MOVING SUM DATA SEGMENTATION FOR STOCHASTIC PROCESSES BASED ON INVARIANCE

Otto-von-Guericke-University Magdeburg

Supplementary Material

S1 Simulation study

In this section, we illustrate the performance of our procedure for multivariate renewal processes by means of a simulation study. Related simulations in addition to a variety of data examples for partial sum processes have been conducted by Eichinger and Kirch (2018); Meier et al. (2019) and for univariate renewal processes by Messer et al. (2014, 2017).

More precisely, we analyze three-dimensional renewal processes with $T = 1600$, where the increments of the inter-event times for each component are $\Gamma-$distributed with intensity changes at 250, 500, 900 and 1150, where the expected time μ between events is given by 1.3, 0.9, 0.6, 0.8 and 1.3. We use a bandwidth of $h = 120$ and the parameter $\eta = 0.75$. Smaller values of η as suggested by Meier et al. (2019) for partial sum processes tend to produce duplicate change point estimators by having

two or more significant local maxima for each change point if the variance is too large (as can be seen in Table 2 below), while larger values of η lead to slightly worse detection rates. For a single-bandwidth MOSUM procedure as suggested here, this should be avoided but can be relaxed if a post-processing procedure is applied as e.g. by Cho and Kirch (2021b) for partial sum processes.

In contrast to partial sum processes, it is natural for renewal processes that the variances change with the intensity. Therefore we consider the following three scenarios: (i) standard deviations of constant value 0.7 (referred to as `constvar`), (ii) standard deviations being $5/6\mu$ (referred to as `smallvar`) and (iii) multivariate Poisson processes (referred to as `Poisson`).

We consider both the case of independence and dependence between the three components. In the latter case, we generate for each regime i an independent (in time) sequence of $\Gamma-$distributed inter-event-times $Y_j = Y_j^{(i)}$, $j = 1, 2, 3$, with a correlation of 0.2 (for all pairs) as $Y_j = X_j + X_4$, where $X_j \sim \Gamma(s, \lambda)$ for $j = 1, 2, 3$ and $X_4 \sim \Gamma(s/4, \lambda)$ for appropriate values of s and λ (resulting in the above intensities and standard deviations for each regime).

In the simulations, we use a threshold as in Remark 1 with $\alpha_T = 0.05$. By Section 2.2 and (3.8) it holds that $\mathbf{\Sigma}_t = \mathrm{Cov}\left((Y_1, Y_2, Y_3)'\right) / E\left(Y_1\right)^3$ while we use the following choices for the matrix $\widehat{\mathbf{A}}_t$ as in (3.7): (A) Diagonal matrix with locally estimated variances $\widehat{\mathbf{\Sigma}}_t(j, j)$ on the diagonal, $j = 1, 2, 3$, (B) with the true variances

(a) `constvar`: Constant standard deviation of 0.7, $\eta = 0.75$.

Change point at	250	500	900	1150	spurious	duplicate
independent, type (A)	1	0.9998	0.9434	1	0.0251	0.0024
independent, type (B)	0.9974	0.9789	0.6271	1	0.0035	0.0007
dependent, type (A)	0.9998	0.9991	0.9219	1	0.0344	0.0027
dependent, type (B)	0.9916	0.9610	0.6351	0.9997	0.0074	0.0008
dependent, type (C)	0.9522	0.8485	0.3670	0.9984	0.0055	0.0019

(b) `smallvar`: Standard deviation of 5/6 the expected time between events, $\eta = 0.75$.

Change point at	250	500	900	1150	spurious	duplicate
independent, type (A)	0.9831	1	0.9735	1	0.0313	0.0033
independent, type (B)	0.9368	1	0.9309	0.9999	0.0038	0.0004
dependent, type (A)	0.9711	0.9999	0.9556	0.9998	0.0386	0.0055
dependent, type (B)	0.9207	0.9986	0.9094	0.9987	0.0073	0.0018
dependent, type (C)	0.7494	0.9890	0.7210	0.9908	0.0052	0.0017

(c) `Poisson`-distributed inter-event times, $\eta = 0.75$.

Change point at	250	500	900	1150	spurious	duplicate
independent, type (A)	0.9054	0.9971	0.8710	0.9983	0.0445	0.0077
independent, type (B)	0.7366	0.9852	0.7188	0.9885	0.0028	0.0014
dependent, type (A)	0.8818	0.9924	0.8418	0.9939	0.0528	0.0091
dependent, type (B)	0.7166	0.9764	0.6978	0.9761	0.0054	0.0020
dependent, type (C)	0.4602	0.8934	0.4289	0.9007	0.0048	0.0013

Table 1: Detection rates for each change point as well as the average number of spurious and duplicate estimators for different distributions of the inter-event times.

	duplicate, `constvar` $\eta = 0.4$	duplicate, `constvar` $\eta = 0.75$	duplicate, `smallvar` $\eta = 0.4$	duplicate, `smallvar` $\eta = 0.75$	duplicate, `poisson` $\eta = 0.4$	duplicate, `poisson` $\eta = 0.75$
independent, type (A)	0.0798	0.0024	0.1046	0.0033	0.1559	0.0077
independent, type (B)	0.0484	0.0007	0.0496	0.0004	0.0526	0.0014
dependent, type (A)	0.1057	0.0027	0.1505	0.0055	0.1880	0.0091
dependent, type (B)	0.0779	0.0008	0.0814	0.0018	0.0755	0.0020
dependent, type (C)	0.0512	0.0019	0.0494	0.0017	0.0349	0.0013

Table 2: Comparison of the average number of duplicate estimators for $\eta = 0.4$ and $\eta = 0.75$.

$\mathbf{\Sigma}_t(j,j)$ on the diagonal and (C) in case of dependent components (non-diagonal) true covariance matrix $\mathbf{\Sigma}_t$. While only (A) is of relevance in applications, this allows us to understand the influence of estimating the variance on the procedure. For dependent data, the distinction between (B) and (C) is important for applications, because a good enough estimator (resulting in a reasonable estimator for the inverse) is often not available for the full covariance matrix as in (C) for moderately high or high dimensions, while it is much less problematic to estimate (B). In (A) the variances at location t are estimated as

$$\widehat{\mathbf{\Sigma}}_t(j,j) = \min\left\{\frac{\widehat{\sigma}^2_{j,-}(t)}{\widehat{\mu}^3_{j,-}(t)}, \frac{\widehat{\sigma}^2_{j,+}(t)}{\widehat{\mu}^3_{j,+}(t)}\right\}, \tag{S1.1}$$

where $\widehat{\sigma}^2_{j,\pm}(t)$ and $\widehat{\mu}_{j,\pm}(t)$ are the sample variance and sample mean respectively based on the inter-event times of the j-th component within the windows $(t-h, t]$ for $'-'$ respectively $(t, t+h]$ for $'+'$. The first and last inter-event times that have been censored by the window are not included. Using the minimum of the left and right local estimators takes into account that the variance can (and typically will) change with the intensity which has already been discussed by Meier et al. (2019) in the context of partial sum processes.

