# Supplementary material to 'Data Segmentation for Time Series Based on a General Moving Sum Approach'

Claudia Kirch and Kerstin Reckruehm

## A. Linear Regression

In this section, we will explore the difference in performance between the MOSUM-Wald and MOSUM-score methodology when applied to linear regression where we use the usual least squares methodology that requires no numerical approximation.
To this end, we consider a time series $Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta}+\varepsilon_i$, $i = 1, \ldots, n$, of length $n = 1000$ with exogenous regressors $\boldsymbol{X}_i = (1, X_{i,1}, X_{i,2})^T$ with $X_{i,1} \sim N(1, 1)$ and $X_{i,2} \sim N(2, 1)$ and i.i.d. standard normal errors. We include three change points at $200, 500$ and $800$ and the regression coefficients $\boldsymbol{\beta}_1 = (1, 2, 2)^T, \boldsymbol{\beta}_2 = (1, 1, 2)^T, \boldsymbol{\beta}_3 = (2, 1, 2)^T$ and $\boldsymbol{\beta}_4 = (2, 1, 1)^T$. We use the global least-squares regression estimator as inspection parameter $\widehat{\boldsymbol{\beta}}_{1,n}$ for the MOSUM-score statistic.
As discussed in Remarks 2 and 5 we can make minimal assumptions on the covariance estimators theoretically, but the small-sample performance crucially depends on this estimator.
As covariance estimators we use $\widehat{\boldsymbol{\Sigma}}^{(j)} = \widehat{v}_n^{(j)} \frac{1}{n} \sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$ with

$$\widehat{v}_n^{(1)} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_{1,n})^2, \tag{\textit{S-global}}$$

$$\widehat{v}_n^{(2)} = \frac{1}{2G}\left(\sum_{i=k-G+1}^{k}(\hat{\epsilon}_i - \bar{\epsilon}_{k-G+1,k})^2 + \sum_{i=k+1}^{k+G}(\hat{\epsilon}_i - \bar{\epsilon}_{k+1,k+G})^2\right),$$

$$\text{with } \hat{\epsilon}_i := Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_{1,n} \text{ and } \bar{\epsilon}_{l,u} := \frac{1}{u-l+1}\sum_{i=l}^{u}\hat{\epsilon}_i, \tag{\textit{S-local}}$$

$$\widehat{v}_n^{(3)} = \frac{1}{2G}\left(\sum_{i=k-G+1}^{k}(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_{k-G+1,k})^2 + \sum_{i=k+1}^{k+G}(Y_i - \boldsymbol{X}_i^T\widehat{\boldsymbol{\beta}}_{k+1,k+G})^2\right). \tag{\textit{W-local}}$$

We use the first two estimators with the MOSUM-score statistics while the last one is only used with the Wald statistics where the estimators $\widehat{\boldsymbol{\beta}}_{t+1,t+G}$ are already available. We do not use this estimator with the MOSUM-score statistic because their usage cancels the computational advantage of the MOSUM-score statistic over the MOSUM-Wald statistic.

The first two estimators $\widehat{\mathbf{\Sigma}}^{(j)}$, $j = 1, 2$, are only consistent in the no-change situation but do not fulfill assumption 8(a) in the presence of change points. Instead of using a threshold as in Remark 5, in this simulation study, we stick to the threshold as in (9) for all methods.

Table 1 gives the estimated number of change points as well as the detection rates for all three change points in the various settings. For a bandwidth of 50 the MOSUM-Wald procedure (with *W-local*) outperforms the MOSUM-score procedure (with *S-local*), while both procedures achieve a similar performance for bandwidth 100. The MOSUM-score procedure with the global estimator (*S-global*) is clearly inferior to both competing methods (in particular for smaller bandwidth and the second change point) emphasizing again the importance of the choice of the covariance estimator for the small sample performance.

In terms of computation time (for $G = n^{2/3}$) the MOSUM-score clearly outperforms the MOSUM-Wald statistics: Even with the local estimator (*S-local*), the MOSUM-score statistic was roughly 22 times faster in our simulations than the MOSUM-Wald statistic (with *W-local*) for a time series of length 1000 (an average (out of 100 runs) of 0.03 seconds as compared to 0.66 seconds). For a length of 8000 it was already more than 31 times faster (0.22 versus 6.91 seconds). The numbers only give a qualitative idea as we did not optimize any of the procedures with respect to computation time but merely used the R-function `rollsum` (from the R-package `zoo` Zeileis and Grothendieck (2005)) to calculate the MOSUM-statistics, where the local estimators for the covariance matrices are implemented naively with a loop and the local regression parameter for the Wald-statistics is calculated with the `lm`-function.

