Inference on breaks in weak location time series models with quasi-Fisher scores

Francq, Christian and Trapani, Lorenzo and Zakoian, Jean-Michel

ENSAE-CREST and Lille University, University of Leicester, ENSAE-CREST and Lille University

2025

Inference on breaks in weak location time series models with quasi-Fisher scores

Christian Francq[*], Lorenzo Trapani[†] and Jean-Michel Zakoïan[‡ §]

Abstract

Based on Godambe's theory of estimating functions, we propose a class of cumulative sum (CUSUM) statistics to detect breaks in the dynamics of time series under weak assumptions. First, we assume a parametric form for the conditional mean, but make no specific assumption about the data-generating process (DGP) or even about the other conditional moments. The CUSUM statistics we consider depend on a sequence of weights that influence their asymptotic accuracy. Data-driven procedures are proposed for the optimal choice of the sequence of weights, in Godambe's sense. We also propose modified versions of the tests that allow to detect breaks in the dynamics even when the conditional mean is misspecified. Our results are illustrated using Monte Carlo experiments and real financial data.

JEL Classification: C12, C13, C52 and C58

Keywords: Break detection in the conditional mean, Change-points, CUSUM, Estimating functions, Quasi-likelihood estimator.

[*]CREST-ENSAE and Lille University. Email: christian.francq@univ-lille3.fr

[†]University of Leicester and Universita' di Pavia. Email: lt285@leicester.ac.uk

[‡]CREST-ENSAE and Lille University. Email: zakoian@ensae.fr

[§]Corresponding author

1 Introduction

In this paper, we study changepoint detection in the context of inference based on the Estimating Functions (EF) approach. The EF approach was originally developed by Godambe (1960) for fully parametric estimation and by Durbin (1960) for estimating a simple autoregressive structure. Several contributions have been developed thereafter, extending the application of EF to more general, non independent and identically distributed (i.i.d.), stochastic processes; whilst a complete review of the literature is beyond the scope of this paper, we refer, *inter alia*, to the papers by Godambe (1985), Godambe & Heyde (1987), Jacod & Sørensen (2018), and Francq & Zakoïan (2023); the reviews by Bera & Bilias (2002), Bera et al. (2006), and Heyde (1997) offer an excellent description of the state of the art on EF-based inference.

In essence, the EF approach is based on estimating a finite dimensional parameter, say $\boldsymbol{\theta}$, by solving the equation $H_n(\boldsymbol{\theta}) = 0$, for some observation-dependent function $H_n(\boldsymbol{\theta})$. A common name for the resulting estimator is Quasi Likelihood Estimator (QLE). Clearly, several popular estimation techniques can be cast in the EF framework. For example, the QMLE is a particular case of QLE, which can be used when the QML criterion is differentiable, and $H_n(\boldsymbol{\theta}) = 0$ corresponds to the first-order conditions of the maximisation problem; however, QMLE can be asymptotically less efficient than QLE (see Francq & Zakoïan (2023) for examples). Similarly, the GMM estimator of Hansen (1982) is related to QLE, with $H_n(\boldsymbol{\theta}) = 0$ corresponding to the moment conditions, although the two estimation techniques differ in the way the observations are weighted. Christensen et al. (2016) show that, in general, the optimal QLE is strictly more efficient than the optimal GMM. Indeed, as Vinod (1997) puts it, it "[...] is remarkable that whenever the OLS or ML estimators do not coincide with the OptEF, it is the OptEF that have superior properties in both large and small samples" (p. 216).[1] On account of these optimality properties, the EF approach has become a very important technique in the toolbox of the applied scientist in such diverse fields as: survey sampling, where Godambe & Thompson (2009) use EF to estimate the response probability parameters (see also the examples in Godambe & Thompson 2009); biostatistics, where the EF approach has been applied to longitudinal datasets in the highly influential paper by Liang & Zeger (1986); in finance, where EF has been suggested for the estimation of (discretely sampled) diffusion-type models (Bibby et al. 2010); and in economics, where, for example, the estimation of the consumption based CAPM - see Tauchen (1986) - which is usually carried out via GMM, with often ad hoc choices of moments,

[1] "OptEF" refers here to the optimal EF estimator.

can instead be based on the EF approach, and Vinod (1997) reports a comprehensive empirical illustration with a comparison between the EF and the GMM methods.

The EF approach is particularly attractive for time series where the dynamics is not fully specified, but the conditional mean is assumed to be a given function of past observations and a finite-dimensional parameter. Unlike "strong" models, which are generally determined by a sequence of innovations (often assumed to be i.i.d.), models characterized only by the first conditional moment (so-called "weak" location models) are not naturally amenable to (quasi) likelihood-based inference. For example, the consistency of the Gaussian quasi-maximum likelihood (QMLE) estimator requires, among other regularity assumptions, the correct specification of the first two conditional moments. In contrast, the QLE, which is obtained by solving estimating equations derived from the first conditional moment, can be consistent and asymptotically normal without the correct specification of the second conditional moment (see Francq & Zakoïan 2023). In this approach, only the conditional mean needs to be correctly specified.

In many applications, however, the conditional mean may undergo a structural change, which highlights the importance of testing for the possible presence of changepoints. This issue is of particular importance in the context of a time series, where, as Hansen (2001) puts it: "Structural change is pervasive in economic time series relationships, and it can be quite perilous to ignore. Inferences about economic relationships can go astray, forecasts can be inaccurate, and policy recommendations can be misleading or worse" (p. 127). The literature on changepoint has a long history, starting with Page (1955), and we refer to Csörgö & Horváth (1997), Aue & Horváth (2013), Casini & Perron (2019) and Horváth & Rice (2024) for state-of-the-art reviews. Whilst numerous contributions have been developed to test for changepoints based on several estimators, to the best of our knowledge nothing is available for the case of EF based inference, and extending the results available for other existing techniques is not a trivial task. Hence, in this paper we investigate the detection of changepoints in the conditional (parametric) mean of a weak location model, using the EF approach to estimate the parameters. We make three main contributions. First, we study the distribution of CUSUM test statistics constructed as the partial sums of the quasi-Fisher scores used in the EF approach. The testing procedure depends on the choice of a sequence of weights, leading to a potentially infinite number of consistent tests, and in this paper we show that the best test is related to Godambe's optimal QLE, also discussing data-driven procedures for this optimal choice of weights. Second, we study inference in the presence of a changepoint, also deriving the limiting distribution of the estimated breakdate. Third, we also study the case where the conditional mean is misspecified, developing Heteroskedasticity and Autocorrelation Consistent (HAC) versions of the tests.

Let us mention a few references on CUSUM-based test statistics that are most relevant to our work. Ploberger & Krämer (1992) introduced an OLS-based CUSUM test for testing the constancy of regression coefficients. Horváth & Parzen (1994) proposed a weighted-CUSUM of Fisher's scores for testing iidness. Berkes et al. (2004) adapted the approach to test parameter constancy in GARCH model, using a QMLE-based score. Negri & Nishiyama (2017) considered more general models with applications to ergodic and non-ergodic diffusion processes. Kutoyants (2016) considered CUSUM statistics based on Fishers's score and studied goodness-of-fit tests for diffusion processes and nonlinear time series models. Lee et al. (2003) considered a CUSUM based on $\widehat{\boldsymbol{\theta}}_k - \widehat{\boldsymbol{\theta}}_n$ for $k = 1, \dots, n$, where $\widehat{\boldsymbol{\theta}}_k$ is an estimator based on the first k observations. Shao & Zhang (2010) proposed a self-normalized Kolmogorov-Smirnov test for a change point in the mean of a time series. Aue & Horváth (2013) showed how CUSUM statistics can be used to detect breaks in the unconditional and conditional means and variances of time series.

The remainder of the paper is organized as follows. Section 2 presents the EF approach. In Section 3, we study the test statistic, its asymptotics under the null (Section 3.1), optimality and inference under the alternative (Section 3.2), the case where the conditional mean may not be correctly specified (Section 3.3), and change-point estimation (Section 3.4). Numerical illustrations based on simulated and real financial data are reported in Section 4.The datasets and R-codes used for the illustration are available here: https://doi.org/10.5281/zenodo.14899445. Examples and proofs are relegated to an appendix.

2 Model, assumptions and the estimating function approach

Consider a real time series $(y_t)_{t\in\mathbb{Z}}$ and the sigma-field $\mathcal{F}_t$ generated by $\{y_u : u \leq t\}$. Assume the existence of a well-defined function $m_t(\boldsymbol{\theta}) = m(\boldsymbol{\theta}; y_{t-1}, y_{t-2}, \dots)$, depending on some parameter $\boldsymbol{\theta} \in \Theta$ where Θ is a compact subset of $\mathbb{R}^d$, such that for some parameter value $\boldsymbol{\theta}_0$ and all $t \in \mathbb{Z}$,

$$m_t := m_t(\boldsymbol{\theta}_0) = E_{t-1}(y_t), \quad \text{where } E_t(\cdot) = E(\cdot \mid \mathcal{F}_t). \tag{1}$$

It is assumed that $m(\boldsymbol{\theta}; \cdot)$ is a well defined measurable function of $\{y_u : u \leq t\}$ and that $m_t(\cdot)$ is almost surely continuous over Θ. It is always possible to write the model as

$$y_t = m_t + \epsilon_t, \tag{2}$$

where $m_t \in \mathcal{F}_{t-1}$ and (ϵ_t) is such that $E_{t-1}(\epsilon_t) \equiv 0$, or equivalently,

$$E_{t-1}\epsilon_t(\boldsymbol{\theta}_0) = 0, \quad \epsilon_t(\boldsymbol{\theta}) = y_t - m_t(\boldsymbol{\theta}). \tag{3}$$

We refer to Model (2) as a *weak location model*, by contrast with models in which strong assumptions are imposed on the error term, such as iidness.

Our main focus will be to test for the possible presence of changes in $\boldsymbol{\theta}$ as time elapses. In other words, assuming the observations $y_1, \ldots, y_n$ satisfy, for $u_0 \in (0, 1]$,

$$y_t = y_{t,n} = \begin{cases} m_t(\boldsymbol{\theta}_1) + \epsilon_t & \text{if } t \leq [nu_0], \\ m_t(\boldsymbol{\theta}_2) + \epsilon_t & \text{if } t > [nu_0], \end{cases} \tag{4}$$

where (ϵ_t) is such that $E_{t-1}(\epsilon_t) \equiv 0$ and $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, we consider testing the hypothesis of constancy over time

$$\mathbf{H_0}: \quad u_0 = 1 \tag{5}$$

versus the alternative hypothesis that a change occurs at an unknown change point

$$\mathbf{H_A}: \quad u_0 \in (0, 1) \quad \text{and} \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2. \tag{6}$$

We would like to point out that, in (6), we consider the alternative hypothesis of At-Most-One-Change (AMOC) for simplicity and only for illustrative purposes; in principle, our approach lends itself to being generalised to the case of multiple changepoints.

2.1 The estimating function approach

We begin by discussing the estimation of $\boldsymbol{\theta}$ via the EF approach. For the sake of a concise discussion, we only report the main aspects of estimation; illustrative examples are presented in Section A of the Appendix.

Assume that n variables $y_1, \ldots, y_n$ have been observed from the time series $\{y_t, -\infty < i < \infty\}$. Consider the following *assumed* proxy of $\sigma_t^2(\boldsymbol{\theta}) := \mathbb{E}_{t-1}\left\{y_t - m_t(\boldsymbol{\theta})\right\}^2$:

$$\kappa_{2t}(\boldsymbol{\theta}) = \kappa_2\left(\boldsymbol{\theta}; y_{t-1}, y_{t-2}, \ldots\right). \tag{7}$$

Note that $m_t(\boldsymbol{\theta})$ and $\kappa_{2t}(\boldsymbol{\theta})$ depend on the non-observed values $\{y_u : u \leq 0\}$. Moreover, $\kappa_{2t}(\boldsymbol{\theta}) = \kappa_{2t}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ may depend on some unknown nuisance parameter $\boldsymbol{\gamma}$ (as well as $\sigma_t^2(\boldsymbol{\theta})$). Let $\mathcal{I}_t$ be the sigma-field generated by $\{y_u : 1 \leq u \leq t\}$, the information available at time t, with $\mathcal{I}_t \subset \mathcal{F}_t$. Let $\widetilde{m}_t(\boldsymbol{\theta})$ be an $\mathcal{I}_t$-measurable approximation of $m_t(\boldsymbol{\theta})$.[2] Similarly, $\widetilde{\kappa}_{2t}$ stands for an $\mathcal{I}_t$-measurable approximation of κ_{2t}, which may be constant or may depend on $\boldsymbol{\theta}$; when $\kappa_{2t}(\boldsymbol{\theta}) = \kappa_{2t}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ depends on a nuisance parameter $\boldsymbol{\gamma} \in \boldsymbol{\Theta}_1$ with $\boldsymbol{\Theta}_1$ a compact set, and $\widehat{\boldsymbol{\gamma}}_n$ is an estimator of $\boldsymbol{\gamma}$, then $\widetilde{\kappa}_{2t} = \widetilde{\kappa}_{2t}(\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}_n)$ denotes an $\mathcal{I}_n$-measurable approximation of $\kappa_{2t}(\boldsymbol{\theta})$.

[2] See Assumption **A3** below for a more precise definition.

Letting, under differentiability assumptions displayed below,

$$\widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta}) = \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\widetilde{\epsilon}_t(\boldsymbol{\theta})}{\widetilde{\kappa}_{2t}}, \quad \text{with} \quad \widetilde{\epsilon}_t(\boldsymbol{\theta}) = y_t - \widetilde{m}_t(\boldsymbol{\theta}),$$

the parameter $\boldsymbol{\theta}_0$ can be estimated by the solution $\widehat{\boldsymbol{\theta}}$ of

$$\sum_{t=1}^{n} \widetilde{\boldsymbol{\Upsilon}}_t\left(\widehat{\boldsymbol{\theta}}\right) = 0. \tag{8}$$

As far as the vocabulary goes, equation (8) is the *estimating equation*, and its left-hand side the *estimating function*. This estimating function corresponds to Fisher's score for particular conditional distributions of y_t (see Example 2 in Section A in the Appendix) and in general it can be interpreted as a quasi-score, hence the title of the paper. Any measurable solution $\widehat{\boldsymbol{\theta}}$ of the estimating equation (8) is the QLE, and all the QLEs that we consider hereafter only differ by the weighting sequence $\widetilde{\kappa}_{2t}$. In Section A.1, we discuss some examples of QLE (see Examples 1 and 2), and illustrate the link between QLE and GMM (Example 3); in Section A.2, we discuss examples of $\widetilde{\kappa}_{2t}$ (see Example 4, where we also discuss data-driven rules to optimally select $\widetilde{\kappa}_{2t}$; see Example 5 for GARCH estimation of $\widetilde{\kappa}_{2t}$).

2.2 Assumptions

To develop a test of the null hypothesis $\mathbf{H_0}$ of no change, we make the following assumptions. Define

$$\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) = \frac{\partial m_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\epsilon_t(\boldsymbol{\theta})}{\kappa_{2t}(\boldsymbol{\theta})}, \tag{9}$$

and, under the assumptions below, consider the matrices

$$\boldsymbol{I} = \mathbb{E}\left(\frac{\sigma_t^2(\boldsymbol{\theta}_0)}{\kappa_{2t}^2(\boldsymbol{\theta}_0)} \frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right), \tag{10}$$

$$\boldsymbol{J} = -\mathbb{E}\left(\frac{1}{\kappa_{2t}(\boldsymbol{\theta}_0)} \frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}\right). \tag{11}$$

A1 The process $(y_t)_{t\in\mathbb{Z}}$ is strictly stationary and ergodic.

A2 There exists $\rho \in [0,1)$ such that, almost surely (a.s.) $\sup_{\boldsymbol{\theta}\in\Theta} |m_t(\boldsymbol{\theta}) - \widetilde{m}_t(\boldsymbol{\theta})| \leq K_t \rho^t$, where K_t is a generic $\mathcal{F}_{t-1}$-measurable random variable such that $\sup_t EK_t^r < \infty$ for some $r > 0$.

A3 Let $\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) = \frac{\partial m_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\epsilon_t(\boldsymbol{\theta})}{\kappa_{2t}(\boldsymbol{\theta})}$. If $E\{\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})\} = 0$ for some $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The parameter $\boldsymbol{\theta}_0$ belongs to the interior of the compact set Θ.

A4 The function $\boldsymbol{\theta} \mapsto m_t(\boldsymbol{\theta})$ is twice continuously differentiable, and

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{\partial m_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \leq K_t \rho^t, \qquad a.s.$$

where K_t is as in **A2**, $\|\cdot\|$ denotes any norm on $\mathbb{R}^d$. Moreover, for some $s>0$, $E|y_t|^s<\infty$ and $E\sup_{\boldsymbol{\theta}\in\Theta}\left\{|m_t(\boldsymbol{\theta})|^s+\left\|\frac{\partial m_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\|^s\right\}<\infty$.