The results of the simulation study can be found in Table 1, where we consider a change point to be detected if there was an estimator in the interval $[c_i - h, c_i + h]$. Additional significant local maxima in such an interval are called *duplicate* change point estimators, while additional significant local maxima outside any of these intervals are called *spurious*.

The procedure performs well throughout all simulations with high detection rate, few spurious and very few duplicate estimators. The results improve further for smaller variance, in which case the signal-to-noise ratio is better.

When the diagonal matrix with the estimated variance is being used, the detection power is larger in all cases than when the true variance is being used. In case of the changes at location 900 this is a substantial improvement, such that the use of this local variance estimator can help boost the signal significantly. This comes at the cost of having an increased but still reasonable amount of spurious and duplicate

change point estimators.

This effect stems from using the minimum in (S1.1), which was introduced to gain detection power if the variance changes with the intensity. Additionally, the use of the true (asymptotic) covariance matrix leads to worse results than only using the corresponding diagonal matrix, which is due to the fact that the theoretical signal term is smaller when using the true (asymptotic) covariance matrix in this example. From a statistical perspective this is advantageous because the local estimation of the inverse of a covariance matrix in moderately large or large dimensions is a very hard problem leading to a loss in precision, while the diagonal elements are far less difficult to estimate consistently.

However, in other examples using the full covariance matrix can also lead to better behavior, namely if the theoretical signal term is bigger in that case. The results for one such example can be found in Table 3. Here, the inter-event times are $Y_j = X_j + \sum_{1\leq k<j} X_{k,j} - \sum_{j<k\leq 3} X_{j,k}$ $j = 1, 2, 3$ where the $X_j = X_j^{(i)}$ are sequences of independent in time $\Gamma(s, \lambda)$-distributed random variables. The $X_{j,k} = X_{j,k}^{(i)}$ are sequences of independent in time $\mathcal{N}(0, s_1^2)$-distributed random variables with s, λ, s_1 appropriately chosen such that the distributions of the inter-event-times have the above average intensities, standard deviations and correlations.

Furthermore, we illustrate the performance of our procedure in the case that Assumption 2 that the bandwidth is less than half the distance to the next change

(a) `constvar`: Constant standard deviation of 0.7, $\eta = 0.75$, $h = 120$.

Change point at	250	500	900	1150	spurious	duplicate
dependent, type (A)	1	1	0.9764	1	0.0666	0.0068
dependent, type (B)	0.9997	0.9936	0.6515	1	0.0037	0.0007
dependent, type (C)	1	0.9999	0.9539	1	0.0049	0.0004

(b) `smallvar`: Standard deviation of 5/6 the expected time between events, $\eta = 0.75$.

Change point at	250	500	900	1150	spurious	duplicate
dependent, type (A)	0.9955	1	0.9905	1	0.0772	0.0077
dependent, type (B)	0.9741	1	0.9628	1	0.0042	0.0006
dependent, type (C)	0.9995	1	0.9989	1	0.0045	0.0003

(c) `Poisson`-distributed inter-event times, $\eta = 0.75$.

Change point at	250	500	900	1150	spurious	duplicate
dependent, type (A)	0.9535	0.9994	0.9192	0.9998	0.1187	0.0203
dependent, type (B)	0.7781	0.9964	0.7433	0.9973	0.0054	0.0012
dependent, type (C)	0.9824	1	0.9776	1	0.0059	0.0008

Table 3: Detection rates for each change point as well as the average number of spurious and duplicate estimators for different distributions of the inter-event times.

point is violated: We analyze three-dimensional renewal processes with $T = 1600$, where the increments of the inter-event times for each component are $\Gamma-$distributed with intensity changes at 250, 500 and 600, where the expected time μ between events is given by 1.3, 0.9, 0.6 and 0.8. We use bandwidths of $h = 60, 90, 120$ and the parameter $\eta = 0.75$. While for the change point at 250 all bandwidths fulfill the assumption, this is true for neither of the other two change points with the bandwidth $h = 120$ being larger than the distance between these two points.

We use the same three scenarios for the standard deviations of the inter-event-times as above. We assume independence between the components and for the matrix $\widehat{\mathbf{A}}_t$, we consider only choice (A) – a matrix with locally estimated variances $\widehat{\mathbf{\Sigma}}_t(j,j)$ on the diagonal, $j = 1, 2, 3$. The results of the simulation study can be found in Table 4, where we consider a change point to be detected if there was an estimator in the interval $[c_i - \min\{h, (c_i - c_{i-1})/2\}, c_i + \min\{h, (c_{i+1} - c_i)/2\}]$.

Clearly, the procedure is performing well even when the model assumptions are mildly violated, as for $h = 60$ and $h = 90$ and the last two change points. For $h = 120$, the detection rates for the change point at 500 slightly increases but the average distance of the estimator to the true change point becomes much larger. For the change at 600, additionally the detection rate clearly decreases. On the other hand, as long as Assumption 2 holds (as for the first change point) or is only mildly violated (as for the last two change points and the two smaller bandwidths), the

detection rate increases with larger bandwidth while at the time the average distances between the estimator and the corresponding true change point decreases. This is due to an increased signal-to-noise ratio due to the larger bandwidths (corresponding to a larger sample size in classical two-sample testing).

In the above situation the changes are *homogeneous* in the sense that the smallest change in intensity is still large enough compared to the smallest distance to neighboring change points (for a detailed definition we refer to Cho and Kirch (2021b), Definition 2.1, or Cho and Kirch (2021a), Definition 2.1). In particular, this guarantees that all changes can be detected with a single bandwidth only.

In some applications with *multiscale* signals, where frequent large changes as well as small isolated changes are present, this is no longer the case as Figure 1 shows. In such cases, several bandwidths need to be used and the obtained candidates are pruned down in a second step (see Cho and Kirch (2021b) for an information criterion based approach for partial sum processes as well as Messer et al. (2014) for a bottom-up-approach for renewal processes). Similarly, if the distance to the neighboring change points is unbalanced MOSUM procedures with asymmetric bandwidths as suggested by Meier et al. (2019) may be necessary.

(a) `constvar`: Constant standard deviation of 0.7, $\eta = 0.75$.

Change point at	250	500	600	spurious	duplicate	Dist. 500	Dist. 600
h=60	0.9847	0.9690	0.6528	0.2436	0.0073	5.52	8.47
h=90	0.9996	0.9970	0.7957	0.1044	0.0025	4.92	8.16
h=120	1	0.9984	0.6368	0.0561	0.0002	9.94	21.43

(b) `smallvar`: Standard deviation of 5/6 the expected time between events, $\eta = 0.75$.

Change point at	250	500	600	spurious	duplicate	Dist. 500	Dist. 600
h=60	0.7534	0.9592	0.6476	0.2689	0.0100	5.28	7.55
h=90	0.9273	0.9978	0.8461	0.1025	0.0054	4.89	7.28
h=120	0.9846	0.9987	0.7237	0.0546	0.0020	10.01	19.66

(c) `Poisson`-distributed inter-event times, $\eta = 0.75$.