More simulation results for the linear regression situation including the false alarm rate in the no-change situation, the results for other bandwidths and covariance estimators and more information on computing times can be found in Reckrühm (2019), Section 4.1.

## B. Simulations for the Poisson Autoregressive Model

In this section we consider the Poisson autoregressive model with the estimation function corresponding to the partial likelihood as in Section 2.3.3. Compared to Section A this includes two additional difficulties: First, due to the serial dependence of the data the true scaling of the procedures depends on the long-run covariance rather than the covariance matrix. This time-dependency is also the reason behind the larger bandwidths compared to Section A. Secondly, there is no analytical solution to the estimating procedure such that numerical methods are required which as expected will greatly increase computation time for the MOSUM-Wald procedure. We consider a time series of length $n = 1000$ with three change points at times 250, 500 as well as 750 with the paramaters $\boldsymbol{\theta}_1 = (1, 0.5)^T$, $\boldsymbol{\theta}_2 = (2.5, 0.5)^T$, $\boldsymbol{\theta}_3 = (2.5, 0.2)^T$ as well as $\boldsymbol{\theta}_4 = (1, 0.5)^T$.

For the MOSUM-score statistics we use the global (partial) maximum likelihood estima-

| | Estimated number $\widehat{q}$ | | | | | Detection rate | | |
|---|---|---|---|---|---|---|---|---|
| $G$ | $\leq 1$ | 2 | 3 | 4 | $\geq 5$ | 200 | 500 | 800 |
| MOSUM-score with *S-global* covariance estimator | | | | | | | | |
| 50 | 0.484 | 0.489 | 0.027 | 0 | 0 | 0.494 | 0.027 | 0.993 |
| 100 | 0.003 | 0.468 | 0.518 | 0.011 | 0 | 0.969 | 0.515 | 0.999 |
| MOSUM-score with *S-local* covariance estimator | | | | | | | | |
| 50 | 0.110 | 0.502 | 0.353 | 0.034 | 0.001 | 0.804 | 0.430 | 1.000 |
| 100 | 0 | 0.049 | 0.918 | 0.033 | 0 | 0.985 | 0.917 | 1.000 |
| MOSUM-Wald with *W-local* covariance estimator | | | | | | | | |
| 50 | 0.018 | 0.445 | 0.501 | 0.035 | 0.001 | 0.963 | 0.539 | 1.000 |
| 100 | 0 | 0.030 | 0.945 | 0.025 | 0 | 0.998 | 0.938 | 1.000 |

Table 1: Number of estimated change points and detection rate for all three change points (i.e. percentage of simulations with a change point estimator in the interval $[k_{j,n} - 20, k_{j,n} + 20]$) for the various scenarios in the linear regression example.

tor as inspection parameter as well as the one based on the observations between time point 300 and 700 (compare also Remark 4). To estimate the covariance matrix we use the following local estimator

$$\begin{aligned}
&\widehat{\boldsymbol{\Sigma}}_{k,n} \\
&= \frac{1}{2G} \sum_{i=k-G+1}^{k} \left( \boldsymbol{H}(\mathbb{Y}_i, \widehat{\boldsymbol{\theta}}_{1,n}) - \overline{\boldsymbol{H}}_{k-G+1,k} \right) \left( \boldsymbol{H}(\mathbb{Y}_i, \widehat{\boldsymbol{\theta}}_{1,n}) - \overline{\boldsymbol{H}}_{k-G+1,k} \right)^T \\
&\quad + \frac{1}{2G} \sum_{i=k+1}^{k+G} \left( \boldsymbol{H}(\mathbb{Y}_i, \widehat{\boldsymbol{\theta}}_{1,n}) - \overline{\boldsymbol{H}}_{k+1,k+G} \right) \left( \boldsymbol{H}(\mathbb{Y}_i, \widehat{\boldsymbol{\theta}}_{1,n}) - \overline{\boldsymbol{H}}_{k+1,k+G} \right)^T,
\end{aligned}$$