A5 If $\boldsymbol{\lambda}^\top\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}=0$ a.s. then $\boldsymbol{\lambda}=\mathbf{0}_d$.

A6 There exists a constant $\underline{\kappa}>0$ such that $\inf_{\boldsymbol{\theta}\in\Theta}\kappa_{2t}(\boldsymbol{\theta})\geq\underline{\kappa}$ a.s.

A7 For all $\boldsymbol{\theta}\in\Theta$ the sequence $\{\kappa_{2t}(\boldsymbol{\theta})\}_{t\in\mathbb{Z}}$ is stationary, ergodic and $\mathcal{F}_{t-1}$-measurable, the function $\boldsymbol{\theta}\mapsto\kappa_{2t}(\boldsymbol{\theta})$ is continuously differentiable, there exist $\rho\in[0,1)$ and K_t as in **A2** such that, almost surely,

$$\sup_{\boldsymbol{\theta}\in\Theta}|\kappa_{2t}(\boldsymbol{\theta})-\widetilde{\kappa}_{2t}(\boldsymbol{\theta})|\leq K_t\rho^t$$

for n large enough.[3]

A8 We have

$$E\sup_{\boldsymbol{\theta}\in\Theta}\|\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})\|^2<\infty\quad\text{and}\quad E\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{\partial\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}\right\|<\infty.$$

Assumptions **A1**-**A8** have similarities with Assumptions A1-A10 in Francq & Zakoïan (2023), and we refer to that paper for comments. In Section A.3 in the Appendix, we illustrate them through some examples (see Examples 6 and 7, where we consider an application to AR(1) models).

3 Change-point tests

Inspired by basic CUSUM statistics used in changepoint problems, we consider the process, defined for $u\in[0,1]$ by

$$\widetilde{\boldsymbol{T}}_n(u)=\frac{1}{\sqrt{n}}\sum_{t=1}^{[nu]}\widetilde{\boldsymbol{\Upsilon}}_t(\widehat{\boldsymbol{\theta}}).$$

By convention $\widetilde{\boldsymbol{T}}_n(0)=0$ and, by definition of the QLE, we also have $\widetilde{\boldsymbol{T}}_n(1)=0$. A natural statistic for testing $\mathbf{H_0}$ is

$$\widetilde{S}_n=\sup_{u\in(0,1)}\widetilde{S}_n(u)=\max_{k\in\{1,\ldots,n-1\}}\widetilde{S}_n(k/n),\qquad\widetilde{S}_n(u)=\widetilde{\boldsymbol{T}}_n^\top(u)\boldsymbol{I}_n^{-1}\widetilde{\boldsymbol{T}}_n(u)$$

where $\boldsymbol{I}_n$ denotes a non singular consistent estimator of $\boldsymbol{I}$. Note that **A5** entails that $\boldsymbol{I}$ is not singular.

[3]Recall that $\widetilde{\kappa}_{2t}$ may depend on $\widehat{\gamma}_n$.

3.1 Asymptotics under the null of no change

Inspired by the references mentioned in the introduction, we will show that, under the null of no break, $\widetilde{\boldsymbol{T}}_n(u)$ converges weakly to a Gaussian process $\boldsymbol{T}(u) = (T_1(u), \dots, T_d(u))^\top$ with covariance structure $E\boldsymbol{T}(u) = 0$ and $\mathrm{Cov}(\boldsymbol{T}(u), \boldsymbol{T}(v)) = \boldsymbol{I}\{\min(u,v) - uv\}$. Thus, each component of the vector $\boldsymbol{I}^{-1/2}\boldsymbol{T}(u)$ is a standard Brownian bridge $\{\boldsymbol{B}(u), u \in [0,1]\}$, with $\boldsymbol{B}(u) = \boldsymbol{W}(u) - u\boldsymbol{W}(1)$ where $\{\boldsymbol{W}(u), u \in [0,1]\}$ denotes a standard Brownian motion on $\mathbb{R}^d$.

Theorem 1. *Under Assumptions* **A1**-**A8***, including* $\mathbf{H_0}$*, we have*

$$\widetilde{S}_n \xrightarrow{\mathcal{L}} S := \sup_{u\in(0,1)} \boldsymbol{T}^\top(u)\boldsymbol{I}^{-1}\boldsymbol{T}(u) = \sup_{u\in(0,1)} \sum_{j=1}^d \{B_j(u)\}^2,$$

where $\boldsymbol{B}(u) = (B_1(u), \dots, B_d(u))^\top$ *is a d-dimensional standard Brownian bridge.*

Note that the distribution of S (for $d \le 10$) is tabulated in Lee et al. (2003). In the proof of this theorem, we establish the following expansion which will be used throughout:

$$\begin{aligned}
\boldsymbol{T}_n(u) :=& \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu]} \boldsymbol{\Upsilon}_t(\widehat{\boldsymbol{\theta}}) \\
=& \frac{1}{\sqrt{n}} \left(\sum_{t=1}^{[nu]} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - u \sum_{t=1}^{n} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) \right) + \frac{1}{n} \sum_{t=1}^{[nu]} \left(\frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\widehat{\boldsymbol{\theta}}^*) - \boldsymbol{J} \right) \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_P(1) \\
:=& \boldsymbol{T}_n^0(u) + \boldsymbol{R}_n(u)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_P(1),
\end{aligned} \tag{12}$$

where $\widehat{\boldsymbol{\theta}}^*$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, and

$$\sup_{u\in(0,1)} \|\boldsymbol{R}_n(u)\| = o_P(1), \qquad \sup_{u\in(0,1)} \|\boldsymbol{T}_n(u) - \widetilde{\boldsymbol{T}}_n(u)\| = o_P(1). \tag{13}$$

Remark 1. The estimation of $\boldsymbol{I}$ cannot be achieved by plug-in, using formula (10). Indeed, the conditional variance function $\sigma_t^2(\cdot)$ is generally unknown. However, a consistent estimator of $\boldsymbol{I} = E\left\{\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0)\boldsymbol{\Upsilon}_t^\top(\boldsymbol{\theta}_0)\right\}$ is

$$\boldsymbol{I}_n = \frac{1}{n} \sum_{t=1}^n \widetilde{\boldsymbol{\Upsilon}}_t(\widehat{\boldsymbol{\theta}}_n)\widetilde{\boldsymbol{\Upsilon}}_t^\top(\widehat{\boldsymbol{\theta}}_n). \tag{14}$$

Remark 2. Nyblom (1989) proposed a general theory for testing the constancy of the parameters involved in the conditional distribution of a time series model. Applied to our semiparametric framework, the Nyblom test[4] replaces the supremum with a mean. More specifically, the test rejects the parameter constancy for large values of

$$\widetilde{S}_n^N := \frac{1}{n} \sum_{k=1}^n \widetilde{S}_n(k/n)$$

[4]We are grateful to P.R. Hansen for pointing out that we can use this test in our framework.

which, by the continuous mapping theorem, has the asymptotic distribution $\int_0^1 \sum_{j=1}^d \{B_j(u)\}^2 du$ under the assumptions of Theorem 1. The Nyblom test, which enjoys some optimality properties under the alternative that the parameter process follows a martingale, is widely used in econometrics (for an example see Hansen et al. (2014), where the test is used to assess the constancy of a correlation).

Remark 3. The CUSUM test also has optimality properties, but for different alternatives than the Nyblom test. At the very beginning of their book, Horváth & Rice (2024) give a univariate example where the (standardized) CUSUM test coincides with a likelihood ratio test and thus enjoys its general good asymptotic properties. See Section A.4 for a multivariate example.

We can now construct a test for $\mathbf{H_0}$. At the significance level $\alpha \in (0,1)$, an asymptotic critical region is given by

$$\max_{1\le k\le n} \widetilde{S}_n(k/n) > S_{1-\alpha}, \tag{15}$$

where $S_{1-\alpha}$ is the $(1-\alpha)$ quantile of the law of S.

3.2 Local Asymptotic Power comparisons

The estimator defined in (8) with $\widetilde{\kappa}_{2t}$ proportional to $\sigma_t^2(\boldsymbol{\theta}_0)$ is optimal in the Godambe sense within the class of EF estimators solving

$$\sum_{t=1}^n \boldsymbol{a}_{t-1}(\boldsymbol{\theta})\widetilde{\epsilon}_t(\boldsymbol{\theta}) = 0, \tag{16}$$

where $\boldsymbol{a}_{t-1}(\boldsymbol{\theta})$ is a $d\times 1$ vector belonging to $\mathcal{F}_{t-1}$ (see Chandra & Taniguchi 2001). In this section we show that Godambe's optimal QLEs lead to optimal tests, in the sense that they optimize some local asymptotic power (LAP).

We will consider a sequence of "local breaks" occurring at a proportion $u_0 \in (0,1)$ of the observations.

3.2.1 Local break in the mean of a sequence of i.i.d. Gaussian variables

The simplest example of local break is obtained by assuming that $y_1,\dots,y_n$ are independent and Gaussian with variance σ^2, and that $y_t = y_{t,n}$ has mean $\theta_0 + \delta_1/\sqrt{[nu_0]}$ when $t \le [nu_0]$ and $\theta_0 + \delta_2/\sqrt{n-[nu_0]}$ when $t > [nu_0]$. We then have

$$\frac{1}{\sqrt{[nu_0]}}\sum_{t=1}^{[nu_0]}(y_t-\theta_0) \sim \mathcal{N}\left(\delta_1,\sigma^2\right), \quad \frac{1}{\sqrt{n-[nu_0]}}\sum_{t=[nu_0]+1}^{n}(y_t-\theta_0) \sim \mathcal{N}\left(\delta_2,\sigma^2\right),$$

and thus

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(y_t-\theta_0)\sim\mathcal{N}\left(\delta_3,\sigma^2\right),\quad \delta_3=\sqrt{u_0}\delta_1+\sqrt{1-u_0}\delta_2.$$

Note that, in this simple example, $\overline{y}=n^{-1}\sum_{t=1}^{n}y_t$ is the Q(M)LE of θ_0 (under the null $\delta_1=\delta_2=0$ of no local break), $\widetilde{T}_n(u)=T_n(u)=n^{-1/2}\sum_{t=1}^{[nu]}(y_t-\overline{y})$ is the usual CUSUM process, and

$$\widetilde{S}_n=\sup_{u\in(0,1)}\frac{1}{n\widehat{\sigma}_y^2}\left\{\sum_{t=1}^{[nu]}(y_t-\overline{y})\right\}^2,\quad \widehat{\sigma}_y^2=\frac{1}{n}\sum_{t=1}^{n}(y_t-\overline{y})^2,$$

is nothing else than the Kolmogorov test statistic. Note also that (12)–(13) hold with $R_n(u)=u-[nu]/n$ and $o_P(1)=0$. The asymptotic distribution of the Kolmogorov test statistic under such local breaks can be obtained as a corollary of the next result.

3.2.2 Local asymptotic power

Let us now return to the general situation. Suppose the conditional distribution of y_t changes at a single point, which is located at a fixed proportion $u_0\in(0,1)$ of the observations. Let $\widehat{\boldsymbol{\theta}}_{(1)}$ be the QLE computed on $y_1,\dots,y_{[u_0n]}$ and $\widehat{\boldsymbol{\theta}}_{(2)}$ the QLE computed on $y_{[u_0n]+1},\dots,y_n$. Recall that $\widehat{\boldsymbol{\theta}}$ is the QLE computed on all the observations $y_1,\dots,y_n$. Let the local alternatives $H_1=H_1(\boldsymbol{\delta}_1,\boldsymbol{\delta}_2)$ such that, for $\boldsymbol{\Upsilon}_t=\boldsymbol{\Upsilon}_{t,n}(\boldsymbol{\theta}_0)$, as $n\to\infty$

$$\sqrt{nu_0}\left(\widehat{\boldsymbol{\theta}}_{(1)}-\boldsymbol{\theta}_0\right)=-\boldsymbol{J}^{-1}\frac{1}{\sqrt{nu_0}}\sum_{t=1}^{[nu_0]}\boldsymbol{\Upsilon}_t+o_P(1)\xrightarrow{\mathcal{L}}\mathcal{N}\left(\boldsymbol{\delta}_1,\boldsymbol{J}^{-1}\boldsymbol{I}\boldsymbol{J}^{-1}\right),\tag{17}$$

$$\sqrt{n(1-u_0)}\left(\widehat{\boldsymbol{\theta}}_{(2)}-\boldsymbol{\theta}_0\right)=-\boldsymbol{J}^{-1}\frac{1}{\sqrt{n(1-u_0)}}\sum_{t=[nu_0]+1}^{n}\boldsymbol{\Upsilon}_t+o_P(1)\xrightarrow{\mathcal{L}}\mathcal{N}\left(\boldsymbol{\delta}_2,\boldsymbol{J}^{-1}\boldsymbol{I}\boldsymbol{J}^{-1}\right).\tag{18}$$

Under mild regularity conditions (for example under mixing conditions), $\widehat{\boldsymbol{\theta}}_{(1)}$ and $\widehat{\boldsymbol{\theta}}_{(2)}$ are asymptotically independent, and we then have

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=-\boldsymbol{J}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{\Upsilon}_t+o_P(1)\xrightarrow{\mathcal{L}}\mathcal{N}\left(\boldsymbol{\delta}_3,\boldsymbol{J}^{-1}\boldsymbol{I}\boldsymbol{J}^{-1}\right)\tag{19}$$

with $\boldsymbol{\delta}_3=\sqrt{u_0}\boldsymbol{\delta}_1+\sqrt{1-u_0}\boldsymbol{\delta}_2$.

Theorem 2. *Assume $H_1(\boldsymbol{\delta}_1,\boldsymbol{\delta}_2)$, regularity conditions ensuring $\boldsymbol{I}_n\to\boldsymbol{I}$ a.s. with $\boldsymbol{I}$ non singular, (12)–(13) and (17)–(19). Then, for all $u\in(0,1)$, $\widetilde{S}_n(u)/u(1-u)$ converges in distribution to a noncentral chi-square distribution with d degrees of freedom. When $\sqrt{1-u_0}\boldsymbol{\delta}_1\neq\sqrt{u_0}\boldsymbol{\delta}_2$, the noncentrality parameter is not equal to 0 and the best LAP is obtained for the optimal QLE.*

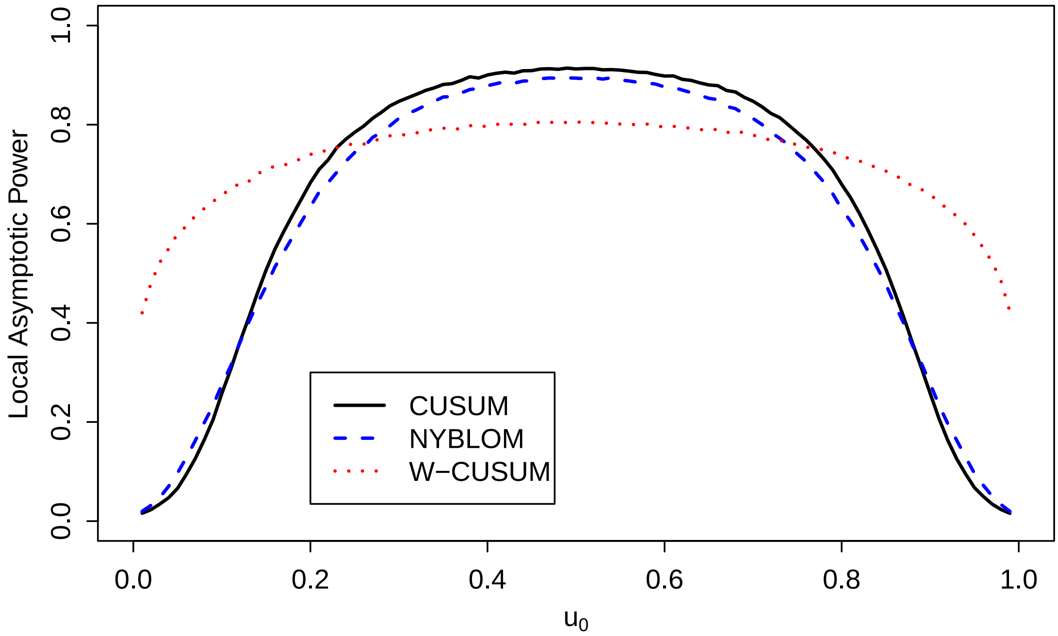

Figure 1: Powers of the CUSUM, Nyblom, and Weighted CUSUM tests as a function of the break date u_0.