Change point at	250	500	600	spurious	duplicate	Dist. 500	Dist. 600
h=60	0.5904	0.8494	0.4698	0.3807	0.0129	7.05	9.36
h=90	0.7838	0.9702	0.6696	0.1457	0.0077	6.65	9.29
h=120	0.9070	0.9807	0.5798	0.0724	0.0046	11.43	20.76

Table 4: Detection rates for each change point, average number of spurious and duplicate estimators for different distributions of the inter-event times as well as the average distances of the change point estimators closest to the true change points in the intervals $[c_i - \min\{h, (c_i - c_{i-1})/2\}, c_i + \min\{h, (c_{i+1} - c_i)/2\}]$ for $c_i = 500, 600$, respectively.

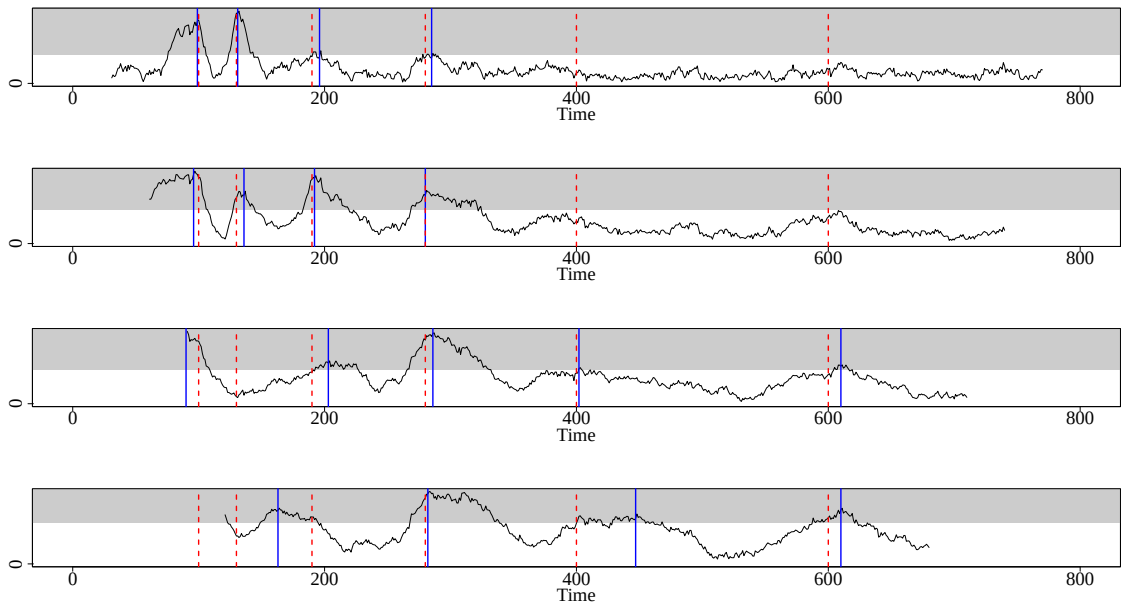


Figure 1: MOSUM statistics with bandwidths of $h = 30, 60, 90, 120$ (top to bottom) for a three-dimensional renewal process with *multiscale* changes with increasing distance between change points in combination with decreasing magnitude of the changes in intensity. The dashed vertical lines indicate the location of the true changes, while the solid lines indicate the change point estimators. In this multiscale situation no single bandwidth can detect all changes: The changes to the left are well estimated by smaller bandwidth, the ones in the middle by medium-sized bandwidths and the one to the right by the largest bandwidth.

S1.1 Further Examples

In this section, we give two more examples fulfilling the model assumptions of Section 3.1 namely partial sum-processes as well as integrals of diffusion processes including Ornstein-Uhlenbeck and Wiener processes with drift.

Partial-Sum-Processes

This first example extends the classical multiple changes in the mean model:

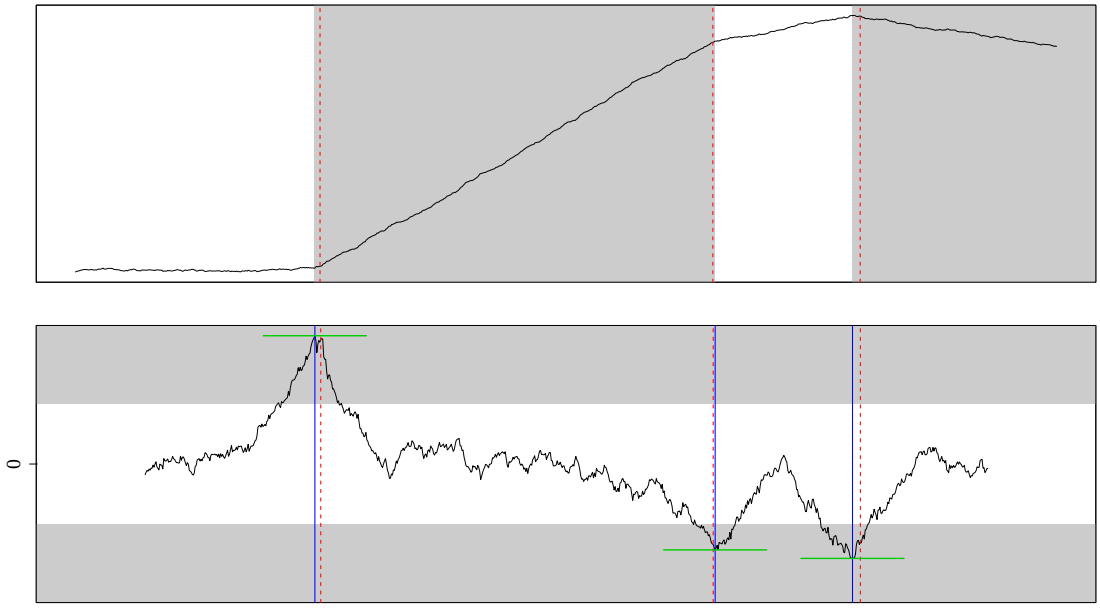
Let $\mathbf{X}_1^{(i)}, \mathbf{X}_2^{(i)}, \ldots$ be a time series with $E\left(\mathbf{X}_1^{(i)}\right) = 0$ and $\mathrm{Cov}\left(\mathbf{X}_1^{(i)}\right) = I_p$ and all $i = 1, \ldots, P$. Let

$$\mathbf{R}_t^{(i)} = \sum_{j=1}^{\lfloor t \rfloor} \left(\boldsymbol{\mu}^{(i)} + \left(\boldsymbol{\Sigma}_T^{(i)}\right)^{1/2} \mathbf{X}_j^{(i)} \right).$$

The upper panel in Figure 2 shows one such realization for illustrational purposes where the noise is a sequence of i.i.d. standard normally distributed random variables. The corresponding process fulfills Assumption 1 in a wide range of situations. For example, Einmahl (1987) shows the validity in the case that $\mathbf{X}_1, \mathbf{X}_2, \ldots$ with $\mathbf{X}_j = \left(\mathbf{X}_j^{(1)}, \ldots, \mathbf{X}_j^{(P)}\right)'$ are i.i.d. with $E\left(\|\mathbf{X}_1\|^{2+\delta}\right) < \infty$ for some $\delta > 0$. Additionally, Kuelbs and Philipp (1980) state an invariance principle for mixing random vectors in Theorem 4. For univariate processes there are many corresponding results under different weak-dependency formulations.