where $\overline{\boldsymbol{H}}_{l,u}$ denotes the sample mean of $\boldsymbol{H}(\mathbb{Y}_l, \widehat{\boldsymbol{\theta}}_{1,n}), \ldots, \boldsymbol{H}(\mathbb{Y}_u, \widehat{\boldsymbol{\theta}}_{1,n})$. However, this estimator does not take the dependence into account and as such estimates the covariance rather than the long-run covariance matrix, such that Assumption 8(a) is not fulfilled. As before we use a threshold as in (9) despite Remark 5.
Motivated by Weiß (2010), equations (7) and (8), we use the estimator $\widetilde{\boldsymbol{\Gamma}}^{-1}_{k,n} = \frac{1}{2}\left( \widetilde{\boldsymbol{\Gamma}}^{-1}_{k-G+1,k} + \widetilde{\boldsymbol{\Gamma}}^{-1}_{k+1,k+G} \right)$ in the MOSUM-Wald procedure, where

$$\widetilde{\boldsymbol{\Gamma}}^{-1}_{l,u} = \frac{1}{G} \sum_{i=l}^{u} \frac{1}{(\boldsymbol{Y}_{i-1}^T \widehat{\boldsymbol{\theta}}^{ML}_{l,u})^2} \begin{pmatrix} Y_i & Y_i Y_{i-1} \\ Y_i Y_{i-1} & Y_i Y_{i-1}^2 \end{pmatrix}.$$

Table 2 gives the estimated number of change points as well as the detection rates for all three change points in the various setting. In this case, the MOSUM-Wald statistics outperforms the score procedures in terms of detection rates for the second and third change point but indeed the MOSUM-score with the restricted estimator (rather than the global one) slightly outperforms the MOSUM-Wald for the first one.

| | **Estimated number $\widehat{q}$** | | | | | **Detection rate** | | |
|---|---|---|---|---|---|---|---|---|
| $G$ | $\leq 1$ | 2 | 3 | 4 | $\geq 5$ | 250 | 500 | 750 |
| MOSUM-score procedure with $\widehat{\boldsymbol{\theta}}_{1,1000}$ | | | | | | | | |
| 80 | 0.619 | 0.288 | 0.063 | 0.028 | 0.002 | 0.713 | 0.135 | 0.242 |
| 150 | 0.056 | 0.321 | 0.449 | 0.137 | 0.037 | 0.921 | 0.583 | 0.623 |
| MOSUM-score procedure with $\widehat{\boldsymbol{\theta}}_{300,700}$ | | | | | | | | |
| 80 | 0.100 | 0.397 | 0.300 | 0.143 | 0.060 | 0.936 | 0.199 | 0.734 |
| 150 | 0.018 | 0.162 | 0.596 | 0.194 | 0.030 | 0.919 | 0.724 | 0.742 |
| MOSUM-Wald procedure | | | | | | | | |
| 80 | 0.069 | 0.295 | 0.373 | 0.199 | 0.064 | 0.890 | 0.603 | 0.645 |
| 150 | 0.001 | 0.040 | 0.629 | 0.261 | 0.069 | 0.896 | 0.809 | 0.803 |

Table 2: Number of estimated change points and detection rate for all three change points (i.e. percentage of simulations with a change point estimator in the interval $[k_{j,n} - 20, k_{j,n} + 20]$) for the various scenarios in the Poisson autoregressive example.

The MOSUM-score is computationally much cheaper and scales better with longer series. Indeed, the median computation time (for 100 runs and $G = n^{2/3}$) for the MOSUM-Wald statistics for a length of $n = 1000$ (running several minutes) was more than 272 times slower than the MOSUM-score statistic (running less than a second). For $n = 8000$ the MOSUM-Wald statistics ran for more than half an hour which was more than 362 times slower than the MOSUM-score (which ran for less than 6 seconds). In particular, the MOSUM-Wald statistic is significantly slowed down by the need of numerical optimization in comparison to the linear regression where no numerical methods are required. The numbers only give a qualitative idea where a naive loop was used for the calculation of the statistics and the global and local parameter estimators for the Poisson regression were calculated with the R-package `tscount` Liboschik, Fokianos, and Fried (2017).