3.2.3 Comparisons with other tests

As an illustration of Theorem 2, let us compare the LAPs of the CUSUM, NYBLOM (see Remark 2) and Weighted CUSUM (W-CUSUM) in the simple case of Section 3.2.1. The 3 tests reject for large values of $\widetilde{S}_n$ $\widetilde{S}_n^N$ and $\widetilde{S}_n^W$ defined by

$$\widetilde{S}_n = \max_{1\leq k<n} \widetilde{S}_n\left(\frac{k}{n}\right), \quad \widetilde{S}_n^N = \frac{1}{n}\sum_{k=1}^{n} \widetilde{S}_n\left(\frac{k}{n}\right), \quad \widetilde{S}_n^W = \max_{1\leq k<n} \frac{n^2}{k(n-k)}\widetilde{S}_n\left(\frac{k}{n}\right)$$

with $\widetilde{S}_n(k/n) = \left\{\sum_{t=1}^{k}(y_t-\overline{y})\right\}^2/(n\widehat{\sigma}_y^2)$. The critical values of the tests as well as the LAPs are evaluated by using $50,000$ independent replications of the test statistics with $n = 1,000$. Figure 1 shows the LAPs for the nominal leavel $\alpha = 1\%$ and the alternatives $H_1(\delta_1, \delta_2)$ with $\delta_1 = -\delta_2 = 3$ for a grid of values of $u_0 \in \{0.01, 0.02, \ldots, 0.99\}$. As expected, the weighted CUSUM is more powerful than the unweighted version when the local break u_0 comes early or late. Note that the CUSUM and Nyblom tests have similar power, often with a slight advantage for the CUSUM.

3.3 Change-point tests with misspecified conditional mean

In Sections 3.1-3.2, we assumed that $m_t(\boldsymbol{\theta}_0)$ is truly the conditional mean of y_t given $\mathcal{F}_{t-1}$. In this section, we propose to relax this assumption. The intuition is that, even if the conditional mean is not correctly specified, its estimated parameter value should not vary too much when the DGP is stable. Replace **A3** and **A5** by:

A3* Let $\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) = \frac{\partial m_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{y_t - m_t(\boldsymbol{\theta})}{\kappa_{2t}(\boldsymbol{\theta})}$ where $m_t(\cdot)$ is $\mathcal{F}_{t-1}$-measurable. If $E\{\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})\} = 0$ for some $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{\theta} = \boldsymbol{\theta}_0^*$, where the so-called pseudo-true parameter $\boldsymbol{\theta}_0^*$ belongs to the interior of the compact set Θ.

Under **A8**, let $\boldsymbol{J}^* = E\frac{\partial}{\partial \boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^*)$ and assume:

A5* $\boldsymbol{J}^*$ is non singular.

Let $\mu_t = E(y_t \mid \mathcal{F}_{t-1})$. Note that we may have $\mu_t \neq m_t(\boldsymbol{\theta}_0^*)$, and more generally $\mu_t \neq m_t(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$. See Example 11 for the case of a misspecified AR(1).

Let $\boldsymbol{\Upsilon}_t^* = \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^*)$. We now need conditions ensuring the Central Limit Theorem (CLT)

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{\Upsilon}_t^* \xrightarrow{\mathcal{L}} \mathcal{N}(0, \boldsymbol{I}^*) \tag{20}$$

for some long-run variance $\boldsymbol{I}^*$. Let $\{\alpha(h)\}_{h\geq 0}$ be the $\alpha-$mixing (strong mixing) coefficients of the process $(\boldsymbol{\Upsilon}_t^*)_{t\in\mathbb{Z}}$, defined by $\alpha(h) = \sup_{A\in\sigma(\boldsymbol{\Upsilon}_u^*, u\leq t), B\in\sigma(\boldsymbol{\Upsilon}_u^*, u\geq t+h)} |P(A\cap B) - P(A)P(B)|$. We reinforce **A1** by the following assumption.

A1* We have $\|\boldsymbol{\Upsilon}_1^*\|_{2+\nu} < \infty$ and $\sum_{h=1}^{\infty}\{\alpha(h)\}^{\nu/(2+\nu)} < \infty$ for some $\nu > 0$.

Note that, by Davydov's inequality, **A1*** entails the existence of the matrix $\boldsymbol{I}^*$.

Theorem 3. *Under Assumptions* **A1**, **A1***, **A2**, **A3***, **A4**, **A6**-**A8**, *there exists a QLE* $\widehat{\boldsymbol{\theta}}$ *satisfying*

$$\sum_{t=1}^{n} \widetilde{\boldsymbol{\Upsilon}}_t(\widehat{\boldsymbol{\theta}}) = 0, \quad \widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta}) = \frac{1}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})} \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \widetilde{\epsilon}_t(\boldsymbol{\theta}),$$

for n large enough. Moreover $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0^*$ a.s. and, under **A5***,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*\right) = -\boldsymbol{J}^{*-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^*) + o_P(1) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \boldsymbol{J}^{*-1}\boldsymbol{I}^*\boldsymbol{J}^{*-1}\right) \; \textit{as } n \to \infty.$$

Standard estimators of a long-run variance of the form $\boldsymbol{I}^*$ are the Heteroskedasticity and Autocorrelation Consistent (HAC) estimators (see Newey & West (1987) and Andrews (1991)) and spectral density estimators (see Den Haan & Levin 1997). Denote by $\boldsymbol{I}_n^*$ a consistent estimator of $\boldsymbol{I}^*$, and consider the process

$$\widetilde{S}_n^* = \sup_{u\in(0,1)} \widetilde{S}_n^*(u), \qquad \widetilde{S}_n^*(u) = \widetilde{\boldsymbol{T}}_n^\top(u)\boldsymbol{I}_n^{*-1}\widetilde{\boldsymbol{T}}_n(u).$$

Theorem 4. *Under the Assumptions of Theorem 3, in particular the non-existence of a break, and if $\boldsymbol{I}^*$ is invertible we have $\widetilde{S}_n^* \xrightarrow{\mathcal{L}} S$.*

Note that $\boldsymbol{\theta}_0^*$ in **A3*** may vary with κ_{2t}, and thus the "optimal" test statistics is not necessarily obtained by choosing κ_{2t} proportional to the conditional variance, as is the case when $m_t(\cdot)$ corresponds to the well-specified conditional mean (see Lemma 1 and Theorem 2).

3.4 Change-point estimation

One of the main goals of change-point analysis is to estimate the location of breaks under the alternative. Results on this issue go back to Hinkley (1970) in the case of iid random variables. To cite just one more recent reference for a general class of (strong) time series models, Ling (2016) derived asymptotic results on estimated change-points.

Assume that, for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ belonging to Θ and for $u_0 \in (0, 1]$, Model (4) holds. Recall that $u_0 = 1$ corresponds to the null hypothesis of no change-point. As in the previous sections, $\widehat{\boldsymbol{\theta}}$ denotes a QLE such that $\sum_{t=1}^n \widetilde{\boldsymbol{\Upsilon}}_t(\widehat{\boldsymbol{\theta}}) = 0$.

We will introduce two stationary processes, $(y_t^{(1)})_{t\in\mathbb{Z}}$ and $(y_t^{(2)})_{t\in\mathbb{Z}}$, which will be used to approximate the observed process before and after the break, respectively. For all $\boldsymbol{\theta} \in \Theta$, let $m_t^{(i)}(\boldsymbol{\theta}) = m(\boldsymbol{\theta}; y_{t-1}^{(i)}, y_{t-2}^{(i)}, \dots)$ and $\kappa_{2t}^{(i)}(\boldsymbol{\theta}) = \kappa_2(\boldsymbol{\theta}; y_{t-1}^{(i)}, y_{t-2}^{(i)}, \dots)$ be stationary approximations of the conditional mean and weight sequence before and after the break.

B1 For $i = 1, 2$, the process $y_t^{(i)} = m_t^{(i)}(\boldsymbol{\theta}_i) + \epsilon_t$, for $t \in \mathbb{Z}$, is strictly stationary and ergodic.

For all $\boldsymbol{\theta} \in \Theta$, let $\boldsymbol{\Upsilon}_t^{(i)}(\boldsymbol{\theta}) = \frac{\partial m_t^{(i)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\epsilon_t^{(i)}(\boldsymbol{\theta})}{\kappa_{2t}^{(i)}(\boldsymbol{\theta})}$ where $\epsilon_t^{(i)}(\boldsymbol{\theta}) = y_t^{(i)} - m_t^{(i)}(\boldsymbol{\theta})$. The pseudo-true parameter value is introduced as follows.

B2 For all $\boldsymbol{\theta}$ in Θ the variables $\boldsymbol{\Upsilon}_t^{(i)}(\boldsymbol{\theta})$ have finite variances, and there is a unique solution $\boldsymbol{\theta}_0^\star = \boldsymbol{\theta}_0^\star(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, belonging to the interior of Θ, to the equation

$$u_0 E\left\{\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta})\right\} + (1 - u_0) E\left\{\boldsymbol{\Upsilon}_t^{(2)}(\boldsymbol{\theta})\right\} = 0.$$

We make the following technical assumptions.

B3 For $i = 1, 2$, the function $\kappa_{2t}^{(i)}(\cdot)$ is continuously differentiable and $m_t^{(i)}(\cdot)$ is twice continuously differentiable. Moreover, there exists $\rho \in (0, 1)$ such that, almost surely, for $1 \le t \le [nu_0]$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\{ |m_t^{(1)}(\boldsymbol{\theta}) - \widetilde{m}_t(\boldsymbol{\theta})| + \left\| \frac{\partial m_t^{(1)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| + |\kappa_{2t}^{(1)}(\boldsymbol{\theta}) - \widetilde{\kappa}_{2t}(\boldsymbol{\theta})| \right\} \le K_t^{(1)} \rho^t,$$

and for $t > [nu_0]$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\{ |m_t^{(2)}(\boldsymbol{\theta}) - \widetilde{m}_t(\boldsymbol{\theta})| + \left\| \frac{\partial m_t^{(2)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| + |\kappa_{2t}^{(2)}(\boldsymbol{\theta}) - \widetilde{\kappa}_{2t}(\boldsymbol{\theta})| \right\} \le K_t^{(2)} \rho^{t-[nu_0]},$$

where $K_t^{(1)}$ is a measurable function of $\{y_u^{(1)} : u < t\}$ and $K_t^{(2)}$ is a measurable function of $\{y_u^{(1)}, y_u^{(2)} : u < t\}$, with $\sup_t E\{K_t^{(i)}\}^r < \infty$ for $i = 1, 2$ and some $r > 0$.

B4 For $i=1,2$ $E|y_t^{(i)}|^s<\infty$ and $E\sup_{\boldsymbol{\theta}\in\Theta}\left\{|m_t^{(i)}(\boldsymbol{\theta})|^s+\left\|\frac{\partial m_t^{(i)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\|^s+\left|\kappa_{2t}^{(i)}(\boldsymbol{\theta})\right|^s\right\}<\infty$, for some $s>0$. Moreover, $\inf_{\boldsymbol{\theta}\in\Theta}\left|\kappa_{2t}^{(i)}(\boldsymbol{\theta})\right|\geq\underline{\kappa}$ a.s. for some constant $\underline{\kappa}>0$.

B5 For $i=1,2$

$$E\sup_{\boldsymbol{\theta}\in\Theta}\left\|\boldsymbol{\Upsilon}_t^{(i)}(\boldsymbol{\theta})\right\|^2<\infty\quad\text{and}\quad E\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{\partial\boldsymbol{\Upsilon}_t^{(i)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^\top}\right\|<\infty.$$

Assumptions **B1-B5** are illustrated in Section A.3 of the Appendix.

Let the change-point estimator

$$\widetilde{k}=\underset{k\in\{1,\ldots,n-1\}}{\arg\max}\ \widetilde{S}_n(k/n),\qquad \widetilde{S}_n(u)=\widetilde{\boldsymbol{T}}_n^\top(u)\boldsymbol{I}_n^{-1}\widetilde{\boldsymbol{T}}_n(u).$$

The consistency of the change-point estimator is established in the following result.

Theorem 5. *Under Assumptions* **B1-B5**, *when* $u_0\in(0,1)$ *and* $E\left\{\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta}_0^\star)\right\}\neq E\left\{\boldsymbol{\Upsilon}_t^{(2)}(\boldsymbol{\theta}_0^\star)\right\}$ *we have*

$$\frac{\widetilde{k}}{n}\to u_0,\quad \textit{in probability as } n\to\infty.$$

4 Numerical illustrations

In this section, we start by comparing on simulations the empirical sizes and powers of the break test under different settings and different choices of the weights in the QLE. Then, we apply our methodology to exchange rates.

4.1 Monte Carlo experiments

Our first Monte Carlo experiments aim to evaluate how the choice of the weighting sequence $\widetilde{\kappa}_{2t}(\cdot)$ impacts the finite sample performance of the test. We simulated a time series (y_t) such that the distribution of y_t conditional on $\mathcal{F}_{t-1}$ is a Gamma law with the shape parameter $k_t=m_t^2/(k\sigma_t^2)$ and the rate parameter $\theta_t=k\sigma_t^2/m_t$, such that $E_{t-1}(y_t)=m_t$ and $\mathrm{Var}_{t-1}(y_t)=k\sigma_t^2$. We took the ARMA(1,1) conditional mean $m_t=c+ay_{t-1}+bm_{t-1}$ and 4 possibilities for the conditional variance: $\sigma_t^2=1$ in DGP A (as for a standard ARMA model), $\sigma_t^2=m_t$ in DGP B (as for an INGARCH count time series model), $\sigma_t^2=m_t^2$ in DGP C (as for an ACD duration model, or the square of a GARCH), $\sigma_t^2=m_t^{3/2}$ in DGP D, which does not correspond to any standard model.

We considered 8 different QLEs solving (8): for the estimators A, B, C and D the weight sequence $\widetilde{\kappa}_{2t}$ is proportional to 1, $\widetilde{m}_t(\boldsymbol{\theta})$, $\widetilde{m}_t^2(\boldsymbol{\theta})$ and $\widetilde{m}_t^{3/2}(\boldsymbol{\theta})$, respectively, and for the other 4 estimators the weights are data driven. More precisely, the estimator Q is the one that minimizes the QLIK_n loss of Example 4 of Appendix A over the 4 weighting sequences of the estimators A-D. The QLE named G estimates the weights using the GARCH(1,1) model (22), with the QLE A as first step estimator $\widehat{\boldsymbol{\theta}}$. The QLEs X1 and X2 estimate the weights by (23) and (24), respectively, also with the QLE A as first step estimator. The left part of Table 1 shows the empirical sizes of the tests based on the 8 estimators for each of the 4 DGPs, when $n = 2000$, $(c, a, b) = (0.01, 0.1, 0.89)$ and the nominal level $\alpha \in \{1\%, 5\%, 10\%\}$. The relative frequencies of rejection of the null $\mathbf{H_0}$ of no break, presented in Table 1 and in the other tables, are computed over 1000 independent replications of each DGP. For a test of level 1% (respectively 5%, or 10%), the empirical relative frequency of rejection over 1000 independent replications should vary between 0.4% and 1.7% (respectively 3.7% and 6.4%, or 8.2% and 11.9%) with a probability of about 95%. Relative frequencies outside these bounds are highlighted (red for overly high rejection rates, blue for overly low rates). The table shows that first-order errors are generally well-controlled by the optimal tests and, more importantly, by the data-estimated optimal tests, with only slight under-rejections in a few cases. However, non-optimal QLEs may have poor empirical sizes, which is the main motivation for using the proposed data-estimated optimal tests.

To compare the power of the different tests, we considered DGPs with a break at $t = 800$. For $t = 1, \dots, 800$ we took $(c, a, b) = (0.01, 0.1, 0.89)$ and for $t = 801, \dots, 2000$ we took $(c, a, b) = (0.15, 0.1, 0.75)$. The other parameters are unchanged, leading to DGPs A*-D*. Note that before and after the break the marginal mean $c/(1 - a - b) = 1$ and the update parameter $a = 0.1$ remain the same for all DGPs, only the persistence parameter b changes. Despite the DGPs being chosen such that it was impossible to visually detect a change point in the trajectories, the right part of Table 1 shows that the tests are often able to detect the break. As expected, for the DGP $\text{X} \in \{\text{A}, \text{B}, \text{C}, \text{D}\}$, the most powerful test is (or is close to) X among $\{\text{A}, \text{B}, \text{C}, \text{D}\}$. Interestingly, the data-selected estimators always perform very well, often as well as the optimal estimator. The poorer performing tests are highlighted in color. These underperforming tests are never the data-selected ones.

Figure 2 shows the empirical distributions of the change point estimates obtained with the 8 different tests. The simulated DGP is DGPA*.The figure, together with the previous table, shows that optimal and data-selected tests outperform others both in detecting breaks and in estimating their positions.