For $\mathbf{X}^{(i)} = \mathbf{X}^{(1)}$ (and $\boldsymbol{\Sigma}^{(i)} = \boldsymbol{\Sigma}^{(1)}$) for all i, then we are back to the classical multiple mean change problem that has been considered in many papers in particular

Figure 2: In the upper panel, a univariate partial sum process with 3 change points (i.e. 4 stationary segments) and standard normally distributed noise are displayed. The gray and white regions mark the estimated segmentation of the data while the red intervals mark the true segmentation.
In the lower panel, the corresponding MOSUM statistic with (relative) bandwidth $h/T = 0.07$ is displayed. The gray areas are the regions where the threshold ($\alpha = 0.05$ as in Remark 1) is exceeded (in absolute value). The blue solid lines indicate the change point estimates obtained as local extrema that fall within the gray area (making them *significant*). The true change points are indicated by the red dashed lines. The green horizontal lines denote ηh-environments around the estimators.



for the univariate situation, see e.g. the recent survey papers by Fearnhead and Rigaill (2020) or Cho and Kirch (2021a). As a proof of concept, the lower panel in Figure 2 shows the corresponding MOSUM statistic with the true variances for $\mathbf{\Sigma}_t$. Similarly to Figure 2 in the main document for renewal processes the statistic fluctuates around 0 away from the change points while local maxima close to the changes are significant.

Diffusion processes

Clearly, switching between independent (or components of a multivariate) Brownian motion with drift is included in this framework. Additionally, Heunis (2003) and Mihalache (2011) derive invariance principles in the context of diffusion processes including Ornstein-Uhlenbeck processes among others. Let $(\mathbf{X}_t)_{t\geq 0}$ be a stochastic process in $\mathbb{R}^N$ satisfying a stochastic differential equation

$$d\mathbf{X}_t = \boldsymbol{\mu}\left(\mathbf{X}_t\right) dt + \mathbf{\Sigma}\left(\mathbf{X}_t\right) d\mathbf{B}_t$$

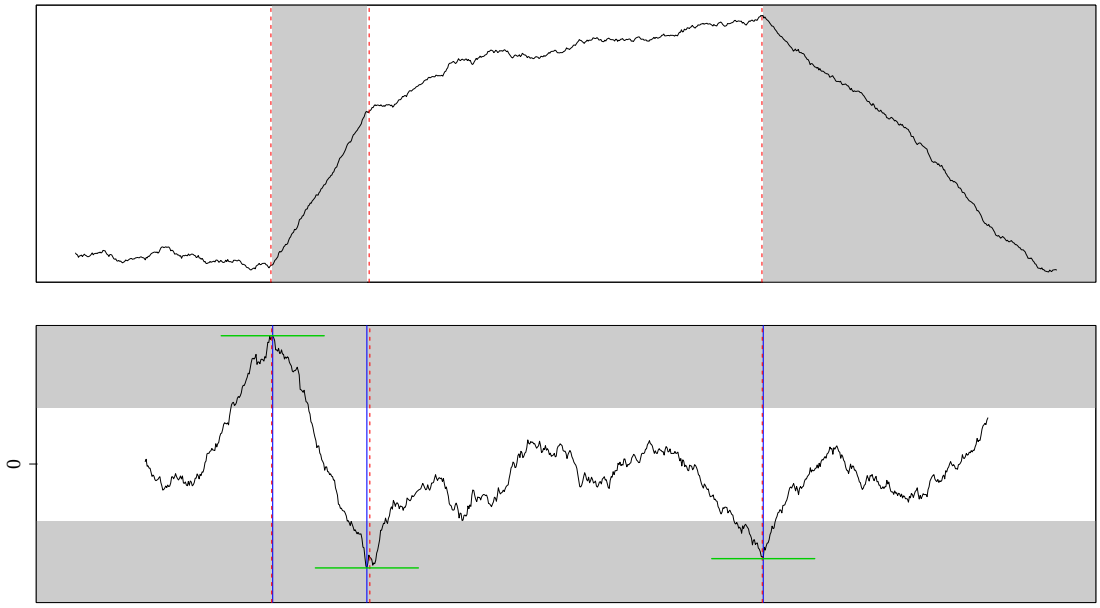
with respect to an n-dimensional standard Wiener process $(\mathbf{B}_t)_{t\geq 0}$ and let $\boldsymbol{\mu}, \mathbf{\Sigma}$ be globally Lipschitz-continuous.

Under some conditions on $f : \mathbb{R}^N \to \mathbb{R}^p$, as given by Heunis (2003), relating to $\boldsymbol{\mu}, \mathbf{\Sigma}$, which in particular guarantee that the function f applied to the (invariant) diffusion results in a centered process, there exists a p-dimensional Wiener process $(\mathbf{W}_t)_{t\geq 0}$ and some $\eta > 0$ such that

$$\left\| \int_0^T f(\mathbf{X}_s)\, ds - \mathbf{W}_T \right\| = O\left(T^{1/2-\eta}\right),$$

Figure 3: In the upper panel, a univariate Wiener process with drift with 3 change points (i.e. 4 stationary segments) is displayed. The gray and white regions mark the estimated segmentation of the data while the red intervals mark the true segmentation.

In the lower panel, the corresponding MOSUM statistic with (relative) bandwidth $h/T = 0.07$ is displayed. The gray areas are the regions where the threshold ($\alpha = 0.05$ as in Remark 1) is exceeded (in absolute value). The blue solid lines indicate the change point estimates obtained as local extrema that fall within the gray area (making them *significant*). The true change points are indicated by the red dashed lines. The green horizontal lines denote ηh-environments around the estimators.



where $(\mathbf{X}_t)_{t\geq 0}$ either is a solution to the SDE with fixed starting value $\mathbf{X}_0 = y_0$ or a strictly stationary solution with respect to an invariant distribution.

Furthermore, in the case of a one-dimensional stochastic diffusion process, Mihalache (2011) showed for some L^2-functions fulfilling constraints depending on $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ that there exists a strong invariance principle for the integrals of diffusion processes with a rate of $O((T \log_2 T)^{1/4} \sqrt{\log T})$. Figure 3 illustrates our method for a univariate Wiener process with changing drift, where we use the true variance for $\boldsymbol{\Sigma}_t$. In general, having suitable estimators for the covariance structure of a diffusion process is a non-trivial problem – even more so in the presence of change points. Again the behavior is very similar to Figures 2 in the main document and 2.

S2 Proofs

We first prove some bounds for the limiting Wiener process that will be used throughout the proofs (for (i)) or are related to the bounds in Assumption 5 (for (ii) and (iii)).