In conclusion, the MOSUM-score statistic can be a good way to generate change point candidates by means of using different inspection parameters (and bandwidths) even in combination with a sub-optimal covariance estimation procedure.

More simulation results including the false alarm rate in the no-change situation, the results for other bandwidths and covariance estimators, results for the least-squares estimators and more information on computing times can be found in Reckrühm (2019), Section 4.2.

# C. Proofs of Section 2.4

*Proof of Theorem 1.* We first prove the assertions for the MOSUM-score statistics ($\ell = 2$): By the invariance principle in Assumption 1 (b) as well as Assumption 3 on the bandwidth we get (with $\mathbf{\Sigma} = \mathbf{\Sigma}_1(\widetilde{\theta})$) that

$$\begin{aligned}
&\max_{G \leq k \leq n-G} \frac{1}{\sqrt{2G}} \left\| \mathbf{\Sigma}^{-1/2} \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}}(k) \right\| \\
&= \max_{G \leq k \leq n-G} \frac{1}{\sqrt{2G}} \left\| \boldsymbol{W}(k+G) - 2\boldsymbol{W}(k) + \boldsymbol{W}(k-G) \right\| + o_P\left(a(n/G)^{-1}\right) \\
&= \sup_{r \in [G, n-G]} \frac{1}{\sqrt{2G}} \left\| \boldsymbol{W}(r+G) - 2\boldsymbol{W}(r) + \boldsymbol{W}(r-G) \right\| + o_P\left(a(n/G)^{-1}\right) \\
&\overset{D}{=} \sup_{t \in [1, n/G-1]} \frac{1}{\sqrt{2}} \left\| \boldsymbol{W}(t+1) - 2\boldsymbol{W}(t) + \boldsymbol{W}(t-1) \right\| + o_P\left(a(n/G)^{-1}\right),
\end{aligned}$$

where we used the self-similarity of the Wiener process in the last step. By an application of Lemma 3.1 in combination with Remark 3.1 of Steinebach and Eastwood (1996) with $\alpha = 1$ and $C_1 = \ldots = C_p = 3/2$, assertion (a) for the MOSUM-score statistics follows. The assertion for the MOSUM-Wald statistics ($\ell = 1$) follows from this and Assumption 2. The assertions in (b) follow successively by an application of the triangular inequality in combination with the consistency of the spectral matrix norm with the Euclidean vector norm (as the corresponding induced norm). The assertions in (c) are obtained analogously. □

# D. Proofs of Section 3.1

*Proof of Proposition 1.* First, by Theorem 1,

$$\begin{aligned}
&P\left( \max_{j=1,\ldots,q+1} \max_{k_{j-1,n}+G \leq k \leq k_{j,n}-G} T_{k,n}^{(1)}(G) \geq D_n(\alpha_n, G) \right) \\
&\leq \sum_{j=1}^{q+1} P\left( a(n/G) \max_{k_{j-1,n}+G \leq k \leq k_{j,n}-G} T_{k,n}^{(1)}(G) - b(n/G) \geq c_{\alpha_n} \right) \\
&\leq \sum_{j=1}^{q+1} (\alpha_n + o(1)) \to 0,
\end{aligned}$$

showing (a) (i). Furthermore, because the spectral matrix norm is induced by the Euclidean vector norm, it holds $\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|$ as well as $\|\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{A}\mathbf{x}\|$. Then, by Assumption 6 (a) applied to the second term (which is dominated by the maximum in the assumption as can be seen by an index shift),

$$
\begin{aligned}
&\min_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} T^{(1)}_{k,n}(G)\\
&\geq \min_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} \sqrt{\frac{G}{2}}\left\|\mathbf{\Gamma}_k^{-1/2}\left(\widehat{\boldsymbol{\theta}}_{k+1,k+G}-\boldsymbol{\theta}_j\right)\right\|\\
&\qquad - \max_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} \sqrt{\frac{G}{2}}\left\|\mathbf{\Gamma}_k^{-1/2}\left(\widehat{\boldsymbol{\theta}}_{k-G+1,k}-\boldsymbol{\theta}_j\right)\right\|\\
&\geq \left\|\mathbf{\Gamma}_{(j)}^{1/2}\right\|^{-1} \min_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} \sqrt{\frac{G}{2}}\left\|\widehat{\boldsymbol{\theta}}_{k+1,k+G}-\boldsymbol{\theta}_j\right\|\\
&\qquad - \left\|\mathbf{\Gamma}_{(j)}^{-1/2}\right\| \max_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} \sqrt{\frac{G}{2}}\left\|\widehat{\boldsymbol{\theta}}_{k-G+1,k}-\boldsymbol{\theta}_j\right\|\\
&= \left\|\mathbf{\Gamma}_{(j)}^{1/2}\right\|^{-1} \min_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} \sqrt{\frac{G}{2}}\left\|\widehat{\boldsymbol{\theta}}_{k+1,k+G}-\boldsymbol{\theta}_j\right\| + O_P\left(\sqrt{\log(n/G)}\right).
\end{aligned}
$$