We conducted another set of Monte Carlo simulations to evaluate the ability of the test to detect structural breaks when the conditional mean is misspecified but the matrix I^* is estimated

α	A	B	C	D	Q	G	X1	X2	A	B	C	D	Q	G	X1	X2
				DGP A								DGP A*				
1%	1.0	2.7	4.9	4.7	0.9	0.6	0.7	0.8	79.8	32.6	20.2	26.0	71.5	77.2	76.8	77.2
5%	5.3	8.8	12.1	12.8	4.9	5.8	5.8	5.6	94.3	53.4	35.4	47.2	88.9	93.3	92.8	92.9
10%	10.1	14.6	18.3	18.7	9.5	11.0	10.2	10.1	97.0	65.7	46.3	60.4	93.1	96.7	96.6	96.6
				DGP B								DGP B*				
1%	0.9	0.7	1.7	1.1	0.7	0.4	0.7	0.7	59.9	80.5	30.0	69.6	79.6	77.2	82.1	82.3
5%	4.1	4.5	7.4	5.0	4.5	2.9	3.9	4.1	80.6	95.6	54.6	89.4	95.5	94.4	95.7	95.8
10%	9.8	8.4	13.0	10.8	8.4	8.4	8.3	8.8	89.7	98.4	68.5	96.0	98.4	98.3	98.7	98.5
				DGP C								DGP C*				
1%	5.8	1.1	0.7	0.7	0.7	0.6	0.8	0.8	49.7	70.9	81.8	83.2	81.8	81.7	83.5	88.8
5%	14.8	4.7	3.9	4.3	3.9	3.2	3.6	4.2	65.5	84.2	95.5	94.6	95.6	95.2	95.8	96.9
10%	22.9	9.9	9.3	8.8	9.3	7.3	8.0	8.7	74.0	91.2	97.9	98.0	97.9	97.6	97.7	98.5
				DGP D								DGP D*				
1%	1.7	1.1	1.7	0.9	0.9	0.7	1.0	1.1	52.3	79.7	67.1	86.7	82.6	81.1	81.6	84.5
5%	7.3	3.9	6.2	5.0	5.0	3.9	4.0	5.2	66.8	94.0	87.6	96.4	95.6	95.6	95.6	96.6
10%	13.6	8.3	10.6	9.7	9.8	8.8	8.7	9.5	76.1	97.2	94.9	98.0	97.7	98.1	98.1	98.1

Table 1: Empirical size (DGP A–D) and power (DGP A*–D*) of 8 QLE-based tests.

by a HAC estimator, as stated in Theorem 4. More specifically, we generated 1000 independent simulations of size $n = 2000$ of an ARMA-GARCH model of the form $y_t = m_t + \epsilon_t$, with $\epsilon_t = \sigma_t \eta_t$, where η_t is an iid noise with a standardized Student distribution with $\nu = 11$ degrees of freedom, $m_t = c + ay_{t-1} + bm_{t-1}$ and $\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \sigma_{t-1}^2$. For the GARCH parameters we took $(\omega, \alpha, \beta) = (0.01, 0.1, 0.83)$, a values close to those typically estimated for real financial return series. Panel $\mathbf{H_0}$ of Table 2 corresponds to a stable DGP with $(c, a, b) = (0.01, -0.5, 0.89)$ for $t = 1, \dots, 2000$. Panel $\mathbf{H_1}$ concerns a DGP with a break $(c, a, b) = (0.01, -0.5, 0.89)$ for $t = 1, \dots, 800$ and $(c, a, b) = (0.01, -0.1, 0.89)$ for $t = 801, \dots, 2000$. To compute the test statistic $\widetilde{S}_n^*$ we considered a misspecified AR(1) model for $\widetilde{m}_t$. In Table 2 the column LSE corresponds to a statistic $\widetilde{S}_n^*$ based on the score $\frac{1}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})}\frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\widetilde{\epsilon}_t(\boldsymbol{\theta})$ where $\widetilde{\kappa}_{2t}(\boldsymbol{\theta}) \equiv 1$, and WLS corresponds to a score when $\widetilde{\kappa}_{2t}(\boldsymbol{\theta})$ is replaced by a GARCH(1,1) estimate of the volatility of $\widetilde{\epsilon}_t(\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is the first-step LSE of the pseudo-true value.

We have done other experiments with larger sample sizes and different values of the parameters. Table 3 shows the results but for simulations of length $n = 8000$. The first panel of the table, devoted to the size, has the same DGP as Table 2 under $\mathbf{H_0}$. For the second panel corresponding to the power we changed the DGP to have a non-degenerated power:

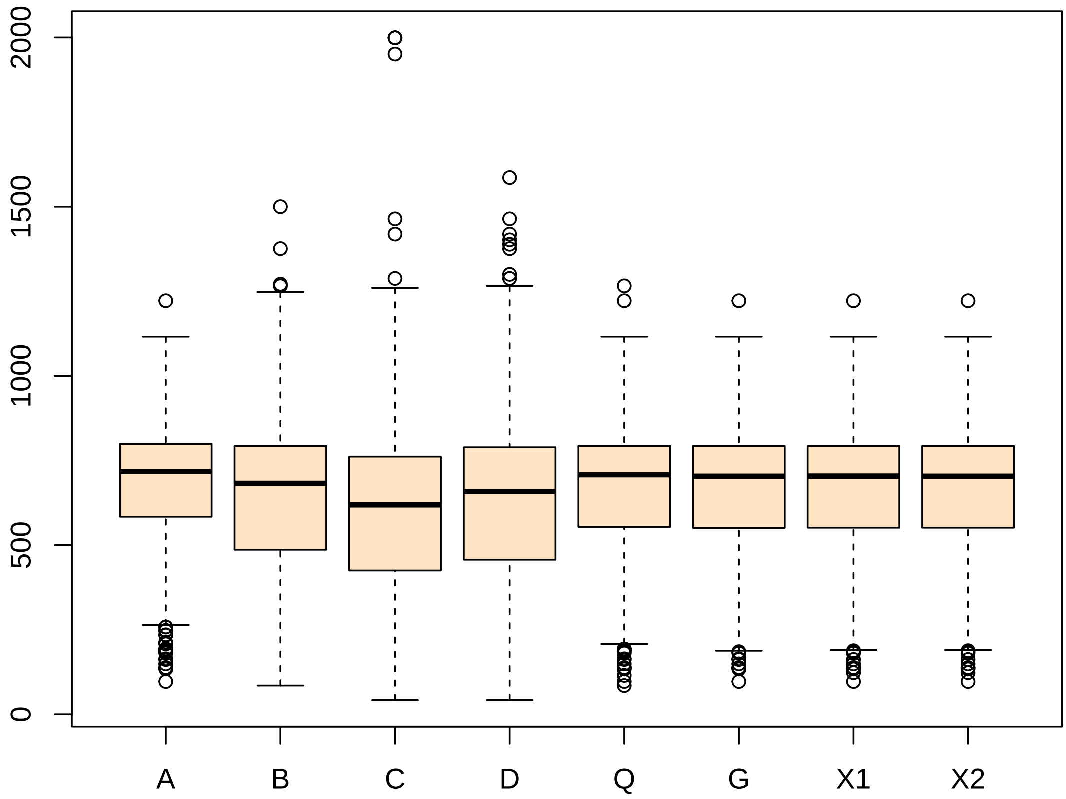

Figure 2: Distributions of the change point estimates

we took $(c, a, b) = (0.01, -0.5, 0.89)$ for $t = 1, \dots, 3200$ and $(c, a, b) = (0.01, -0.3, 0.89)$ for $t = 3201, \dots, 8000$. Tables 2-3 show that, as expected, the first-order error is not well controlled when $\boldsymbol{I}^*$ is not estimated by a HAC estimator. The different versions of the HAC estimator are those of Andrews (1991), with pre-whitening. All performed similarly. Even with HAC estimation of $\boldsymbol{I}^*$, the type I errors remain far from their nominal values, especially for WLS in Table 2. This is not surprising since long-run variances are notoriously difficult to estimate accurately. The bottom panels of Tables 2-3 confirm the ability of the tests to detect breaks, even in a misspecified conditional mean model. The WLS test appears to be slightly more powerful in Table 2, but this does not hold in other settings (see Table 3). For a well-specified conditional mean, Theorem 2 implies that WLS should be more powerful than LSE, but the result is likely to be false in misspecified models.

	$\mathbf{H_0}$											
	no HAC		spectral		Bartlett		Parzen		Tukey H.		Quadratic	
	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS
1%	0.0	0.0	1.6	0.2	1.2	0.2	1.2	0.2	1.2	0.2	1.2	0.2
5%	0.1	0.1	7.2	1.9	5.3	1.9	5.6	1.9	5.5	1.9	5.4	1.9
10%	0.1	0.8	11.7	5.9	10.5	5.2	10.3	5.1	10.2	5.2	10.2	5.1
	$\mathbf{H_1}$											
	no HAC		spectral		Bartlett		Parzen		Tukey H.		Quadratic	
	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS
1%	84.5	97.6	97.8	99.5	96.8	99.5	96.8	99.5	96.8	99.5	96.8	99.5
5%	95.1	99.6	99.9	100.0	99.7	100.0	99.7	100.0	99.7	100.0	99.7	100.0
10%	97.2	99.9	100.0	100.0	99.9	100.0	99.9	100.0	99.9	100.0	99.9	100.0

Table 2: Size and power of the test when $\widetilde{\kappa}_{2t}$ is constant (LSE) or is an estimated GARCH volatility (WLS), and $\boldsymbol{I}_n^*$ is a short-run empirical variance (no HAC) or a HAC estimator.

$\mathbf{H_0}$

	no HAC		spectral		Bartlett		Parzen		Tukey H.		Quadratic	
	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS
1%	0.0	0.0	1.0	0.6	0.8	0.6	0.8	0.6	0.8	0.6	0.9	0.6
5%	0.4	0.9	4.7	4.0	4.4	3.9	4.4	3.9	4.4	3.9	4.4	3.9
10%	1.1	1.6	9.4	8.9	8.8	8.7	8.9	8.7	8.9	8.7	8.9	8.7

$\mathbf{H_1}$

	no HAC		spectral		Bartlett		Parzen		Tukey H.		Quadratic	
	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS	LSE	WLS
1%	65.2	91.8	98.3	98.8	98.1	98.8	98.1	98.8	98.1	98.8	98.1	98.8
5%	83.5	97.3	99.8	99.8	99.7	99.8	99.7	99.8	99.7	99.8	99.7	99.8
10%	88.8	98.5	100.0	99.8	100.0	99.8	100.0	99.8	100.0	99.8	100.0	99.8

Table 3: As Table 2, but for simulations of length $n = 8000$ instead of $n = 2000$, and also different parameter values under $\mathbf{H_1}$.

4.2 Application to exchange rates

As a simple real data illustration, consider the returns of the daily exchange rates of the US dollar (USD) and the Swiss franc (CHF) against the euro from 1999-01-04 to 2022-07-12 (corresponding to 6025 observations).

We estimated GARCH(1,1) models on the log-returns (i.e. ARMA(1,1) on the squared log-returns y_t) by QLEs. Tests for breaks were performed using the test statistic $\widetilde{S}_n$, for which the optimal weights were estimated by the data-driven procedure (QLIK or based on the 3 GARCH models defined in the previous section).

Figure 3 shows that there is no evidence of breaks for USD, but strong evidence of breaks for CHF. The breakpoints are September 6, 2011 and January 15, 2015. In fact, the Swiss franc exchange rate was pegged to the euro between these two dates.

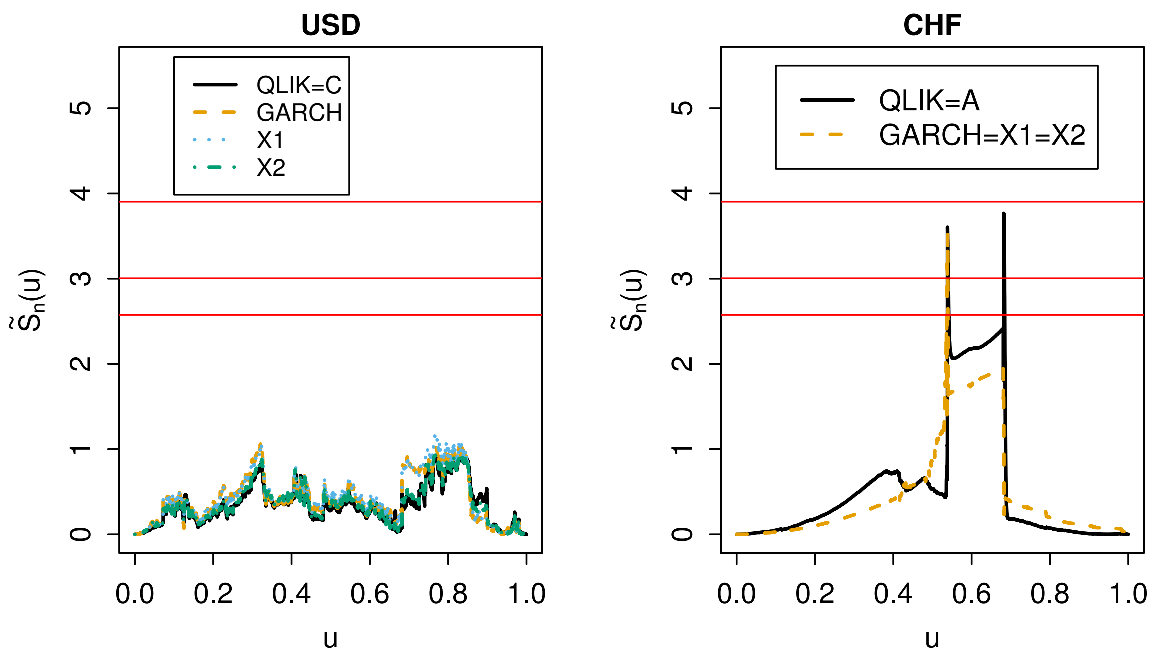

Figure 3: Trajectories of the CUSUM statistics $\widetilde{S}_n(u)$ for different QLEs and 2 exchange rates. The red lines indicate the asymptotic critical values of the tests at the 1%, 5% and 10% levels.

An alternative approach to CUSUM tests for break detection involves minimizing the OLS sum of squared residuals in a linear model where the beta coefficient is constrained to remain constant over $m+1$ subperiods of some minimum length. Bai & Perron (2003) showed that this is feasible, even for $m > 1$, using a dynamic programming algorithm implemented in the R package `strucchange`. Of course the results depend on the choice of the linear model. In our analysis, we experimented with AR(p) for different values of p. Although the dynamic programming algorithm is powerful, the computation time of this method is substantial[5] because we allowed up to $m = 5$ breaks. Thus, the AR delay was limited to $p \leq 7$. For the USD series, we used the R function `breakpoints()` with its default values. For the CHF series, we chose a minimum segment size of 10% by specifying the parameter value `h=0.1` instead of the default value `h=0.15` (which is too large to allow breaks on September 6, 2011 and January 15, 2015). The number of breaks, m, was estimated by BIC minimization. Table 4 shows that the estimated number and timing of breaks vary considerably with p. For the CHF series, a break on September 6, 2011 is often detected, but unlike the CUSUM test, the break around January 15, 2015 is not clearly identified. The most tricky output is that for the USD series, 3 breaks are often detected in Table 4, while the CUSUM test does not detect any break (see the left panel of Figure 3).

To further explore this discrepancy, Table 5 presents the QMLE estimates of GARCH(1,1) models for the four subperiods defined by the three breaks identified in the AR(p) models with $p \in \{2, \dots, 5\}$ (with each period beginning and ending five days before the detected breaks). The estimated GARCH parameters across the four periods do not show substantial variation, given their standard errors. A Wald test of the null hypothesis that the GARCH coefficients are the same across the four periods yields a p-value of 7.8%. This result does not provide sufficient evidence to reject H_0.[6] One explanation for the possible failure of the Bai-Perron test is that the regression framework on which it is based is unable to capture the volatility persistence that characterizes the dynamics of financial returns. For example, a long period of low volatility followed by a period of high volatility could be seen as evidence of breaks, as the AR(p) model for squared returns is inconsistent with this type of behaviour. To assess the possibility that the Bai-Perron test misidentifies breaks, we estimated a GARCH on the USD series and re-ran the tests from Table 4 on two independent simulations of this GARCH model without breaks. Table 5 shows that, contrary to our EF-based CUSUM tests, the Bai-Perron test does indeed detect spurious breaks. The failure of the Bai-Perron test can be explained by the irrelevance of the AR(p) model for squared returns (or, equivalently, of the ARCH(p) model for returns). On the other hand, extensions of the Bai-Perron dynamic programming algorithm to persistent models (such as GARCH) do not seem to exist.

[5] It took 8 hours on a 4 year old PC to get Table 4.

[6] For this Wald test, we write H_0 as $\boldsymbol{R}\boldsymbol{\theta}_0 = \mathbf{0}_9$, where $\boldsymbol{\theta}_0$ is the vector obtained by stacking the GARCH parameters of the four subperiods. Let $\widehat{\boldsymbol{\theta}}_i$ be the GARCH QMLE of subperiod i of length n_i. The variance of $\widehat{\boldsymbol{\theta}}_i$ is approximated by $\boldsymbol{\Sigma}_i/n_i$. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^\top, \dots, \widehat{\boldsymbol{\theta}}_4^\top)^\top$. Under H_0, the distribution of $\boldsymbol{R}\widehat{\boldsymbol{\theta}}$ is approximated by the Gaussian distribution with mean $\mathbf{0}$ and block diagonal variance $\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}'$, where $\boldsymbol{\Sigma}$ has block diagonal elements $\boldsymbol{\Sigma}_i/n_i$. The Wald-type statistic follows.