Proposition S2.1. *Let Assumption 1 hold with a rate of convergence as in Assumption 2 with the notation of Assumption 5. Let* $0 < \xi_T \leq h_T$ *and* $D_T \geq 1$ *be arbitrary sequences (bounded or unbounded).*

(a) The following bounds hold for the Wiener processes as in Assumption 1:

$$(i) \quad \max_{i=1,\dots,q_T} \sup_{0\le t\le \xi_T} \frac{1}{\sqrt{\xi_T}} \|\mathbf{W}_{\theta_i}^{(\theta_{i+1})} - \mathbf{W}_{\theta_i\pm t}^{(\theta_{i+1})}\| = O_P\left(\sqrt{\log 2q_T}\right),$$

$$(ii) \quad \sup_{\frac{D_T}{\|\mathbf{d}_i\|^2}\le s\le h_T} \frac{\sqrt{D_T}\left\|\mathbf{W}_{\theta_i}^{(\theta_i)} - \mathbf{W}_{\theta_i\pm s}^{(\theta_i)}\right\|}{s\,\|\mathbf{d}_i\|} = O_P(1),$$

$$(iii) \quad \max_{i=1,\dots,q_T} \sup_{\frac{D_T}{\|\mathbf{d}_i\|^2}\le s\le h_T} \frac{\sqrt{D_T}\left\|\mathbf{W}_{\theta_i}^{(\theta_i)} - \mathbf{W}_{\theta_i\pm s}^{(\theta_i)}\right\|}{s\,\|\mathbf{d}_i\|} = O_P\left(\sqrt{\log 2q_T}\right),$$

where neither the rates, nor the constants depend on D_T.

(b) The bound in (i) carries over to the centered increments of the original process:

$$\max_{i=1,\dots,q_T} \sup_{0\le t\le \xi_T} \frac{1}{\sqrt{\xi_T}} \|\widetilde{\mathbf{R}}_{\theta_i}^{(\theta_{i+1})} - \widetilde{\mathbf{R}}_{\theta_i\pm t}^{(\theta_{i+1})}\| = O_P\left(\sqrt{\log 2q_T}\right).$$

The bound in (ii) carries over if a forward and backward invariance principle as in Remark 2 exists starting in an arbitrary point θ_i. In this case (iii) carries over if $q_T = O(1)$.

For a single change point (instead of taking the maximum over all) the bound in (a) (i) and (b) is given by $O_P(1)$.

Proof. (a) Let $\mathbf{B}_t^{(j)} = (\mathbf{\Sigma}_T^{(j)})^{-1/2}\,\mathbf{W}_t^{(j)}$. Then by the self-similarity of Wiener processes it holds

$$\begin{aligned}
&\max_{i=1,\dots,q_T} \sup_{0\le t\le \xi_T} \frac{1}{\sqrt{\xi_T}} \|\mathbf{W}_{\theta_i}^{(\theta_{i+1})} - \mathbf{W}_{\theta_i+t}^{(\theta_{i+1})}\| \\
&\le O(1) \max_{j=1,\dots,P} \|(\mathbf{\Sigma}_T^{(j)})^{1/2}\| \max_{i=1,\dots,q_T} \sup_{0\le t\le 1} \|\mathbf{B}_{\theta_i}^{(\theta_{i+1})} - \mathbf{B}_{\theta_i+t}^{(\theta_{i+1})}\|.
\end{aligned}$$

By the uniform boundedness of the covariance matrices as in Assumption 1,

$$\max_{j=1,\ldots,P} \|(\mathbf{\Sigma}_T^{(j)})^{1/2}\| = \max_{j=1,\ldots,P} \sqrt{\|\mathbf{\Sigma}_T^{(j)}\|} = O(1).$$

The reflection principle in combination with tail probabilities for Gaussian random variables shows that with appropriate constants D_1, D_2 (not depending on i) it holds for all $D \geq 1$

$$P\left(\sup_{0\leq t\leq 1} \|\mathbf{B}_{\theta_i}^{(\theta_{i+1})} - \mathbf{B}_{\theta_i+t}^{(\theta_{i+1})}\| \geq D_1\sqrt{D\,\log 2q_T}\right) \leq \frac{D_2}{2^D\, q_T^D},$$

which in combination with subadditivity shows that

$$\max_{i=1,\ldots,q_T} \sup_{0\leq t\leq 1} \|\mathbf{B}_{\theta_i}^{(\theta_{i+1})} - \mathbf{B}_{\theta_i+t}^{(\theta_{i+1})}\| = O_P\left(\sqrt{\log 2q_T}\right).$$

The assertion without the maximum follows analogously.

Clearly, (ii) follows from (iii) so we will only prove the latter. As above it is sufficient to prove the assertion for $\{\mathbf{B}_t\}$. Due to the self-similarity of Wiener processes and its stationary and independent increments, it holds

$$\max_{i=1,\ldots,q_T} \sup_{\frac{D_T}{\|\mathbf{d}_i\|^2}\leq s\leq h_T} \frac{\sqrt{D_T}\left\|\mathbf{B}_{\theta_i+s}^{(\theta_i)} - \mathbf{B}_{\theta_i}^{(\theta_i)}\right\|}{s\,\|\mathbf{d}_i\|} \stackrel{\mathcal{D}}{=} \max_{j=1,\ldots,q_T} \sup_{1\leq t\leq h_T\|\mathbf{d}_j\|^2/D_T} \frac{\left\|\mathbf{B}_t^{(j)}\right\|}{t},$$

where $\{\mathbf{B}_t^{(j)}\}$, $j = 1, 2, \ldots$, are independent standard Wiener processes. Similar assertions hold for the other expressions. By the reflection principle and tail probabilities

for Gaussian random variables it holds for any $C > 4$

$$\begin{aligned} P\left(\sup_{t\geq 1}\frac{\|\mathbf{B}_t\|}{t} \geq \sqrt{C\ \log 2q_T}\right) &\leq \sum_{l\geq 1} P\left(\sup_{2^l\leq t<2^{l+1}}\frac{\|\mathbf{B}_t\|}{t} \geq \sqrt{C\ \log 2q_T}\right) \\ &\leq \sum_{l\geq 1} P\left(\sup_{0\leq t\leq 1}\|\mathbf{B}_t\| \geq \frac{2^l}{\sqrt{2^{l+1}}}\sqrt{C\ \log 2q_T}\right) \leq\ O(1)\sum_{l\geq 1}\left(O(1)2Cq_T\right)^{-2^l} \\ &= O\left(\frac{1}{4C^2q_T^2}\right), \end{aligned}$$

which shows the assertion in combination with the sub-additivity.

(b) By the invariance principle and (a) (i) it holds

$$\begin{aligned} \max_{i=1,\ldots,q_T}\ \sup_{\theta_i-h_T<t\leq\theta_i+h_T}\left\|\mathbf{\Lambda}_t(\widetilde{\mathbf{R}}_t)\right\| &\leq O_P\left(\frac{T^{1/2}\nu_T}{\sqrt{h_T}}\right) + \max_{i=1,\ldots,q_T}\ \sup_{\theta_i-h_T<t\leq\theta_i+h_T}\|\mathbf{\Lambda}_t(\mathbf{W})\| \\ &= O_P(\sqrt{\log 2q_T}). \end{aligned}$$

The other statement can be proven analogously. For the assertion in (ii) the invariance principle starting in θ_i backward or forward applied from θ_i to $\theta_i \pm h_T$ yields a rate of $h_T^{1/2}\nu_{h_T}$, which is strong enough to prove the rate analogously to above.

