A.20) and (A.21):

(i) Let E_σ denote the event that

$$\underline{\sigma} \leq \hat{\sigma}_f(u) \leq \overline{\sigma} \text{ for all } u \in [0,1] \text{ and } f \in \mathcal{F}.$$

Since by assumption, the event E_σ occurs with probability approaching one, we can restrict attention to E_σ throughout the proof. In particular, it is sufficient to show that (A.20) and (A.21) hold true on the event E_σ.

(ii) Since $u_0 < 1$ by assumption and $D(u,v,f) = 0$ for any $u \geq u_0$, it holds that $\mathcal{D}_T^{\text{sc}}(u) = 0$ for any $u \geq u_0$ on the event E_σ. Moreover, $\mathcal{D}_T^{\text{sc}}$ has the same degree of smoothness at the point u_0 as the non-scaled version $\mathcal{D}$.

Keeping the above remarks in mind, (A.20) and (A.21) follow by the same arguments as in the proof of Theorem 6.3. □

References

[1] Andersen, T., Bollerslev, T., Diebold, F. and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, **71** 579–625.

[2] Bosq, D. (1996). *Nonparametric statistics for stochastic processes.* New York, Springer.

[3] Carlstein, E. (1988). Nonparametric change-point estimation. *Annals of Statistics*, **16** 188–197.

[4] Chen, Y., Härdle, W. K. and Pigorsch, U. (2010). Localized realized volatility modeling. *Journal of the American Statistical Association*, **105** 1376–1393.

[5] Dahlhaus, R. (1996). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Processes and their Applications*, **62** 139–168.

[6] Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Annals of Statistics*, **25** 1–37.

[7] Dahlhaus, R. and Subba Rao, S. (2006). Statistical inference for time-varying ARCH processes. *Annals of Statistics*, **34** 1075–1114.

[8] Dette, H., Preuss, P. and Vetter, M. (2011). A measure of stationarity in locally stationary processes with applications to testing. *Journal of the American Statistical Association*, **106** 1113–1124.

[9] Drees, H. and Stărică, C. (2003). A simple non-stationary model for stock returns. *Preprint.*

[10] Dwivedi, Y. and Subba Rao, S. (2010). A test for second-order stationarity of a time series based on the discrete fourier transform. *Journal of Time Series Analysis*, **32** 68–91.

[11] Fryzlewicz, P., Sapatinas, T. and Subba Rao, S. (2006). A Haar-Fisz technique for locally stationary volatility estimation. *Biometrika*, **93** 687–704.

[12] Fryzlewicz, P. and Subba Rao, S. (2011). Mixing properties of ARCH and time-varying ARCH processes. *Bernoulli*, **17** 320–346.

[13] Heber, G., Lunde, A., Shephard, N. and Sheppard, K. (2009). *Oxford-Man Institute's realized library. Library version 0.2.* Oxford-Man Institute, University of Oxford.

[14] Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57** 1–17.

[15] Koo, B. and Linton, O. (2012). Estimation of semiparametric locally stationary diffusion models. *Journal of Econometrics*, **170** 210–233.

[16] Liebscher, E. (1996). Central limit theorems for sums of α-mixing random variables. *Stochastics and Stochastic Reports*, **59** 241–258.

[17] Lin, Z. and Bai, Z. (2010). *Probability inequalities*. New York, Springer.

[18] Mallik, A., Banerjee, M. and Sen, B. (2013). Asymptotics for p-value based threshold estimation in regression settings. *Preprint*.

[19] Mallik, A., Sen, B., Banerjee, M. and Michailidis, G. (2011). Threshold estimation based on a p-value framework in dose-response and regression settings. *Biometrika*, **98** 887–900.

[20] Mikosch, T. and Stărică, C. (2004). Non-stationarities in financial time series, the long-range dependence, and the IGARCH effects. *Review of Economics and Statistics*, **86** 378–390.

[21] Müller, H.-G. (1992). Change-points in nonparametric regression analysis. *Annals of Statistics*, **20** 737761.

[22] Paparoditis, E. (2009). Testing temporal constancy of the spectral structure of a time series. *Bernoulli*, **15** 1190–1221.

[23] Paparoditis, E. (2010). Validating stationarity assumptions in time series analysis by rolling local periodograms. *Journal of the American Statistical Association*, **105** 839–851.

[24] Preuss, P., Vetter, M. and Dette, H. (2012). A test for stationarity based on empirical processes. *Forthcoming in Bernoulli*.

[25] Subba Rao, S. (2006). On some nonstationary, nonlinear random processes and their stationary approximations. *Advances in Applied Probability*, **38** 1155–1172.

[26] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York, Springer.

[27] Vogt, M. (2012). Nonparametric regression for locally stationary time series. *Annals of Statistics*, **40** 2601–2633.

[28] von Sachs, R. and Neumann, M. H. (2000). A wavelet-based test for stationarity. *Journal of Time Series Analysis*, **21** 597–613.

[29] Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, **73** 91–104.

[30] Wu, J. S. and Chu, C. K. (1993). Kernel-type estimators of jump points and values of a regression function. *Annals of Statistics*, **21** 1545–1566.