	USD								CHF							
p	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
m	3	3	3	3	3	3	2	2	0	1	2	2	0	0	1	1
Dates	04-05-17			05-06-14			05-06-14			15-01-13	11-09-06	11-09-06			11-09-06	
	08-08-07			08-12-19			08-12-19				15-01-12	15-01-09				
	12-02-02			12-06-29												

Table 4: Estimation of the number m of breaks and of their dates (year-month-day) by the R function `breakpoints()` based on the algorithm of Bai & Perron (2003). The model is an AR(p) on the squared returns.

Period	ω	α	β
1999-01-04 to 2005-06-07	0.021 (0.027)	0.048 (0.048)	0.937 (0.056)
2005-06-21 to 2008-12-12	0.000 (0.001)	0.036 (0.017)	0.972 (0.013)
2008-12-30 to 2012-06-22	0.013 (0.006)	0.000 (0.007)	0.980 (0.012)
2012-07-06 to 2022-07-11	0.001 (0.002)	0.061 (0.044)	0.960 (0.024)

Table 5: For the USD series, GARCH(1,1) models fitted to 4 sub-periods, with estimated standard deviations in brackets.

	Simulation 1								Simulation 2							
p	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
m	3	2	2	2	2	2	2	2	4	3	3	3	3	3	3	0
Dates	914				1858				904	904			904		904	
	1858				2763				2349	2349			2349		2349	
	2763								3330	3304			3280		3317	
									4760							

Table 6: As Table 4 but on two simulated trajectories of a GARCH without break.

5 Concluding remarks

This paper contributes to the time series literature on break detection by addressing models where the conditional distributions are not fully specified. We propose a novel econometric methodology based on the CUSUM of quasi-scores to detect structural breaks in the conditional mean of time series. A key advantage of this approach is its reliance on EF estimators, which require only weak, semi-parametric assumptions about the data-generating process (DGP). This stands in contrast to traditional CUSUM methods based on Fisher's scores, which necessitate full specification of the conditional distribution.

We establish the asymptotic distribution of the proposed CUSUM statistics under the null hypothesis of no change point and explore optimality considerations through LAP comparisons. Notably, the weights that are optimal for estimation in Godambe's sense are also shown to be optimal for hypothesis testing. To enhance robustness, we develop modified test versions that incorporate long-run matrix estimation, mitigating the risks of conditional mean misspecification. Finally, the empirical applications demonstrate the effectiveness of our testing procedures in identifying structural breaks in financial time series dynamics, underscoring their practical utility. An interesting avenue for future research would be extending our methodology to test for breaks in a broader range of conditional moments.

References

Andrews, D. W. (1991), 'Heteroskedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica* **59**, 817–858.

Aue, A. & Horváth, L. (2013), 'Structural breaks in time series', *Journal of Time Series Analysis* **34**, 1–16.

Bai, J. & Perron, P. (2003), 'Computation and analysis of multiple structural change models', *Journal of Applied Econometrics* **18**, 1–22.

Bera, A. K. & Bilias, Y. (2002), 'The MM, ME, ML, EL, EF andGMM approaches to estimation: a synthesis', *Journal of Econometrics* **107**, 51–86.

Bera, A. K., Bilias, Y., Simlai, P. et al. (2006), Estimating functions and equations: An essay on historical developments with applications to econometrics, *in* 'Palgrave Handbook of Econometrics', Vol. 1, pp. 427–476.

Berkes, I., Horváth, L. & Kokoszka, P. (2004), 'Testing for parameter constancy in garch (p, q) models', *Statistics & Probability Letters* **70**, 263–273.

Bibby, B. M., Jacobsen, M. & Sørensen, M. (2010), Estimating functions for discretely sampled diffusion-type models, *in* 'Handbook of Financial Econometrics: Tools and Techniques', Elsevier, pp. 203–268.

Billingsley, P. (1986), *Probability and measure*, John Wiley & Sons.

Casini, A. & Perron, P. (2019), Structural breaks in time series, *in* 'Oxford Research Encyclopedia of Economics and Finance', Oxford University Press.

Chandra, S. A. & Taniguchi, M. (2001), 'Estimating functions for nonlinear time series models', *Annals of the Institute of Statistical Mathematics* **53**, 125–141.

Christensen, B. J., Posch, O. & Van Der Wel, M. (2016), 'Estimating dynamic equilibrium models using mixed frequency macro and financial data', *Journal of Econometrics* **194**, 116–137.

Csörgö, M. & Horváth, L. (1997), *Limit theorems in change-point analysis*, Chichester: Wiley.

Den Haan, W. J. & Levin, A. T. (1997), 'A practitioner's guide to robust covariance matrix estimation', *Handbook of Statistics* **15**, 299–342.

Durbin, J. (1960), 'Estimation of parameters in time-series regression models', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **22**, 139–153.

Francq, C. & Zakoïan, J.-M. (2023), 'Optimal estimating function for weak location-scale dynamic models', *Journal of Time Series Analysis* **44**, 533–555.

Godambe, V. P. (1960), 'An optimum property of regular maximum likelihood estimation', *The Annals of Mathematical Statistics* **31**, 1208–1211.

Godambe, V. P. (1985), 'The foundations of finite sample estimation in stochastic processes', *Biometrika* **72**, 419–428.

Godambe, V. P. & Heyde, C. C. (1987), 'Quasi-likelihood and optimal estimation', *International Statistical Review* **55**, 231–244.

Godambe, V. P. & Thompson, M. E. (2009), Estimating functions and survey sampling, *in* 'Handbook of Statistics', Vol. 29, Elsevier, pp. 83–101.

Hansen, B. E. (2001), 'The new econometrics of structural change: Dating breaks in us labor productivity', *Journal of Economic Perspectives* **15**, 117–128.

Hansen, L. P. (1982), 'Large sample properties of generalized method of moments estimators', *Econometrica* **50**, 1029–1054.

Hansen, P. R., Lunde, A. & Voev, V. (2014), 'Realized beta garch: A multivariate garch model with realized measures of volatility', *Journal of Applied Econometrics* **29**, 774–799.

Herrndorf, N. (1984), 'A functional central limit theorem for weakly dependent sequences of random variables', *The Annals of Probability* **12**, 141–153.

Heyde, C. C. (1997), *Quasi-likelihood and its application: a general approach to optimal parameter estimation*, Springer New York.

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

Horváth, L. & Parzen, E. (1994), 'Limit theorems for fisher-score change processes', *Lecture Notes-Monograph Series* pp. 157–169.

Horváth, L. & Rice, G. (2024), *Change Point Analysis for Time Series*, Springer.

Jacod, J. & Sørensen, M. (2018), 'A review of asymptotic theory of estimating functions', *Statistical Inference for Stochastic Processes* **21**, 415–434.

Kutoyants, Y. A. (2016), 'On score-functions and goodness-of-fit tests for stochastic processes', *Mathematical Methods of Statistics* **25**, 99–120.

Lee, S., Ha, J., Na, O. & Na, S. (2003), 'The cusum test for parameter change in time series models', *Scandinavian Journal of Statistics* **30**, 781–796.

Liang, K.-Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

Ling, S. (2016), 'Estimation of change-points in linear and nonlinear time series models', *Econometric Theory* **32**, 402–430.

Negri, I. & Nishiyama, Y. (2017), 'Z-process method for change point problems with applications to discretely observed diffusion processes', *Statistical Methods & Applications* **26**, 231–250.

Newey, W. K. & West, K. D. (1987), 'A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix', *Econometrica* **55**, 703–708.

Nyblom, J. (1989), 'Testing for the constancy of parameters over time', *Journal of the American Statistical Association* **84**, 223–230.

Page, E. S. (1955), 'A test for a change in a parameter occurring at an unknown point', *Biometrika* **42**, 523–527.

Ploberger, W. & Krämer, W. (1992), 'The CUSUM test with ols residuals', *Econometrica* **60**, 271–285.

Shao, X. & Zhang, X. (2010), 'Testing for change points in time series', *Journal of the American Statistical Association* **105**, 1228–1240.

Tauchen, G. (1986), 'Statistical properties of generalized method-of-moments estimators of structural parameters obtained from financial market data', *Journal of Business & Economic Statistics* **4**, 397–416.

Vinod, H. (1997), 'Using godambe-durbin estimating functions in econometrics', *Lecture Notes-Monograph Series* pp. 215–237.

Appendix: Discussion, Examples and Proofs

A Discussion and examples

In this section, we complement the assumptions spelt out in Section 2 by offering several illustrative examples; we also refer to the article by Francq & Zakoïan (2023) for further examples and discussion.

A.1 Examples of QLE

Example 1 (Examples of QLE)**.** *When $\widetilde{\kappa}_{2t}$ is a non zero constant, it is easy to see that the solution of* (8) *is the Least Squares (LS) estimator. When $y_t \geq 0$ and $\widetilde{\kappa}_{2t}$ is proportional to $\widetilde{m}_t(\boldsymbol{\theta}) > 0$ (respectively, to $\widetilde{m}_t^2(\boldsymbol{\theta}) > 0$), then it can be verified that the solution of* (8) *is the Poisson (respectively, exponential) QMLE, obtained by minimising*

$$\sum_{t=1}^{n} \widetilde{m}_t(\boldsymbol{\theta}) - y_t \log \widetilde{m}_t(\boldsymbol{\theta}) \quad \textit{(respectively } \textstyle\sum_{t=1}^{n} y_t/\widetilde{m}_t(\boldsymbol{\theta}) + \log \widetilde{m}_t(\boldsymbol{\theta})\textit{)}.$$

Example 2 (An example when QLE is the MLE)**.** *Assume that the distribution of y_t given $\mathcal{F}_{t-1}$ belongs to the one-parameter exponential family. This means that, with respect to a σ-finite measure, the conditional distribution admits a density of the form*

$$g_{m_t}(y) = k(y) \exp\left\{\eta(m_t)y - a(m_t)\right\}, \tag{21}$$

for some positive function $k(\cdot)$, and twice differentiable functions $\eta(\cdot)$ and $a(\cdot)$. It is well-known that $\eta'(m_t) = a'(m_t)/m_t = 1/\sigma_t^2$. It follows that

$$\frac{\partial \log g_{m_t(\boldsymbol{\theta})}(y_t)}{\partial \theta} = \frac{\partial m_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\epsilon_t(\boldsymbol{\theta})}{\sigma_t^2(\boldsymbol{\theta})}.$$

Hence it follows that the QLE coincides with the MLE (only approximately when $m_t \neq \widetilde{m}_t$).

Example 3 (Link with GMM estimation)**.** *The Generalised Method of Moments (GMM) method developed by Hansen (1982) is based on moment conditions of the form*

$$\mathbb{E}\boldsymbol{g}_t(\boldsymbol{\theta}) = \mathbf{0},$$

iff $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, where $\boldsymbol{g}_t(\boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}^m$ with $m \geq d$ and

$$\boldsymbol{g}_t(\boldsymbol{\theta}) = \boldsymbol{z}_t \epsilon_t(\boldsymbol{\theta}),$$

and $\boldsymbol{z}_t$ is an m-dimensional vector of instruments belonging to $\mathcal{F}_{t-1}$. Let

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\theta}) = n^{-1} \sum_{t=1}^{n} \boldsymbol{g}_t(\boldsymbol{\theta}),$$

be an empirical estimator of $\mathbb{E}\boldsymbol{g}_t(\boldsymbol{\theta})$. The GMM estimator minimises $\overline{\boldsymbol{g}}_n'(\boldsymbol{\theta})\widehat{\boldsymbol{S}}^{-1}\overline{\boldsymbol{g}}_n(\boldsymbol{\theta})$, where $\widehat{\boldsymbol{S}}$ is a positive definite weight matrix. The first-order conditions give the EF

$$\sum_{t=1}^{n} \widehat{\boldsymbol{\Omega}}_t(\boldsymbol{\theta}) \widehat{\boldsymbol{S}}^{-1} \overline{\boldsymbol{g}}_n(\theta) = 0,$$

where

$$\widehat{\boldsymbol{\Omega}}_t(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{g}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}, \quad \text{and} \quad \widehat{\boldsymbol{\Omega}}(\boldsymbol{\theta}) = n^{-1} \sum_{t=1}^{n} \widehat{\boldsymbol{\Omega}}_t(\boldsymbol{\theta}).$$

Thus, the GMM estimators are QLEs, and Godambe's results imply that the optimal QLE is always at least as efficient (in Godambe's sense and asymptotically) as the optimal GMM. Indeed, as mentioned in the Introduction, Christensen et al. (2016) show that in general the optimal QLE is strictly more efficient than the optimal GMM.

A.2 Examples of weighing sequences $\widetilde{\kappa}_{2t}$

Example 4 (Selection of $\widetilde{\kappa}_{2t}$ by QLIK)**.** *In view of Example 1, several natural candidates exist for the weighting sequence, such as: $\widetilde{\kappa}_{2t} \propto 1$, $\widetilde{\kappa}_{2t} \propto \widetilde{m}_t(\theta)$ (for positive data) or $\widetilde{\kappa}_{2t} \propto \widetilde{m}_t^2(\boldsymbol{\theta})$, among an infinite number of other possibilities. Thus, assume we want to select the weights over a finite set of potential weighting sequences, say $\left\{\widetilde{\kappa}_{2t}^{(h)}(\boldsymbol{\theta})\right\}$ for $h \in \{1, \ldots, \mathcal{H}\}$. The optimal weighting* *sequence is the conditional variance, up to any non zero multiplicative constant; further, the conditional variance is solution to the quasi-likelihood (QLIK) loss function. Hence, Francq & Zakoïan (2023) propose a data-driven selection of the weights by minimizing over h the empirical QLIK loss function defined as*

$$QLIK_n\left(\widetilde{\kappa}_{2t}^{(h)}\left(\widehat{\theta}\right)\right) = \frac{1}{n}\sum_{t=1}^{n}\left\{\frac{\widetilde{\epsilon}_t^2\left(\widehat{\boldsymbol{\theta}}\right)}{\widehat{c}_n^{(h)}\widetilde{\kappa}_{2t}^{(h)}\left(\widehat{\boldsymbol{\theta}}\right)} + \log\left(\widehat{c}_n^{(h)}\widetilde{\kappa}_{2t}^{(h)}\left(\widehat{\boldsymbol{\theta}}\right)\right)\right\},$$

where

$$\widehat{c}_n^{(h)} = \frac{1}{n}\sum_{t=1}^{n}\frac{\widetilde{\epsilon}_t^2\left(\widehat{\boldsymbol{\theta}}\right)}{\widetilde{\kappa}_{2t}^{(h)}\left(\widehat{\boldsymbol{\theta}}\right)},$$

and $\widehat{\boldsymbol{\theta}}$ is a first step estimator of $\boldsymbol{\theta}_0$.