□

S2.1 Proofs of Section 3.3

Proof of Theorem 1.

(a) Because $\widehat{\mathbf{A}}_t$ is symmetric and positive definite such that the minimal eigenvalue of $\widehat{\mathbf{A}}_t^{-1}$ is given by $1/\|\widehat{\mathbf{A}}_t\|$ it holds

$$\mathbf{m}_t\widehat{\mathbf{A}}_t^{-1}\mathbf{m}_t \geq \frac{1}{\|\widehat{\mathbf{A}}_t\|}\|\mathbf{m}_t\|^2 = \frac{1}{\|\widehat{\mathbf{A}}_t\|}\frac{(h-|c_i-t|)^2}{2h}\|\mathbf{d}_i\|^2.$$

(b) By the invariance principle from Assumption 1 it holds by Assumption 2 that

$$\sup_{h\leq t\leq T-h} \|\mathbf{\Lambda}_t - \mathbf{\Lambda}_t(\mathbf{W})\| = O_P\left(\frac{T^{1/2}\nu_T}{\sqrt{h}}\right) = o_P\left(\sqrt{\log(T/h)}^{-1}\right), \tag{S2.1}$$

where $\mathbf{\Lambda}_t(\mathbf{W}_t)$ is the MOSUM statistics defined in (3.1) with $\{\mathbf{Z}_t\}$ there replaced by $\{\mathbf{W}_t\}$. Assertion (b)(i) follows immediately by the 1/2-self-similarity of the Wiener process with $\mathbf{B}_t = \mathbf{\Sigma}_t^{-1/2}\mathbf{W}_t$.

For the sub-linear case as in (ii) we get by (S2.1)

$$\begin{aligned} a\left(\frac{T}{h}\right) \sup_{h\leq t\leq T-h} \left\|\mathbf{\Sigma}_t^{-1/2}\mathbf{\Lambda}_t\right\| &= a\left(\frac{T}{h}\right) \sup_{h\leq t\leq T-h} \|\mathbf{\Lambda}_t(\mathbf{B})\| + o_P(1) \\ &\overset{\mathcal{D}}{=} \frac{1}{\sqrt{2}} \sup_{0\leq s\leq \frac{T}{h}-2} \|\mathbf{B}_{s+2} - 2\mathbf{B}_{s+1} + \mathbf{B}_s\| + o_P(1), \end{aligned}$$

where $(\mathbf{\Lambda}_t)_{t\geq 0}$ is a stationary process. Assertion (b)(ii) follows by Steinebach and Eastwood (1996), Lemma 3.1 in combination with Remark 1 with $\alpha = 1$ and $C_1 = \ldots = C_p = \frac{3}{2}$.

Replacing $\mathbf{\Sigma}_t$ by $\widehat{\mathbf{\Sigma}}_t$ does not change any of the above assertions by standard arguments.

(c) By splitting $\mathbf{\Lambda}_t(\widetilde{\mathbf{R}})$ into increments of length at most $2h$ anchored at the change points c_i we get by Proposition S2.1(b)(i)

$$\begin{aligned} &\max_{i=1,\ldots,q_T} \sup_{c_i-h<t\leq c_i+h} \left\|\mathbf{\Lambda}_t(\widetilde{\mathbf{R}})\right\| \\ &= O(1) \max_{i=1,\ldots,q_T} \sup_{0\leq t\leq h} \frac{1}{\sqrt{h}} \|\widetilde{\mathbf{R}}_{c_i}^{(c_{i+1})} - \widetilde{\mathbf{R}}_{c_i+t}^{(c_{i+1})}\| + O(1) \max_{i=1,\ldots,q_T} \sup_{0\leq t\leq h} \frac{1}{\sqrt{h}} \|\widetilde{\mathbf{R}}_{c_i}^{(c_i)} - \widetilde{\mathbf{R}}_{c_i-t}^{(c_i)}\| \\ &= O_P\left(\sqrt{\log 2q_T}\right). \end{aligned}$$

This shows that

$$\max_{i=1,\ldots,q_T} \sup_{c_i-h<t\leq c_i+h} \boldsymbol{\Lambda}_t' \boldsymbol{\Sigma}_t^{-1} \boldsymbol{\Lambda}_t = O_P(\log 2q_T),$$

In combination with (b) and the fact that there are only finitely many regimes (c) follows.

□

S2.2 Proofs of Section 4

We first prove consistency of the segmentation procedure.

Proof of Theorem 2. Define for $0 < \tau < 1$ the following set

$$S_T = S_T^{(1)} \cap S_T^{(2)} \cap \bigcap_{j=1}^{q_T} \left(S_T^{(3)}(j,\tau) \cap S_T^{(4)}(j,\tau) \right), \tag{S2.2}$$

where

$$\begin{aligned}
S_T^{(1)} &= \left\{ \max_{j=1,\ldots,q_T} \sup_{|t-c_j|>h} \mathbf{M}_t' \widehat{\mathbf{A}}_t^{-1} \mathbf{M}_t < \beta \right\}, \\
S_T^{(2)} &= \left\{ \min_{j=1,\ldots,q_T} \mathbf{M}_{c_j}' \widehat{\mathbf{A}}_{c_j}^{-1} \mathbf{M}_{c_j} \geq \beta \right\}, \\
S_T^{(3)}(j,\tau) &= \bigcap_{k=1}^{\lceil \frac{1}{\tau} \rceil - 1} \left\{ \sup_{c_j-h\leq t\leq c_j-k\tau h} \|\mathbf{M}_t\| < \|\mathbf{M}_{c_j-(k-1)\tau h}\| \right\}, \\
S_T^{(4)}(j,\tau) &= \bigcap_{k=1}^{\lceil \frac{1}{\tau} \rceil - 1} \left\{ \sup_{c_j+k\tau h\leq t\leq c_j+h} \|\mathbf{M}_t\| < \|\mathbf{M}_{c_j+(k-1)\tau h\|} \right\}.
\end{aligned}$$

On $S_T^{(1)}$ there are asymptotically no significant points outside of h-environments of the change points. On $S_T^{(2)}$ there is at least one significant time point for each change

point. On $S_T^{(3)}(j,\tau) \cap S_T^{(4)}(j,\tau)$ with $\tau < \eta/2$, there are no local extrema (within the h-environment of c_j) that are outside the interval $(c_j - \tau h, c_j + \tau h)$. Additionally, on $S_T^{(2)} \cap S_T^{(3)}(j,\tau) \cap S_T^{(4)}(j,\tau)$ the global extremum within that interval will be the only significant local extremum within the h-environment of c_j such that

$$\left\{ \max_{i=1,\ldots,\min(\hat{q}_T,q_T)} |\hat{c}_i - c_i| \leq \tau h, \hat{q}_T = q_T \right\} \supset S_T.$$

We will conclude the proof by showing that S_T is an asymptotic one set.

Indeed, $P(S_T^{(1)}) \to 1$ by Theorem 1 (c) on noting that

$$\mathbf{M}_t' \widehat{\mathbf{A}}_t^{-1} \mathbf{M}_t \leq \|\widehat{\mathbf{A}}_t^{-1}\| \, \|\mathbf{M}_t\|^2$$

and $P(S_T^{(2)}) \to 1$ by Theorem 1 (a) and (c).