We get by Assumption 5 and 6 (b)

$$
\begin{aligned}
&P\left(\min_{j=1,\ldots,q}\min_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} T^{(1)}_{k,n}(G) < D_n(\alpha_n,G)\right)\\
&\le \sum_{j=1}^{q} P\left(\min_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}} \sqrt{\frac{G}{2}}\left\|\widehat{\boldsymbol{\theta}}_{k+1,k+G}-\boldsymbol{\theta}_j\right\| < O_P\left(\sqrt{\log(n/G)}\right)\right)\\
&\to 0.
\end{aligned}
$$

Similar arguments deal with the minimum over $k_{j,n} < k \le k_{j,n} + (1-\varepsilon)\,G$, such that assertion (a)(ii) follows.
The proof of (b) follows along the same lines taking Assumption 7 into account. □

*Proof of Remark 2.* The proof is analogous to the above proofs where we use that by Theorem 1

$$
\max_{j=1,\ldots,q+1}\max_{k_{j-1,n}+G\le k\le k_{j,n}-G}\left\|T^{(1)}_{k,n}(G)\right\| = O_P\left(\sqrt{\log(n/G)}\right).
$$

□

*Proof of Theorem 2.* The assertion follows immediately from Proposition 1 on noting that

$$
\begin{aligned}
&\left\{\max_{j=1,\ldots,q+1}\max_{k_{j-1,n}+G\le k\le k_{j,n}-G} T^{(1)}_{k,n}(G) < D_n(\alpha_n,G)\right\}\\
&\quad\cap\left\{\min_{k_{j,n}-(1-\varepsilon)\,G\le k\le k_{j,n}+(1-\varepsilon)\,G} T^{(1)}_{k,n}(G) \ge D_n(\alpha_n,G)\right\}\\
&\subset\left\{\widehat{q}^{(1)}_n = q\right\}\cap\left\{\max_{1\le j\le q}\left|\widehat{k}^{(1)}_{j,n}-k_{j,n}\right| < G\right\}.
\end{aligned}
$$

□

# E. Proofs of Section 3.2

*Proof of Lemma 1.* We will show by contradiction that $\boldsymbol{EH}(\mathbb{X}_1^{(j)}, \widetilde{\boldsymbol{\theta}}_{0,1}) \neq \boldsymbol{EH}(\mathbb{X}_1^{(j+1)}, \widetilde{\boldsymbol{\theta}}_{0,1})$ holds for at least one $j \in \{1, \dots, q\}$. Assume that all these expectations are equal, then by definition of $\widetilde{\boldsymbol{\theta}}_{0,1}$ we get for all $j = 1, \dots, q+1$

$$\mathbf{0} = \sum_{l=1}^{q+1} (\lambda_l - \lambda_{l-1}) \, E\boldsymbol{H}(\mathbb{X}_1^{(l)}, \widetilde{\boldsymbol{\theta}}_{0,1}) = E\boldsymbol{H}(\mathbb{X}_1^{(j)}, \widetilde{\boldsymbol{\theta}}_{0,1}),$$

which by the identifiability of $\boldsymbol{\theta}_j$ implies $\boldsymbol{\theta}_j = \widetilde{\boldsymbol{\theta}}_{0,1}$, for all $j = 1, \dots, q+1$, contradicting the assumption. If there are only two possible regimes, then clearly if one change is detectable all of them are. □

*Proof of Proposition 2.* Analogously to the proof of Proposition 1 (a) (i) we get

$$P\left( \max_{j=1,\dots,q+1} \max_{k_{j-1,n}+G \le k \le k_{j,n}-G} T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}}) \ge D_n(\alpha_n, G) \right) \to 0.$$