Example 5 (Estimation of $\widetilde{\kappa}_{2t}$ by GARCH-X)**.** *As a second example, another natural estimator of the conditional variance can be based on fitting a GARCH-type model on $\widetilde{\epsilon}_t = y_t - \widetilde{m}_t\left(\widehat{\boldsymbol{\theta}}\right)$, for $t = 1, \ldots, n$, where $\widehat{\boldsymbol{\theta}}$ is a first step consistent estimator of $\boldsymbol{\theta}_0$. This leads to the simple GARCH(1,1) estimator*

$$\widetilde{\kappa}_{2t} = \widehat{\omega} + \widehat{\alpha}\widetilde{\epsilon}_{t-1}^2 + \widehat{\beta}\widetilde{\kappa}_{2,t-1} \tag{22}$$

or to extended GARCH-X estimators like

$$\widetilde{\kappa}_{2t} = \widehat{\omega} + \widehat{\alpha}\widetilde{\epsilon}_{t-1}^2 + \widehat{\beta}\widetilde{\kappa}_{2,t-1} + \widehat{\pi}_1\left|\widetilde{m}_t\left(\widehat{\boldsymbol{\theta}}\right)\right| \tag{23}$$

or

$$\widetilde{\kappa}_{2t} = \widehat{\omega} + \widehat{\alpha}\widetilde{\epsilon}_{t-1}^2 + \widehat{\beta}\widetilde{\kappa}_{2,t-1} + \widehat{\pi}_1\left|\widetilde{m}_t\left(\widehat{\boldsymbol{\theta}}\right)\right| + \widehat{\pi}_2\widetilde{m}_t^2\left(\widehat{\boldsymbol{\theta}}\right). \tag{24}$$

For instance, (24) *allows weights proportional to the conditional mean or its square, and thus can target the Poisson and exponential QMLEs (see also Example 1).*

A.3 Examples and discussion of Assumptions A1-A8 and B1-B5

Example 6 (Discussion of Assumptions **A1-A8**)**.** *Assumptions* **A1-A8** *can be made explicit and studied for particular models, such as ARMA or GARCH. As an illustration, assume an INGARCH model, obtained when y_t given $\mathcal{F}_{t-1}$ follows a Poisson distribution with intensity parameter*

$$m_t = c_0 + a_0 y_{t-1} + b_0 m_{t-1},$$

with obvious notation. In this case, it is known that Assumption **A1** *holds true when $a_0 + b_0 < 1$, and that the strictly stationary solution of the INGARCH model even admits moments at any* *order. It is easy to see that, when* $|b| < 1$ *for all* $\boldsymbol{\theta} \in \Theta$*, Assumption* **A2** *holds with*

$$K_t = \sup_{\boldsymbol{\theta} \in \Theta} \{|a||y_0| + |b||m_0(\boldsymbol{\theta})|\} ;$$

indeed, in this example $K_t = K$*, but in general this variable can be time-varying in the case of other models. Similarly, Assumption* **A7** *can be shown to hold (although with another expression of* $K_t = K$*). By the same token, it can be verified that Assumption* **A5** *holds if*

$$a_0 \neq 0 \quad \textit{and} \quad \inf_{\boldsymbol{\theta} \in \Theta} c > 0.$$

Further, since y_t *admits moments of any order, Assumption* **A8** *is always satisfied. Is is also clear that all the other assumptions hold true for many weighting sequences* $\widetilde{\kappa}_{2t}$*.*

Example 7 (Conditional mean of the weak AR(1) process)**.** *Suppose that*

$$y_t = \theta_0 y_{t-1} + \epsilon_t, \quad i \in \mathbb{Z}, \quad |\theta_0| < 1,$$

where ϵ_t *is strictly stationary, ergodic and satisfies* $\mathbb{E}_{t-1}(\epsilon_t) = 0$ *(note the notational change, by unbolding the scalars). Then, Assumption* **A1** *is immediately satisfied and it is clear that, with* $\Theta = [-1, 1]$*, Assumptions* **A2** *and* **A4** *are also satisfied. Moreover, we have*

$$\mathbb{E}\Upsilon_t(\theta) = \mathbb{E}\left(\frac{y_{t-1}^2}{\kappa_{2t}(\theta)}\right)(\theta - \theta_0),$$

showing that Assumption **A3** *holds true. Under the assumption* $\sigma_t^2(\theta_0) > 0$ *in the first part of* **A5***, the second part holds true. It can be verified via standard, if tedious, arguments, that a sufficient condition for Assumption* **A8** *to hold is*

$$\mathbb{E}\sup_{\theta}\left(\frac{\sigma_t^2(\theta_0)}{\kappa_{2t}^2(\theta)}y_{t-1}^2 + \frac{y_{t-1}^4}{\kappa_{2t}^2(\theta)} + \left|\frac{1}{\kappa_{2t}^2(\theta)}\frac{\partial \kappa_{2t}}{\partial \theta}\right|(|y_{t-1}| + y_{t-1}^2)\right) < \infty.$$

It should be noted that if, for instance, κ_{2t} *is of the form* $a + by_{t-1}^2$ *with* $a, b > 0$*, and similarly for* $\sigma_t^2(\theta_0)$*, the latter conditions may only require* $\mathbb{E}y_t^2 < \infty$*. Finally, note that*

$$\boldsymbol{I} = E\left(\frac{\sigma_t^2(\boldsymbol{\theta}_0)}{\kappa_{2t}^2(\boldsymbol{\theta}_0)}y_{t-1}^2\right).$$

Example 8 (Illustration of Assumptions **B1-B5**)**.** *Consider the simple case where*

$$y_t^{(j)} = \theta_j y_{t-1}^{(j)} + \epsilon_t$$

with ϵ_t *a strong white noise process, and* $j = 1, 2$*. If* $|\theta_j| < 1$ *for* $j = 1, 2$*, then Assumption* **B1** *holds. Taking* $\Theta \subset (-1, 1)$*, and assume for example that* $\widetilde{\kappa}_{2t} = \kappa_{2t} \propto 1$*,* $\widetilde{m}_1(\theta) = 0$*, and*

$$\widetilde{m}_t(\theta) = m_t(\theta) = \theta y_{t-1} \quad \textit{for} \quad t \geq 2,$$

then Assumption **B3** *readily holds. To show this result, we note that* $y_t = y_t^{(1)}$ *for* $i \leq [nu_0]$, *that* $y_{[nu_0]+1}^{(j)} = \sum_{h=0}^{\infty} \theta_j^h \epsilon_{[nu_0]+1-h}$ *and* $|y_{[nu_0]+1+k}^{(2)} - y_{[nu_0]+1+k}| = |\theta_2|^k |y_{[nu_0]+1}^{(2)} - y_{[nu_0]+1}^{(1)}|$ *for* $k \geq 0$. *So we can choose* $K_t^{(1)} = |y_0|$ *and*

$$K_t^{(2)} = \sum_{h=0}^{\infty} \left|\theta_1^h - \theta_2^h\right| \left|\epsilon_{[nu_0]+1-h}\right|,$$

in Assumption **B3**. *Assumption* **B4** *is always satisfied and if* $\mathbb{E}\epsilon_t^4 < \infty$ *then Assumption* **B5** *also holds. Finally note that Assumption* **B2** *holds with*

$$\mathbb{E}\left\{\Upsilon_t^{(1)}(\theta)\right\} = (\theta_j - \theta)\frac{\mathbb{E}\epsilon_1^2}{1-\theta_j^2}, \qquad \theta_0^{\star} = \frac{\dfrac{u_0\theta_1}{1-\theta_1^2} + \dfrac{(1-u_0)\theta_2}{1-\theta_2^2}}{\dfrac{u_0}{1-\theta_1^2} + \dfrac{1-u_0}{1-\theta_2^2}}.$$

Example 9 (Illustration of Assumption **B2**)**.** *Assume that* $y_1, \dots, y_n$ *are independent and Gaussian with variance* σ^2, *and mean equal to* θ_1 *when* $t \leq \lfloor nu_0 \rfloor$ *and to* θ_2 *when* $t > \lfloor nu_0 \rfloor$. *We have*

$$\widetilde{T}_n(u) = T_n(u) = n^{-1/2} \sum_{t=1}^{\lfloor (n+1)u \rfloor} (y_t - \overline{y}),$$

where $\overline{y} = n^{-1}\sum_{t=1}^{n} y_t$. *Then Assumption* **B2** *holds with* $\theta_0^* = u_0\theta_1 + (1-u_0)\theta_2$. *We also have* $\boldsymbol{\Upsilon}_t^{(j)}(\theta) = y_t^{(j)} - \theta$ *for* $j = 1, 2$. *Thus*

$$\boldsymbol{\Delta}(\theta_1, \theta_2) = \theta_2 - \theta_1.$$

A.4 Link between the weighted CUSUM and the LR statistics

Inspired by the example given by Horváth & Rice (2024), consider a sequence of independent and Gaussian vectors $Y_1, \dots, Y_k, \dots, Y_n$ such that $Y_t \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for $t \leq k$ and $Y_t \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ for $t > k$, where $\boldsymbol{\Sigma}$ is a known non singular variance matrix. Let the null $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ and the alternative $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. The unknown parameters of interest are k and $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$. Note that, at $\boldsymbol{\theta}_0 = (\boldsymbol{\mu}_0, \boldsymbol{\mu}_0) \in H_0$, the likelihood $L_n(Y_1, \dots, Y_n; \boldsymbol{\theta}_0, k)$ does not depend on k. With obvious notations, the standard likelihood ratio leads to reject H_0 for large values of $LR = \sup_{1 \leq k \leq n} LR(k)$, where

$$LR(k) = \log \frac{L_n(Y_1, \dots, Y_n; \widehat{\boldsymbol{\theta}}, k)}{L_n(Y_1, \dots, Y_n, \widehat{\boldsymbol{\theta}}_0, k)} = \frac{nk}{2(n-k)}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^\top \boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0),$$

up to unimportant additive constants, noting that $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_1 = \frac{n}{n-k}(\widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1)$ and $\widehat{\boldsymbol{\mu}}_2 - \widehat{\boldsymbol{\mu}}_0 = \frac{k}{n-k}(\widehat{\boldsymbol{\mu}}_0 - \widehat{\boldsymbol{\mu}}_1)$. This likelihood ratio is directly related to the weighted CUSUM by

$$2LR = \sup_{u \in (0,1)} \frac{\widetilde{S}_n(u)}{u(1-u)} + o_P(1).$$

A.5 Examples of local breaks, misspecification and change-point estimation

Example 10 (Example of a local break). *The simplest example of a local break is obtained by assuming that $y_1, \dots, y_n$ are independent and Gaussian with variance σ^2, and that $y_t = y_{t,n}$ has mean $\theta_0 + \delta_1/\sqrt{\lfloor nu_0 \rfloor}$ when $t \le \lfloor nu_0 \rfloor$ and $\theta_0 + \delta_2/\sqrt{n - \lfloor nu_0 \rfloor}$ when $t > \lfloor nu_0 \rfloor$. Note that we then have*

$$\frac{1}{\sqrt{\lfloor nu_0 \rfloor}} \sum_{t=1}^{\lfloor nu_0 \rfloor} (y_t - \theta_0) \sim \mathcal{N}\left(\delta_1, \sigma^2\right), \qquad \frac{1}{\sqrt{n - \lfloor nu_0 \rfloor}} \sum_{t=\lfloor nu_0 \rfloor + 1}^{n} (y_t - \theta_0) \sim \mathcal{N}\left(\delta_2, \sigma^2\right),$$

and

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} (y_t - \theta_0) \sim \mathcal{N}\left(\delta_3, \sigma^2\right), \quad \delta_3 = \sqrt{u_0}\delta_1 + \sqrt{1-u_0}\delta_2.$$

note that, in this simple example, $\overline{y} = n^{-1}\sum_{t=1}^n y_t$ is the Q(M)LE of θ_0 (under the null $\delta_1 = \delta_2 = 0$ of no local break), $\widetilde{T}_n(u) = T_n(u) = n^{-1/2}\sum_{t=1}^{\lfloor (n+1)u \rfloor}(y_t - \overline{y})$ is the usual CUSUM process, and

$$\left(\widetilde{T}_n(u)\right)' \widehat{\boldsymbol{I}}^{-1} \widetilde{T}_n(u) = \sup_{u \in (0,1)} \frac{1}{n\widehat{\sigma}_y^2} \left\{ \sum_{t=1}^{\lfloor (n+1)u \rfloor} (y_t - \overline{y}) \right\}^2, \quad \widehat{\sigma}_y^2 = \frac{1}{n} \sum_{t=1}^{n} (y_t - \overline{y})^2.$$

Example 11 (Misspecification). *Assume, perhaps wrongly, that $m_t(\boldsymbol{\theta}) = a + by_{t-1}$ with $\boldsymbol{\theta} = (a,b)'$. We then have*

$$\mathring{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta}) = \begin{pmatrix} 1 \\ y_{t-1} \end{pmatrix} \frac{1}{\kappa_{2t}} (y_t - a - by_{t-1})$$

Then Assumptions **A3*** *and* **A5*** *are satisfied with*

$$\mathring{\theta}_0 = \boldsymbol{A}^{-1}\boldsymbol{b}, \quad \boldsymbol{b} = \begin{pmatrix} \mathbb{E}\left(\frac{y_t}{\kappa_{2t}}\right) \\ \mathbb{E}\left(\frac{y_t y_{t-1}}{\kappa_{2t}}\right) \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} \mathbb{E}\left(\frac{1}{\kappa_{2t}}\right) & \mathbb{E}\left(\frac{y_{t-1}}{\kappa_{2t}}\right) \\ \mathbb{E}\left(\frac{y_{t-1}}{\kappa_{2t}}\right) & \mathbb{E}\left(\frac{y_{t-1}^2}{\kappa_{2t}}\right) \end{pmatrix}$$

when $\boldsymbol{b}$ and $\boldsymbol{A}$ exist and $\boldsymbol{A} = -\mathring{\boldsymbol{J}}$ is invertible (which is for instance the case when κ_{2t} is constant and $Var(y_t) > 0$).

Assumption A1 is satisfied if, for instance, $\kappa_{2t} \equiv \kappa > 0$, $\|y_1\|_{4+2\nu} < \infty$ and $\sum_{h=1}^{\infty}\{\alpha_y(h)\}^{\nu/(2+\nu)} < \infty$ for some $\nu > 0$, where $\{\alpha_y(h)\}$ denotes the sequence of the $\alpha-$mixing coefficients of (y_t). A non constant weighting sequence (κ_{2t}) can reduce the moment requirement. In particular, if κ_{2t} is the volatility of an ARCH(1), or more generally $\kappa_{2t} > c_1 + c_2 y_{t-1}^2$ with positive constants c_1 and c_2, then only $\|y_1\|_{2+\nu} < \infty$ is required.*

Example 12 (Unconditional mean of Gaussian variables). *Assume that $y_1, \dots, y_n$ are independent and Gaussian with variance σ^2, and mean equal to θ_1 when $t \le [nu_0]$ and to θ_2 when*

$t > [nu_0]$. We have $\widetilde{T}_n(u) = T_n(u) = n^{-1/2}\sum_{t=1}^{[nu]}(y_t - \overline{y})$ where $\overline{y} = n^{-1}\sum_{t=1}^{n} y_t$. Assumption **B2** *is thus satisfied with $\theta_0^* = u_0\theta_1 + (1-u_0)\theta_2$. We also have $\boldsymbol{\Upsilon}_t^{(i)}(\theta) = y_t^{(i)} - \theta$ for $i = 1, 2$. Thus $\Delta(\theta_1, \theta_2) = \theta_2 - \theta_1$.*

B Technical results and proofs

Denote by $X_n \xrightarrow{\mathcal{L}} X$, or simply $X_n \xrightarrow{\mathcal{L}} P_X$, when a sequence of random vectors X_n converges in distribution to a random vector X of distribution P_X. For a sequence of random functions $\{X_n(u), u \in [0,1]\}$ tending weakly to $\{X(u), u \in [0,1]\}$, we denote $X_n(\cdot) \Longrightarrow X(\cdot)$. The following result establishes the asymptotic distribution of the QLE.

Lemma 1 (Francq & Zakoïan (2023))**.** *Under Assumptions* **A1**-**A8***, a QLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$, such that*

$$\sum_{t=1}^{n} \widetilde{\boldsymbol{\Upsilon}}_t(\widehat{\boldsymbol{\theta}}) = 0, \qquad \widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta}) = \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\widetilde{\epsilon}_t(\boldsymbol{\theta})}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})},$$

exists[7] *for n large enough, and as $n \to \infty$ we have $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ a.s. and*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = -\boldsymbol{J}^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) + o_P(1) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma} := \boldsymbol{J}^{-1}\boldsymbol{I}\boldsymbol{J}^{-1}\right)$$

with

$$\boldsymbol{J} = E\left(\frac{-1}{\kappa_{2t}(\boldsymbol{\theta}_0)}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top}\right), \qquad \boldsymbol{I} = E\left(\frac{\sigma_t^2(\boldsymbol{\theta}_0)}{\kappa_{2t}^2(\boldsymbol{\theta}_0)}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top}\right).$$

If $\kappa_{2t}(\boldsymbol{\theta}_0) \propto \sigma_t^2(\boldsymbol{\theta}_0)$ then the asymptotic variance of the QLE, which is equal to

$$\boldsymbol{\Sigma}_{op} = \left\{E\frac{1}{\sigma_t^2(\boldsymbol{\theta}_0)}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\frac{\partial m_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top}\right\}^{-1},$$

is optimal in the sense that $\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{op}$ is semi positive definite.

B.1 Proof of Theorem 1

A Taylor expansion around $\boldsymbol{\theta}_0$ yields

$$\boldsymbol{T}_n(u) = \frac{1}{\sqrt{n}}\sum_{t=1}^{[nu]} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) + u\left(\frac{1}{nu}\sum_{t=1}^{[nu]}\frac{\partial}{\partial \boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\widehat{\boldsymbol{\theta}}^*)\right)\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

where $\widehat{\boldsymbol{\theta}}^*$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. By Lemma 1 we deduce (12). It follows from the functional central limit theorem for stationary, ergodic martingale differences (see e.g. Theorem 18.3 of Billingsley 1986) that

$$\boldsymbol{T}_n^0(\cdot) \Longrightarrow \boldsymbol{T}(\cdot).$$

[7] uniqueness has been shown under the extra contraction assumption **A10** of Francq & Zakoïan (2023).