Similarly, for $c_i - h \leq t \leq c_i$, we obtain that

$$\begin{aligned} &\|\mathbf{M}_{c_i-(k-1)\tau h}\| - \|\mathbf{M}_{c_i-k\tau h}\| \geq \|\mathbf{m}_{c_i-(k-1)\tau h}\| - \|\mathbf{m}_{c_i-k\tau h}\| + O_P\left(\sqrt{\log(T/h)}\right) \\ &\geq \frac{\tau}{\sqrt{2}} \sqrt{h} \, \|\mathbf{d}_i\| (1 + o_P(1)), \end{aligned}$$

where the o_P-term is uniform in i. This shows that $P\left(\bigcap_{j=1}^{q_T} S_T^{(3)}(j,\tau)\right) \to 1$. The assertion $P\left(\bigcap_{j=1}^{q_T} S_T^{(4)}(j,\tau)\right) \to 1$ follows analogously. □

With the above proposition we are ready to prove the localization rates for the change point estimators.

Proof of Theorem 3. On S_T as in (S2.2) it holds for any sequence ξ_T

$$\left\{\hat{c}_i - c_i < -C\xi_T^2/\|\mathbf{d}_i\|^2\right\} = \left\{\sup_{c_i-h\le t\le c_i-C\xi_T^2/\|\mathbf{d}_i\|^2} \|\mathbf{M}_t\|^2 \ge \sup_{c_i-C\xi_T^2/\|\mathbf{d}_i\|^2\le t\le c_i+h} \|\mathbf{M}_t\|^2\right\}$$
$$\subset \left\{\sup_{c_i-h\le t\le c_i-C\xi_T^2/\|\mathbf{d}_i\|^2} 2h\left(\|\mathbf{M}_t\|^2 - \|\mathbf{M}_{c_i}\|^2\right) \ge 0\right\}.$$

We will now show that the probability for the last set becomes arbitrarily small for C sufficiently large with $\xi_T = \omega_T$ as well as that the probability for the union of these sets over all change points $i = 1, \ldots, q_T$ becomes arbitrarily small for $\xi_T = \tilde{\omega}_T$. An analogous assertion can be shown for $\hat{c}_i > c_i + C\xi_T^2/\|\mathbf{d}_i\|^2$, completing the proof. For $c_i - h \le t < c_i$ the following decomposition holds

$$\mathbf{V}_t = \|\mathbf{M}_t\|^2 - \|\mathbf{M}_{c_i}\|^2 = -(\mathbf{m}_{c_i} - \mathbf{m}_t + \mathbf{\Lambda}_{c_i} - \mathbf{\Lambda}_t)'(\mathbf{m}_{c_i} + \mathbf{m}_t + \mathbf{\Lambda}_{c_i} + \mathbf{\Lambda}_t)$$
$$= -\frac{1}{2h}\left(D_{1,t}\,\mathbf{d}_i + \mathbf{N}_{1,t}\right)'\left(D_{2,t}\,\mathbf{d}_i + \mathbf{N}_{2,t}\right), \tag{S2.3}$$

where $D_{1,t} = c_i - t > 0, \quad D_{2,t} = 2h + t - c_i \ge h,$

$$\mathbf{N}_{1,t} = (\widetilde{\mathbf{R}}^{(c_i)}_{c_i-h} - \widetilde{\mathbf{R}}^{(c_i)}_{t-h}) + (\widetilde{\mathbf{R}}^{(c_{i+1})}_{c_i+h} - \widetilde{\mathbf{R}}^{(c_{i+1})}_{t+h}) - 2\left(\widetilde{\mathbf{R}}^{(c_i)}_{c_i} - \widetilde{\mathbf{R}}^{(c_i)}_{t}\right)$$
$$\mathbf{N}_{2,t} = (\widetilde{\mathbf{R}}^{(c_{i+1})}_{c_i+h} - \widetilde{\mathbf{R}}^{(c_{i+1})}_{t+h}) + 2\left(\widetilde{\mathbf{R}}^{(c_{i+1})}_{t+h} - \widetilde{\mathbf{R}}^{(c_{i+1})}_{c_i}\right) - (\widetilde{\mathbf{R}}^{(c_i)}_{c_i-h} - \widetilde{\mathbf{R}}^{(c_i)}_{t-h})$$
$$- 2\left(\widetilde{\mathbf{R}}^{(c_i)}_{t} - \widetilde{\mathbf{R}}^{(c_i)}_{c_i-h}\right).$$

We will concentrate on the proof of (b), where the proof of (a) is done analogously without the maximum over the change points and with the (possibly) tighter rate ω_T as in Assumption 5 (a) instead of $\tilde{\omega}_T$ as in (b). Indeed, Assumption 5 (b) immediately

implies that for any $\epsilon > 0$ there exists a C such that for any $y > 0$ it holds

$$P\left(\max_{i=1,\ldots,q_T} \sup_{c_i - h \le t \le c_i - C\tilde{\omega}_T^2/\|\mathbf{d}_i\|^2} \frac{\|\mathbf{N}_{1,t}\|}{D_{1,t}\,\|\mathbf{d}_i\|} \ge y\right)$$
$$= P\left(\max_{i=1,\ldots,q_T} \sup_{C\tilde{\omega}_T^2/\|\mathbf{d}_i\|^2 \le c_i - t \le h} \sqrt{C\tilde{\omega}_T^2}\,\frac{\|\mathbf{N}_{1,t}\|}{\|\mathbf{d}_i\||c_i - t|} \ge \sqrt{C}\,y\,\tilde{\omega}_T\right) \le \epsilon.$$

Similarly, by Proposition S2.1 (b)(i), it holds

$$\max_{i=1,\ldots,q_T} \sup_{c_i - h \le t \le c_i - C\tilde{\omega}_T^2/\|\mathbf{d}_i\|^2} \frac{\|\mathbf{N}_{2,t}\|}{D_{2,t}\,\|\mathbf{d}_i\|} = O_P\left(\sqrt{\frac{\log 2q_T}{h\|\mathbf{d}_i\|^2}}\right) = o_P(1),$$

where the last statement follows by Assumption 2 on noting that $q_T \le T/(2h)$.