The statement remains true when a sequence of inspection parameters $\widetilde{\boldsymbol{\theta}}_n$ is used because by Assumption 4 (i) it holds for any $j = 1, \dots, q$ that

$$\begin{aligned} & a(n/G) \max_{k_{j-1,n}+G \le k \le k_{j,n}-G} T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}}_n) \\ & = a(n/G) \max_{k_{j-1,n}+G \le k \le k_{j,n}-G} T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}}) + o_P(1). \end{aligned}$$

Additionally, we need the statement for environments of non-detectable change points $k_{j,n}$ with $j \notin \tilde{Q}$. Indeed, it holds for $|k - k_{j,n}| \le G$ with $j \notin \tilde{Q}$ that $E\boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}}(k) = 0$ by definition of $\tilde{Q}$ as in (10). Consequently, by Assumptions 1 and 3 it holds for $j \notin \tilde{Q}$

$$\begin{aligned} & \max_{k_{j,n} \le k < k_{j,n}+G} T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}}) \\ & = o_P(1) + O_P\left( \max_{k_{j,n} \le k < k_{j,n}+G} \frac{1}{\sqrt{G}} \left\| \boldsymbol{W}(k+G) - 2\boldsymbol{W}(k) + \boldsymbol{W}(k_{j,n}) \right\| \right) \\ & \qquad + O_P\left( \max_{k_{j,n} \le k < k_{j,n}+G} \frac{1}{\sqrt{G}} \left\| \boldsymbol{\Sigma}_{(j+1)}^{-1/2} \boldsymbol{\Sigma}_{(j)}^{1/2} \left( \boldsymbol{W}(k_{j,n}) - \boldsymbol{W}(k-G) \right) ) \right\| \right) \\ & = O_P(1) = o_P(D_n(\alpha_n, G)), \end{aligned} \tag{1}$$

where the last line follows by the self-similarity of Wiener processes, the stationarity of its increments and the continuous sample paths. An analogous assertion holds for $k_{j,n} - G \le k < k_{j,n}$ showing that

$$P\left( \max_{j \notin \tilde{Q}} \max_{|k-k_{j,n}|<G} T_{k,n}^{(2)}(G, \widetilde{\boldsymbol{\theta}}) \ge D_n(\alpha_n, G) \right) \to 0,$$

completing the proof of (a) (i) for a fixed inspection parameter $\widetilde{\boldsymbol{\theta}}$. Here, the statement remains true for a sequence of inspection parameters, because by Assumption 4 (ii) it holds for $j \notin \tilde{Q}$

$$\max_{|k-k_{j,n}|<G} T^{(2)}_{k,n}(G,\widetilde{\boldsymbol{\theta}}_n) = \max_{|k-k_{j,n}|<G} T^{(2)}_{k,n}(G,\widetilde{\boldsymbol{\theta}}) + o_P(\sqrt{\log(n/G)}) \tag{2}$$
$$= o_P(D_n(\alpha_n, G)).$$

The assertion with estimated long-run covariances as in (b) can be obtained along the same lines by using the consistency of the spectral matrix norm with the Euclidean vector norm and Assumptions 8 (b).
Concerning (ii) first observe that for $\tilde{k}_{j,n} < k \leq \tilde{k}_{j,n} + (1-\varepsilon)G$, $j = 1, \ldots, \tilde{q}(\widetilde{\boldsymbol{\theta}})$ it holds

$$\begin{aligned}
&E\boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}}(k)\\
&= \sum_{i=k+1}^{k+G} E\boldsymbol{H}(\mathbb{X}_i^{(j+1)}, \widetilde{\boldsymbol{\theta}}) - \sum_{i=k-G+1}^{\tilde{k}_{j,n}} E\boldsymbol{H}(\mathbb{X}_i^{(j)}, \widetilde{\boldsymbol{\theta}}) - \sum_{i=\tilde{k}_{j,n}+1}^{k} E\boldsymbol{H}(\mathbb{X}_i^{(j+1)}, \widetilde{\boldsymbol{\theta}})\\
&= \left(G - |k - \tilde{k}_{j,n}|\right) \boldsymbol{d}_j,
\end{aligned}$$

where we denote the signal by

$$\boldsymbol{d}_j = E\boldsymbol{H}(\mathbb{X}_1^{(j+1)}, \widetilde{\boldsymbol{\theta}}) - E\boldsymbol{H}(\mathbb{X}_1^{(j)}, \widetilde{\boldsymbol{\theta}}), \quad j = 1, \ldots \tilde{q}(\widetilde{\boldsymbol{\theta}}). \tag{3}$$