It remains to show (13). Let $V_k(\boldsymbol{\theta}_0)$ be the ball of center $\boldsymbol{\theta}_0$ and radius $1/k$. The strong consistency of $\widehat{\boldsymbol{\theta}}$ entails that

$$\|\boldsymbol{R}_n(u)\| \leq \left\| \frac{1}{n} \sum_{t=1}^{[nu]} \left(\frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - \boldsymbol{J} \right) \right\| + \sup_{\boldsymbol{\theta} \in V_k(\boldsymbol{\theta}_0)} \left\| \frac{1}{n} \sum_{t=1}^{[nu]} \left(\frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) \right) \right\|.$$

We have,

$$\sup_{\boldsymbol{\theta} \in V_k(\boldsymbol{\theta}_0)} \sup_{u \in (0,1)} \left\| \frac{1}{n} \sum_{t=1}^{[nu]} \left(\frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) \right) \right\| \leq \frac{1}{n} \sum_{t=1}^{n} \sup_{\boldsymbol{\theta} \in V_k(\boldsymbol{\theta}_0)} \left\| \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) \right\|$$

which tends, as $n \to \infty$, to

$$E \sup_{\boldsymbol{\theta} \in V_k(\boldsymbol{\theta}_0)} \left\| \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) \right\|.$$

by the ergodic theorem. By Fatou's lemma, and using the continuity assumptions in **A2** and **A7**, the latter expectation can be made arbitrarily small by choosing k sufficiently large. Let (u_n) be a deterministic sequence converging to infinity slower than n (i.e. $n/u_n \to \infty$). Let $\boldsymbol{Y}_t = \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - \boldsymbol{J}$. We have

$$\sup_{u \in (0,1)} \left\| \frac{1}{n} \sum_{t=1}^{[nu]} \left(\frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0) - \boldsymbol{J} \right) \right\| \leq \sup_{1 \leq k \leq u_n} \left\| \frac{1}{n} \sum_{t=1}^{k} \boldsymbol{Y}_t \right\| + \sup_{u_n \leq k \leq n} \left\| \frac{1}{n} \sum_{t=1}^{k} \boldsymbol{Y}_t \right\|.$$

We have $\frac{1}{k} \sum_{t=1}^{k} \boldsymbol{Y}_t \to 0$ a.s. as $k \to \infty$, hence the last term in the previous inequality converges to 0 a.s. Moreover, by Markov inequality, for any $\iota > 0$,

$$P\left(\sup_{1 \leq k \leq u_n} \left\| \frac{1}{n} \sum_{t=1}^{k} \boldsymbol{Y}_t \right\| > \iota \right) \leq P\left(\frac{1}{n} \sum_{t=1}^{u_n} \|\boldsymbol{Y}_t\| > \iota \right) \leq \frac{u_n}{n\iota} E \|\boldsymbol{Y}_1\| \to 0,$$

as $n \to \infty$, from which we deduce that the first convergence in (13) holds.

Now, we have

$$\begin{aligned}
\sup_{u \in (0,1)} \|\boldsymbol{T}_n(u) - \widetilde{\boldsymbol{T}}_n(u)\| &\leq \sup_{u \in (0,1)} \sup_{\boldsymbol{\theta} \in \Theta} \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu]} \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\widetilde{\epsilon}_t(\boldsymbol{\theta})}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})} - \frac{\partial m_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\epsilon_t(\boldsymbol{\theta})}{\kappa_{2t}(\boldsymbol{\theta})} \right\| \\
&\leq \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sup_{\boldsymbol{\theta} \in \Theta} \left(\left\| \frac{\partial m_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \frac{|\epsilon_t(\boldsymbol{\theta})|}{\kappa_{2t}(\boldsymbol{\theta})} \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sup_{\boldsymbol{\theta} \in \Theta} \left(|m_t(\boldsymbol{\theta}) - \widetilde{m}_t(\boldsymbol{\theta})| \left\| \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \frac{1}{\kappa_{2t}(\boldsymbol{\theta})} \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \sup_{\boldsymbol{\theta} \in \Theta} \left(\left| \frac{1}{\kappa_{2t}(\boldsymbol{\theta})} - \frac{1}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})} \right| \left\| \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| |\widetilde{\epsilon}_t(\boldsymbol{\theta})| \right).
\end{aligned}$$

By Assumptions **A6-A7**, the first term in the right-hand side is bounded by

$$\frac{1}{\underline{\kappa}\sqrt{n}}\sum_{t=1}^{\infty}\sup_{\boldsymbol{\theta}\in\Theta}|\epsilon_t(\boldsymbol{\theta})|K_t\rho^t = O\left(\frac{1}{\sqrt{n}}\right), \quad a.s.$$

because the summands have finite s-th order moment by **A4**. The same upper bound holds for the other two terms of the right-hand side of the previous inequality. Hence the second convergence in (13) is established. □

B.2 Proof of Theorem 2

Since $\boldsymbol{\Upsilon}_t = \boldsymbol{\Upsilon}_{t,n}$ is such that

$$\frac{1}{\sqrt{nu_0}}\sum_{t=1}^{[nu_0]}\boldsymbol{\Upsilon}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\boldsymbol{J}\boldsymbol{\delta}_1, \boldsymbol{I}\right), \quad \frac{1}{\sqrt{n(1-u_0)}}\sum_{t=[nu_0]+1}^{n}\boldsymbol{\Upsilon}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\boldsymbol{J}\boldsymbol{\delta}_2, \boldsymbol{I}\right),$$

for $u \leq u_0$ we have

$$\frac{1}{\sqrt{nu}}\sum_{t=1}^{[nu]}\left(\boldsymbol{\Upsilon}_t - \frac{1}{\sqrt{nu_0}}\boldsymbol{J}\boldsymbol{\delta}_1\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \boldsymbol{I}\right),$$

and thus

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{[nu]}\boldsymbol{\Upsilon}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{u}{\sqrt{u_0}}\boldsymbol{J}\boldsymbol{\delta}_1, u\boldsymbol{I}\right), \quad \frac{1}{\sqrt{n}}\sum_{t=[nu]+1}^{[nu_0]}\boldsymbol{\Upsilon}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{u_0-u}{\sqrt{u_0}}\boldsymbol{J}\boldsymbol{\delta}_1, (u_0-u)\boldsymbol{I}\right),$$

$$\frac{1}{\sqrt{n}}\sum_{t=[nu_0]+1}^{n}\boldsymbol{\Upsilon}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\sqrt{1-u_0}\boldsymbol{J}\boldsymbol{\delta}_2, (1-u_0)\boldsymbol{I}\right),$$

and for $u \geq u_0$ we have

$$\frac{1}{\sqrt{n}}\sum_{t=[nu_0]+1}^{[nu]}\boldsymbol{\Upsilon}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{u-u_0}{\sqrt{1-u_0}}\boldsymbol{J}\boldsymbol{\delta}_2, (u-u_0)\boldsymbol{I}\right), \quad \frac{1}{\sqrt{n}}\sum_{t=[nu]+1}^{n}\boldsymbol{\Upsilon}_t \xrightarrow{\mathcal{L}} \mathcal{N}\left(\frac{1-u}{\sqrt{1-u_0}}\boldsymbol{J}\boldsymbol{\delta}_2, (1-u)\boldsymbol{I}\right).$$

We therefore have

$$\begin{aligned}
\boldsymbol{T}_n^0(u) &= \frac{1}{\sqrt{n}}\left(\sum_{t=1}^{[nu]}\boldsymbol{\Upsilon}_t - u\sum_{t=1}^{n}\boldsymbol{\Upsilon}_t\right) \\
&= \frac{1}{\sqrt{n}}\sum_{t=1}^{[nu]}(1-u)\boldsymbol{\Upsilon}_t - u\frac{1}{\sqrt{n}}\sum_{t=[nu]+1}^{[nu_0]}\boldsymbol{\Upsilon}_t - u\frac{1}{\sqrt{n}}\sum_{t=[nu_0]+1}^{n}\boldsymbol{\Upsilon}_t \\
&\xrightarrow{\mathcal{L}} \boldsymbol{T}_{u_0}(u) \sim \mathcal{N}\left\{\boldsymbol{J}\boldsymbol{\delta}_{u_0}(u), u(1-u)\boldsymbol{I}\right\}, \quad \boldsymbol{\delta}_{u_0}(u) = \frac{u(1-u_0)}{\sqrt{u_0}}\boldsymbol{\delta}_1 - u\sqrt{1-u_0}\boldsymbol{\delta}_2,
\end{aligned}$$

when $u \leq u_0$, and

$$\begin{aligned}
\boldsymbol{T}_n^0(u) &= \frac{1}{\sqrt{n}}\sum_{t=1}^{[nu_0]}(1-u)\boldsymbol{\Upsilon}_t + (1-u)\frac{1}{\sqrt{n}}\sum_{t=[nu_0]+1}^{[nu]}\boldsymbol{\Upsilon}_t - u\frac{1}{\sqrt{n}}\sum_{t=[nu]+1}^{n}\boldsymbol{\Upsilon}_t \\
&\xrightarrow{\mathcal{L}} \boldsymbol{T}_{u_0}(u) \sim \mathcal{N}\left\{\boldsymbol{J}\boldsymbol{\delta}_{u_0}(u), u(1-u)\boldsymbol{I}\right\}, \quad \boldsymbol{\delta}_{u_0}(u) = \sqrt{u_0}(1-u)\boldsymbol{\delta}_1 - \frac{u_0(1-u)}{\sqrt{1-u_0}}\boldsymbol{\delta}_2,
\end{aligned}$$

when $u \geq u_0$. It follows that, for all $u \in (0,1)$, $\boldsymbol{T}_{u_0}^\top(u)\boldsymbol{I}^{-1}\boldsymbol{T}_{u_0}(u)/u(1-u)$ follows a chi-square distribution with d degrees of freedom and noncentrality parameter

$$\frac{1}{u(1-u)}\boldsymbol{\delta}_{u_0}^\top(u)\boldsymbol{J}\boldsymbol{I}^{-1}\boldsymbol{J}\boldsymbol{\delta}_{u_0}(u),$$

which is maximal for the optimal QLE. We conclude by noting that the noncentral chi-squared distribution satisfies the stochastic-equal-mean order property: the larger the mean (*i.e.* the noncentrality parameter) is, the larger is the cdf, at any point. Note that the noncentrality parameter is maximal at u_0. □

B.3 Proof of Theorem 3

By standard arguments, it can be shown that **A2**, **A4** and **A7** entail

$$\sup_{\boldsymbol{\theta}\in\Theta}\left\|\sum_{t=1}^{n}\widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta})-\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})\right\| \leq \sum_{t=1}^{\infty}K_t\rho^t < \infty \quad \text{a.s.} \tag{25}$$

This entails that the initial values that are generally used to compute recursively $\widetilde{m}_t(\boldsymbol{\theta})$ and $\widetilde{\kappa}_{2t}(\boldsymbol{\theta})$ have no consequence on the asymptotic behavior of the QLEs. In particular (25) and the ergodic theorem entail that, for any neighborhood $V(\boldsymbol{\theta}_1)$ of $\boldsymbol{\theta}_1 \in \Theta$,

$$\begin{aligned}\lim_{n\to\infty}\inf_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_1)\cap\Theta}\left\|n^{-1}\sum_{t=1}^{n}\widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta})\right\| &\geq \lim_{n\to\infty}\inf_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_1)\cap\Theta}\left\|n^{-1}\sum_{t=1}^{n}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})\right\| \\ &\geq \|E\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_1)\| - E\sup_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_1)\cap\Theta}\|\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})-\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_1)\|.\end{aligned}$$

If $V_m(\boldsymbol{\theta})$ denotes the ball of center $\boldsymbol{\theta}$ and radius $1/m$, by the dominated convergence theorem

$$E\sup_{\boldsymbol{\theta}\in V_m(\boldsymbol{\theta}_1)\cap\Theta}\|\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})-\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_1)\|$$

is arbitrarily small when m is large enough. By **A3***, we also have $\|E\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_1)\| > 0$ when $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0^*$. We thus have shown that for any $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0^*$, there exists a neighborhood $V(\boldsymbol{\theta}_1)$ of $\boldsymbol{\theta}_1$ such that

$$\liminf_{n\to\infty}\inf_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_1)\cap\Theta}\left\|n^{-1}\sum_{t=1}^{n}\widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta})\right\| > 0, \quad \text{a.s.}$$

and that for any neighbourhood $V(\boldsymbol{\theta}_0^*)$ of $\boldsymbol{\theta}_0^*$

$$\limsup_{n\to\infty}\inf_{\boldsymbol{\theta}\in V(\boldsymbol{\theta}_0^*)\cap\Theta}\left\|n^{-1}\sum_{t=1}^{n}\widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta})\right\| = 0, \quad \text{a.s.}$$

By compactness of Θ, the existence and consistency of the QLE then follow.

A first order Taylor expansion, (25), and the consistency of $\widehat{\boldsymbol{\theta}}$ imply

$$\mathbf{0}_d = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \widetilde{\boldsymbol{\Upsilon}}_t(\widehat{\boldsymbol{\theta}}) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{\Upsilon}_t(\widehat{\boldsymbol{\theta}}) + o_P(1) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^*) + \boldsymbol{J}_n^* \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*) + o_P(1),$$

where the row i of $\boldsymbol{J}_n^*$ is of the form $n^{-1} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_{it}(\boldsymbol{\theta}^*)$, and $\boldsymbol{\theta}^*$ is such that $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0^*\| \leq \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*\|$. Using the consistency of $\widehat{\boldsymbol{\theta}}$, for n large enough we have

$$\|\boldsymbol{J}_n^* - \boldsymbol{J}^*\| \leq \frac{1}{n} \sum_{t=1}^{n} \sup_{\boldsymbol{\theta} \in V_m(\boldsymbol{\theta}_0^*) \cap \Theta} \left\| \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) - \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^*) \right\| + \left\| \boldsymbol{J}^* - \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^*) \right\|$$

for all m. The ergodic theorem entails that the a.s. limit as $n \to \infty$ of the right-hand side is arbitrarily small when m is large. The Bahadur representation, that is the expression of $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^*\right)$, follows. The last result follows from (20), which comes from **A1*** and the CLT of Herrndorf (1984). □

B.4 Proof of Theorem 4

By the functional CLT of Herrndorf (1984) and the Cramer-Wold device, **A1*** entails

$$\frac{(\boldsymbol{I}^*)^{1/2}}{\sqrt{n}} \sum_{t=1}^{[n\cdot]} \boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^*) \Longrightarrow \boldsymbol{W}(\cdot),$$

where $\boldsymbol{W}(\cdot)$ denotes a standard d-dimensional Brownian motion. The proof is therefore like that of Theorem 1. □

B.5 Proof of Theorem 5

Let

$$\boldsymbol{T}_n(u) = \begin{cases} \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu]} \boldsymbol{\Upsilon}_t^{(1)}(\widehat{\boldsymbol{\theta}}) & \text{if } u \leq u_0, \\ \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu_0]} \boldsymbol{\Upsilon}_t^{(1)}(\widehat{\boldsymbol{\theta}}) + \frac{1}{\sqrt{n}} \sum_{t=[nu_0]+1}^{[nu]} \boldsymbol{\Upsilon}_t^{(2)}(\widehat{\boldsymbol{\theta}}) & \text{if } u > u_0. \end{cases}$$

i) We start by showing that

$$\sup_{u \in (0,1)} \|\boldsymbol{T}_n(u) - \widetilde{\boldsymbol{T}}_n(u)\| = o_P(1). \tag{26}$$

We first consider the supremum over $u \in (u_0, 1)$, which is the most complicated term. We have

$$\begin{aligned}
\sup_{u\in(u_0,1)} \|\boldsymbol{T}_n(u) - \widetilde{\boldsymbol{T}}_n(u)\| &\leq \sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu_0]} \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\widetilde{\epsilon}_t(\boldsymbol{\theta})}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})} - \frac{\partial m_t^{(1)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\epsilon_t^{(1)}(\boldsymbol{\theta})}{\kappa_{2t}^{(1)}(\boldsymbol{\theta})} \right\| \\
&+ \sup_{u\in(u_0,1)} \sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{\sqrt{n}} \sum_{t=1+[nu_0]}^{[nu]} \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\widetilde{\epsilon}_t(\boldsymbol{\theta})}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})} - \frac{\partial m_t^{(2)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\epsilon_t^{(2)}(\boldsymbol{\theta})}{\kappa_{2t}^{(2)}(\boldsymbol{\theta})} \right\| \\
&\leq \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu_0]} \sup_{\boldsymbol{\theta}\in\Theta} \left(\left\| \frac{\partial m_t^{(1)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \frac{|\epsilon_t^{(1)}(\boldsymbol{\theta})|}{\kappa_{2t}^{(1)}(\boldsymbol{\theta})} \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu_0]} \sup_{\boldsymbol{\theta}\in\Theta} \left(|m_t^{(1)}(\boldsymbol{\theta}) - \widetilde{m}_t(\boldsymbol{\theta})| \left\| \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \frac{1}{\kappa_{2t}^{(1)}(\boldsymbol{\theta})} \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{t=1}^{[nu_0]} \sup_{\boldsymbol{\theta}\in\Theta} \left(\left| \frac{1}{\kappa_{2t}^{(1)}(\boldsymbol{\theta})} - \frac{1}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})} \right| \left\| \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| |\widetilde{\epsilon}_t(\boldsymbol{\theta})| \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{t=1+[nu_0]}^{n} \sup_{\boldsymbol{\theta}\in\Theta} \left(\left\| \frac{\partial m_t^{(2)}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \frac{|\epsilon_t^{(2)}(\boldsymbol{\theta})|}{\kappa_{2t}^{(2)}(\boldsymbol{\theta})} \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{t=1+[nu_0]}^{n} \sup_{\boldsymbol{\theta}\in\Theta} \left(|m_t^{(2)}(\boldsymbol{\theta}) - \widetilde{m}_t(\boldsymbol{\theta})| \left\| \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| \frac{1}{\kappa_{2t}^{(2)}(\boldsymbol{\theta})} \right) \\
&+ \frac{1}{\sqrt{n}} \sum_{t=1+[nu_0]}^{n} \sup_{\boldsymbol{\theta}\in\Theta} \left(\left| \frac{1}{\kappa_{2t}^{(2)}(\boldsymbol{\theta})} - \frac{1}{\widetilde{\kappa}_{2t}(\boldsymbol{\theta})} \right| \left\| \frac{\partial \widetilde{m}_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| |\widetilde{\epsilon}_t(\boldsymbol{\theta})| \right).
\end{aligned}$$