Combining the above assertions with $P(S_T^c) = o_P(1)$ we obtain using the Cauchy-Schwarz inequality

$$P\left(\|\mathbf{d}_i\|^2(\hat{c}_i - c_i) < -C\tilde{\omega}_T^2 \text{ for some } i = 1,\ldots,q_T\right)$$
$$\le o_P(1) + P\Bigg(\max_{i=1,\ldots,q_T} \sup_{c_i - h \le t \le c_i - \frac{C\tilde{\omega}_T^2}{\|\mathbf{d}_i\|^2}} - D_{1,t}D_{2,t}\|\mathbf{d}_i\|^2$$
$$\cdot\left(1 + \frac{\mathbf{N}_{1,t}'\mathbf{d}_i}{D_{1,t}\|\mathbf{d}_i\|^2} + \frac{\mathbf{d}_i'\mathbf{N}_{2,t}}{D_{2,t}\|\mathbf{d}_i\|^2} + \frac{\mathbf{N}_{1,t}'\mathbf{N}_{2,t}}{D_{1,t}D_{2,t}\,\|\mathbf{d}_i\|^2}\right) \ge 0\Bigg)$$
$$\le o_P(1) + \mathrm{P}\left(\max_{i=1,\ldots,q_T} \sup_{c_i - h \le t \le c_i - \frac{C\tilde{\omega}_T^2}{\|\mathbf{d}_i\|^2}} \left|\frac{\mathbf{N}_{1,t}'\mathbf{d}_i}{D_{1,t}\|\mathbf{d}_i\|^2} + \frac{\mathbf{d}_i'\mathbf{N}_{2,t}}{D_{2,t}\|\mathbf{d}_i\|^2} + \frac{\mathbf{N}_{1,t}'\mathbf{N}_{2,t}}{D_{1,t}D_{2,t}\,\|\mathbf{d}_i\|^2}\right| \ge 1\right)$$
$$\le o_P(1) + P\left(\max_{i=1,\ldots,q_T} \sup_{c_i - h \le t \le c_i - \frac{C\tilde{\omega}_T^2}{\|\mathbf{d}_i\|^2}} \frac{\|\mathbf{N}_{1,t}\|}{D_{1,t}\|\mathbf{d}_i\|} \ge \frac{1}{3}\right) \le \epsilon$$

for C large enough (and ϵ arbitrary). This concludes the proof. □

Proof of Theorem 4. For $0 \leq c_i - t \leq D/\|\mathbf{d}_i\|^2$ it holds by Proposition S2.1 (b) (i) (the result without the maximum over all change points) with the notation as in (S2.3)

$$\max_{0\leq c_i - t \leq D/\|\mathbf{d}_i\|^2} \|\mathbf{N}_{1,t}\| = O_P\left(\frac{1}{\|\mathbf{d}_i\|}\right), \qquad \max_{0\leq c_i - t \leq D/\|\mathbf{d}_i\|^2} \|\mathbf{N}_{2,t}\| = O_P\left(\sqrt{h}\right),$$
$$\max_{0\leq c_i - t \leq D/\|\mathbf{d}_i\|^2} |D_{1,t}| = O\left(\frac{1}{\|\mathbf{d}_i\|^2}\right), \quad \max_{0\leq c_i - t \leq D/\|\mathbf{d}_i\|^2} |D_{2,t} - 2h| = O\left(\frac{1}{\|\mathbf{d}_i\|^2}\right).$$

Together with (S2.3) this shows

$$\mathbf{V}_t = -\|\mathbf{d}_i\|^2 |c_i - t| - \|\mathbf{d}_i\| \, \mathbf{N}'_{1,t}\mathbf{u}_i + O_P\left(\frac{1}{\sqrt{h}\,\|\mathbf{d}_i\|}\right)$$

By Assumption 2 it holds $\|\mathbf{d}_i\|^2 h \to \infty$ such that with the substitution $s = (t - c_i)\|\mathbf{d}_i\|^2$ with $-D \leq s \leq 0$ we get

$$\begin{aligned}\mathbf{V}_s = -|s| &+ \|\mathbf{d}_i\| \left(\mathbf{Y}^{(1)}_{D+s} - \mathbf{Y}^{(1)}_{D}\right)' \mathbf{u}_i - 2\|\mathbf{d}_i\| \left(\mathbf{Y}^{(21)}_{D+s} - \mathbf{Y}^{(21)}_{D}\right)' \mathbf{u}_i \\ &+ \|\mathbf{d}_i\| \left(\mathbf{Y}^{(3)}_{D+s} - \mathbf{Y}^{(3)}_{D}\right)' \mathbf{u}_i + o_P(1).\end{aligned}$$

Similarly, for $0 \leq t - c_i \leq D/\|\mathbf{d}_i\|^2$ and the same substitution now leading to $0 \leq s \leq D$ we get

$$\begin{aligned}\mathbf{V}_s = -|s| &+ \|\mathbf{d}_i\| \left(\mathbf{Y}^{(1)}_{D+s} - \mathbf{Y}^{(1)}_{D}\right)' \mathbf{u}_i - 2\|\mathbf{d}_i\| \left(\mathbf{Y}^{(22)}_{D+s} - \mathbf{Y}^{(22)}_{D}\right)' \mathbf{u}_i \\ &+ \|\mathbf{d}_i\| \left(\mathbf{Y}^{(3)}_{D+s} - \mathbf{Y}^{(3)}_{D}\right)' \mathbf{u}_i + o_P(1).\end{aligned}$$

Note that for $\|\mathbf{d}_i\|^2 |\hat{c}_i - c_i| \leq D$ it holds

$$\|\mathbf{d}_i\|^2(\hat{c}_i - c_i) \leq x \quad \iff \quad \max_{-D\leq s\leq x} \mathbf{V}_s \geq \max_{x<s\leq D} \mathbf{V}_s.$$

Now, first applying the functional central limit theorem from Assumption 6 and then letting $D \to \infty$ (in combination with Theorem 3, where now by assumption $\omega_T = 1$) yields the result. □

Bibliography

Cho, H. and C. Kirch (2021a). Data segmentation algorithms: Univariate mean change and beyond. *Econometrics and Statistics*, 2452–3062.

Cho, H. and C. Kirch (2021+b). Two-stage data segmentation permitting multiscale changepoints, heavy tails and dependence. *Annals of the Institute of Statistical Mathematics, to appear*.

Eichinger, B. and C. Kirch (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli 24*, 526–564.

Einmahl, U. (1987). Strong invariance principles for partial sums of independent random vectors. *Ann. Probab. 15*, 1419–1440.

Fearnhead, P. and G. Rigaill (2020). Relating and comparing methods for detecting changes in mean. *Stat*, e291.

Heunis, A. J. (2003). Strong invariance principle for singular diffusions. *Stochastic Processes and their Applications. 104*, 57–80.

Kuelbs, J. and W. Philipp (1980). Almost sure invariance principles for partial sums of mixing b-valued random variables. *Ann. Probab. 8*, 1003–1036.

Meier, A., H. Cho, and C. Kirch (2019). mosum: A package for moving sums in change point analysis. *Journal of Statistical Software 97(8)*, 1–42.

Messer, M., K. M. Costa, J. Roeper, and G. Schneider (2017). Multi-scale detection of rate changes in spike trains with weak dependencies. *Journal of Computational Neuroscience 42*, 187–201.

Messer, M., M. Kirchner, J. Schiemann, J. Roeper, R. Neininger, and G. Schneider (2014). A multiple filter test for the detection of rate changes in renewal processes with varying variance. *The Annals of Applied Statistics 8(4)*, 2027–2067.

Mihalache, S.-R. (2011). *Sequential Change-Point Detection for Diffusion Processes*. dissertation, Universität zu Köln.

Steinebach, J. and V. R. Eastwood (1996). Extreme value asymptotics for multivariate renewal processes. *Journal of multivariate analysis 56(2)*, 284–302.