By Assumptions **B3-B4**, the first term in the right-hand side is bounded by

$$\frac{1}{\underline{\kappa}\sqrt{n}} \sum_{t=1}^{\infty} \sup_{\boldsymbol{\theta}\in\Theta} |\epsilon_t^{(1)}(\boldsymbol{\theta})| K_t^{(1)} \rho^t = O\left(\frac{1}{\sqrt{n}}\right), \quad a.s.$$

using the existence of a bound for a small-order moment for $\sup_{\boldsymbol{\theta}\in\Theta} |\epsilon_t^{(1)}(\boldsymbol{\theta})| K_t^{(1)}$. The other terms can be handled in the same way. We similarly show that $\sup_{u\in(0,u_0)} \|\boldsymbol{T}_n(u) - \widetilde{\boldsymbol{T}}_n(u)\| = o_P(1)$. Thus (26) is established.

ii) Now we prove that

$$\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0^\star \quad \text{a.s. as } n \to \infty. \tag{27}$$

We note that $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\| n^{-1} \sum_{t=1}^{n} \widetilde{\boldsymbol{\Upsilon}}_t(\boldsymbol{\theta}) \right\|$. Let $\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) = \boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta})$ if $t \leq [nu]$, and $\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}) = \boldsymbol{\Upsilon}_t^{(2)}(\boldsymbol{\theta})$ otherwise. For any neighborhood $V(\boldsymbol{\theta}_3)$ of $\boldsymbol{\theta}_3 \in \Theta$, using the fact that

$\sup_{\boldsymbol\theta\in\Theta}\left\|\sum_{t=1}^{\infty}\widetilde{\boldsymbol\Upsilon}_t(\boldsymbol\theta)-\boldsymbol\Upsilon_t(\boldsymbol\theta)\right\|<\infty$ a.s., we have

$$\begin{aligned}
&\lim_{n\to\infty}\inf_{\boldsymbol\theta\in V(\boldsymbol\theta_3)\cap\Theta}\left\|n^{-1}\sum_{t=1}^{n}\widetilde{\boldsymbol\Upsilon}_t(\boldsymbol\theta)\right\|\\
&\geq\lim_{n\to\infty}\left\|n^{-1}\sum_{t=1}^{n}\boldsymbol\Upsilon_t(\boldsymbol\theta_3)\right\|-\lim_{n\to\infty}\sup_{\boldsymbol\theta\in V(\boldsymbol\theta_3)\cap\Theta}\left\|n^{-1}\sum_{t=1}^{n}\boldsymbol\Upsilon_t(\boldsymbol\theta)\right\|\\
&\geq\left\|u_0E\left\{\boldsymbol\Upsilon_t^{(1)}(\boldsymbol\theta_3)\right\}+(1-u_0)E\left\{\boldsymbol\Upsilon_t^{(2)}(\boldsymbol\theta_3)\right\}\right\|-\sup_{i=1,2}E\sup_{\boldsymbol\theta\in V(\boldsymbol\theta_3)\cap\Theta}\left\|\boldsymbol\Upsilon_t^{(i)}(\boldsymbol\theta)-\boldsymbol\Upsilon_t^{(i)}(\boldsymbol\theta_3)\right\|,
\end{aligned}$$

where the first term in the r.h.s. is positive for $\boldsymbol\theta_3\neq\boldsymbol\theta_0^*$, while the second term can be made arbitrarily small when the neighborhood shrinks to the singleton $\{\boldsymbol\theta_3\}$ by arguments already given. The consistency of $\widehat{\boldsymbol\theta}$ follows as in the proof of Theorem 3.

iii) Under **B5** let

$$\begin{aligned}
\boldsymbol I^\star&=u_0E\left\{\boldsymbol\Upsilon_t^{(1)}(\boldsymbol\theta_0^\star)\left(\boldsymbol\Upsilon_t^{(1)}(\boldsymbol\theta_0^\star)\right)^\top\right\}+(1-u_0)E\left\{\boldsymbol\Upsilon_t^{(2)}(\boldsymbol\theta_0^\star)\left(\boldsymbol\Upsilon_t^{(2)}(\boldsymbol\theta_0^\star)\right)^\top\right\},\\
\boldsymbol J^\star&=u_0E\left\{\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t^{(1)}(\boldsymbol\theta_0^\star)\right\}+(1-u_0)E\left\{\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t^{(2)}(\boldsymbol\theta_0^\star)\right\}.
\end{aligned}$$

We will show that

$$\sup_{u\in(0,1)}\left|\frac{1}{n}\widetilde{S}_n(u)-L(u)\right|=o_P(1)\tag{28}$$

where

$$L(u)=\begin{cases}\{u(1-u_0)\}^2\boldsymbol\Delta'(\boldsymbol\theta_1,\boldsymbol\theta_2)\boldsymbol I^{\star-1}\boldsymbol\Delta(\boldsymbol\theta_1,\boldsymbol\theta_2) & \text{if } u\leq u_0\\ \{u_0(1-u)\}^2\boldsymbol\Delta'(\boldsymbol\theta_1,\boldsymbol\theta_2)\boldsymbol I^{\star-1}\boldsymbol\Delta(\boldsymbol\theta_1,\boldsymbol\theta_2) & \text{if } u>u_0\end{cases}$$

and, recalling that $\boldsymbol\theta_0^\star$ depends on $\boldsymbol\theta_1$ and $\boldsymbol\theta_2$,

$$\boldsymbol\Delta(\boldsymbol\theta_1,\boldsymbol\theta_2)=E\left\{\boldsymbol\Upsilon_t^{(1)}(\boldsymbol\theta_0^\star)\right\}-E\left\{\boldsymbol\Upsilon_t^{(2)}(\boldsymbol\theta_0^\star)\right\}.$$

By the arguments of the proof of Theorem 1 we have, with $\widehat{\boldsymbol\theta}_u^*$ between $\widehat{\boldsymbol\theta}$ and $\boldsymbol\theta_0^\star$,

$$\frac{1}{\sqrt n}\boldsymbol T_n(u)=\frac{1}{n}\sum_{t=1}^{[nu]}\boldsymbol\Upsilon_t(\boldsymbol\theta_0^\star)+u\left(\frac{1}{nu}\sum_{t=1}^{[nu]}\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t(\widehat{\boldsymbol\theta}_u^*)\right)(\widehat{\boldsymbol\theta}-\boldsymbol\theta_0^\star)$$

Given that $\widetilde{\boldsymbol T}_n(1)=0$ in view of (26) we have, for $\widehat{\boldsymbol\theta}^*$ between $\widehat{\boldsymbol\theta}$ and $\boldsymbol\theta_0^\star$,

$$\boldsymbol T_n(1)=o_P(1)=\frac{1}{\sqrt n}\sum_{t=1}^{n}\boldsymbol\Upsilon_t(\boldsymbol\theta_0^\star)+\boldsymbol J_n^\star\sqrt n(\widehat{\boldsymbol\theta}-\boldsymbol\theta_0^\star),\quad \boldsymbol J_n^\star=\frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t(\widehat{\boldsymbol\theta}^*).$$

Moreover, we have

$$\begin{aligned}
\boldsymbol J_n^\star=&\frac{1}{nu}\sum_{t=1}^{[nu]}\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t(\widehat{\boldsymbol\theta}_u^\star)+\frac{1}{nu}\sum_{t=1}^{[nu]}\left\{\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t(\widehat{\boldsymbol\theta}^\star)-\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t(\widehat{\boldsymbol\theta}_u^\star)\right\}\\
&-\frac{1}{nu}\sum_{t=1}^{[nu]}\left(\frac{\partial}{\partial\boldsymbol\theta^\top}\boldsymbol\Upsilon_t(\widehat{\boldsymbol\theta}^\star)-\boldsymbol J_n^\star\right)+\left(1-\frac{[nu]}{nu}\right)\boldsymbol J_n^\star.
\end{aligned}$$

Thus, by already given arguments

$$\begin{aligned}\frac{1}{\sqrt{n}}\boldsymbol{T}_n(u) =& \frac{1}{n}\left(\sum_{t=1}^{[nu]}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^\star) - u\sum_{t=1}^{n}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^\star)\right) + \frac{1}{n}\sum_{t=1}^{[nu]}\left(\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\widehat{\boldsymbol{\theta}}_u^*) - \boldsymbol{J}_n^\star\right)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^\star) + uo_P(1) \\ :=& \frac{1}{\sqrt{n}}\boldsymbol{T}_n^0(u) + \boldsymbol{R}_n(u)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0^\star) + o_P(1), \end{aligned} \tag{29}$$

where the reminder term is independent of u. We will show that, in contrast with the first convergence in (13) of the proof of Theorem 1, we have

$$\sup_{u\in(0,1)}\|\boldsymbol{R}_n(u)\| = O_P(1). \tag{30}$$

We have, using the consistency of $\widehat{\boldsymbol{\theta}}_u^*$ to $\boldsymbol{\theta}_0^\star$,

$$\begin{aligned}\|\boldsymbol{R}_n(u)\| \leq& \left\|\frac{1}{n}\sum_{t=1}^{[nu]}\left(\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^\star) - \boldsymbol{J}^\star\right)\right\| + \sup_{\boldsymbol{\theta}\in V_k(\boldsymbol{\theta}_0^\star)}\left\|\frac{1}{n}\sum_{t=1}^{[nu]}\left(\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^\star) - \frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta})\right)\right\| \\ &+ \|\boldsymbol{J}_n^\star - \boldsymbol{J}^\star\|.\end{aligned}$$

As in the proof of Theorem 1, it can be shown that the second term in the r.h.s. converges to 0 in probability, uniformly in $u \in (0,1)$. It can also be shown that the third term converges to 0 in probability. Moreover,

$$\begin{aligned}\sup_{u\in(0,u_0)}\left\|\frac{1}{n}\sum_{t=1}^{[nu]}\left(\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^\star) - \boldsymbol{J}^\star\right)\right\| \leq& \sup_{u\in(0,u_0)}\left\|\frac{1}{n}\sum_{t=1}^{[nu]}\left(\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t(\boldsymbol{\theta}_0^\star) - E\left\{\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta}_0^\star)\right\}\right)\right\| \\ &+ \left\|E\left\{\frac{\partial}{\partial\boldsymbol{\theta}^\top}\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta}_0^\star)\right\} - \boldsymbol{J}^\star\right\|\end{aligned}$$

where the first term in the r.h.s. converges to 0 in probability by already given arguments. It can also be shown that a similar bound holds when the supremum of the l.h.s. term is taken over $(u_0, 1)$. Thus (30) is established, from which it follows that the second term in the r.h.s. of (29) converges to 0 in probability as $n \to \infty$.

Now we have, for $u \leq u_0$,

$$\begin{aligned}\frac{1}{\sqrt{n}}\boldsymbol{T}_n^0(u) =& u(1-u)\frac{1}{[nu]}\sum_{t=1}^{[nu]}\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta}_0^\star) - u(u_0-u)\frac{1}{[nu_0]-[nu]}\sum_{t=[nu]+1}^{[nu_0]}\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta}_0^\star) \\ &- u(1-u_0)\frac{1}{n-[nu_0]}\sum_{t=[nu_0]+1}^{n}\boldsymbol{\Upsilon}_t^{(2)}(\boldsymbol{\theta}_0^\star) + o_P(1) \\ &\to u(1-u)E\left\{\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta}_0^\star)\right\} - u(u_0-u)E\left\{\boldsymbol{\Upsilon}_t^{(1)}(\boldsymbol{\theta}_0^\star)\right\} \\ &- u(1-u_0)E\left\{\boldsymbol{\Upsilon}_t^{(2)}(\boldsymbol{\theta}_0^\star)\right\}, \quad \text{in probability as } n\to\infty.\end{aligned}$$

Thus, for $u \leq u_0$,

$$\frac{1}{\sqrt{n}}\boldsymbol{T}_n^0(u) \to u(1-u_0)\boldsymbol{\Delta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2), \quad \text{in probability as } n\to\infty,$$

and we similarly show that, for $u > u_0$,

$$\frac{1}{\sqrt{n}}\boldsymbol{T}_n^0(u) \to u_0(1-u)\boldsymbol{\Delta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2), \quad \text{in probability as } n\to\infty.$$

Thus, (28) is not yet established but we have shown that $\frac{1}{\sqrt{n}}\boldsymbol{T}_n^0(u) \to \boldsymbol{T}(u)$ for all $u \in (0,1)$, where

$$\boldsymbol{T}(u) = \begin{cases} u(1-u_0)\boldsymbol{\Delta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) & \text{if } u \leq u_0, \\ u_0(1-u)\boldsymbol{\Delta}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) & \text{if } u > u_0. \end{cases}$$

Now we have, letting $\boldsymbol{Y}_t^{(i)} = \boldsymbol{\Upsilon}_t^{(i)}(\boldsymbol{\theta}_0^\star) - E\boldsymbol{\Upsilon}_t^{(i)}(\boldsymbol{\theta}_0^\star)$ for $i=1,2$,

$$\begin{aligned} \sup_{u\in(0,u_0)}\left\|\frac{1}{\sqrt{n}}\boldsymbol{T}_n^0(u) - \boldsymbol{T}(u)\right\| \leq \sup_{u\in(0,u_0)} & \left\|\frac{u}{[nu]}\sum_{t=1}^{[nu]}\boldsymbol{Y}_t^{(1)}\right\| + \left\|\frac{u_0-u}{[nu_0]-[nu]}\sum_{t=[nu]+1}^{[nu_0]}\boldsymbol{Y}_t^{(1)}\right\| \\ & + (1-u_0)\left\|\frac{1}{n-[nu_0]}\sum_{t=[nu_0]+1}^{n}\boldsymbol{Y}_t^{(2)}\right\| + o_p(1). \end{aligned} \tag{31}$$

The third term in the r.h.s. converges to 0 in probability as $n\to\infty$ in view of the stationarity and ergodicity of $\boldsymbol{Y}_t^{(2)}$. Moreover, by the arguments used in the proof of Theorem 1,

$$\sup_{u\in(0,u_0)}\left\|\frac{u}{[nu]}\sum_{t=1}^{[nu]}\boldsymbol{Y}_t^{(1)}\right\| \leq \sup_{1\leq k\leq k_n}\left\|\frac{2}{n}\sum_{t=1}^{k}\boldsymbol{Y}_t^{(1)}\right\| + \sup_{k_n\leq k\leq n}\left\|\frac{2}{n}\sum_{t=1}^{k}\boldsymbol{Y}_t^{(1)}\right\|$$

where the last term in the r.h.s. converges to 0 a.s. and, by the Markov inequality, for any $\iota > 0$, and $k_n/n \to 0$,

$$P\left(\sup_{1\leq k\leq u_n}\left\|\frac{1}{n}\sum_{t=1}^{k}\boldsymbol{Y}_t^{(1)}\right\| > \iota\right) \leq P\left(\frac{1}{n}\sum_{t=1}^{k_n}\left\|\boldsymbol{Y}_t^{(1)}\right\| > \iota\right) \leq \frac{k_n}{n\iota}E\left\|\boldsymbol{Y}_1^{(1)}\right\| \to 0,$$

as $n\to\infty$, from which we deduce that the first term in the r.h.s. of (31) converges in probability to 0. The second termcan be handled similarly. It follows that the l.h.s. of (31) converges in probability to 0. We similarly show that the same convergence holds when the supremum is taken over $(u_0,1)$. We thus have shown that

$$\sup_{u\in(0,1)}\left\|\frac{1}{\sqrt{n}}\boldsymbol{T}_n^0(u) - \boldsymbol{T}(u)\right\| \to 0, \quad \text{in probability as } n\to\infty.$$

In view of equations (26), (29) and (30), we also have

$$\sup_{u\in(0,1)}\left\|\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{T}}_n(u) - \boldsymbol{T}(u)\right\| \to 0, \quad \text{in probability as } n\to\infty. \tag{32}$$

Noting that the matrix $\boldsymbol{I}_n$ converges in probability to $\boldsymbol{I}^\star$, the convergence in (28) is established. To conclude, it suffices to apply the argmax theorem (see Theorem 3.2.2 of Van der Vaart and Wellner, 1996). □