For $\tilde{k}_{j,n} - (1-\varepsilon)G \leq k \leq \tilde{k}_{j,n}$ we arrive at the same conclusion. Consequently, it holds for all $j \in \tilde{Q}$ and $|k - \tilde{k}_{j,n}| \leq (1-\varepsilon)G$ by the consistency of the spectral matrix with the Euclidean vector norm

$$\left\| \boldsymbol{\Sigma}_k^{-1/2} E\boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}}(k) \right\| \geq \left\| \boldsymbol{\Sigma}_k^{1/2} \right\|^{-1} \left(G - |k - \tilde{k}_{j,n}|\right) \|\boldsymbol{d}_j\| \geq c\,G,$$

for some $c > 0$ (depending on $\varepsilon$, the difference in expectation and the long-run covariances, noting that $\boldsymbol{\Sigma}_k$ is constant on each segment).
By analogous arguments as in (1) (but involving the necessary centering due to $\tilde{Q}$) and (2) it holds

$$\begin{aligned}
&\min_{|k-\tilde{k}_{j,n}|\leq(1-\varepsilon)\,G} T^{(2)}_{k,n}(G, \widetilde{\boldsymbol{\theta}}_n)\\
&= \min_{|k-\tilde{k}_{j,n}|\leq(1-\varepsilon)\,G} \frac{1}{\sqrt{2G}} \left\| \boldsymbol{\Sigma}_k^{-1/2} E\boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}}(k) \right\| + o_P(\sqrt{\log(n/G)})\\
&\geq c\sqrt{\frac{G}{2}} + o_P\left(\sqrt{\log(n/G)}\right).
\end{aligned}$$

The proof can now be concluded as in the proof of Proposition 1 (a) and (b) (ii). □

*Proof of Theorem 3 and Remark 5.* The proofs are completely analogous to the proofs of Theorem 2 respectively Remark 2 and therefore omitted. □

The following lemma helps simplify several arguments.

**Lemma 1.** *For a sequence of real random variables $\{X_n\}$ it holds for $n \to \infty$*

$$X_n = O_P(1) \qquad \Longleftrightarrow \qquad P\left(|X_n| > \xi_n\right) \to 0 \quad \textit{for any } \xi_n \to \infty.$$

*Proof.* The proof of the only-if-part is straightforward. We prove the if-part by contradiction. If $X_n$ is not stochastically bounded, then there exists $\eta > 0$ such that for any bound $C > 0$ and any $n_0 \geq 0$, there exists $n_1(n_0, C) > n_0$ such that $P\left(|X_{n_1}| > C\right) > \eta$. Setting $N_0 = 0$ and recursively $N_l = n_1(N_{l-1}, l)$ as well as $\xi_{N_{l-1}+1} = \ldots = \xi_{N_l} = l$, we get $N_n \to \infty$, $\xi_n \to \infty$ as well as by construction

$$P\left(|X_{N_l}| > \xi_{N_l}\right) > \eta,$$

which is a contradiction. □

The proof technique of the below proof is well known in change point analysis, for example it has been used in the context of MOSUM statistics for the mean change problem by Eichinger and Kirch (2018) (Proof of Theorem 3.2).

*Proof of Theorem 4.* By finiteness of $q$ and Lemma 1 it is sufficient to prove that for any sequence $\xi_n \to \infty$ (arbitrarily slow) it holds

$$\begin{aligned}
&P\left(\widehat{k}^{(2)}_{j,n}(\widetilde{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\Psi}}_{j,n}) < \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}}) - \xi_n\right) \to 0,\\
&P\left(\widehat{k}^{(2)}_{j,n}(\widetilde{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\Psi}}_{j,n}) > \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}}) + \xi_n\right) \to 0.
\end{aligned}$$

We will prove the first assertion in detail, the second one follows analogously. For simplicity of notation denote $\tilde{k}_{j,n} = \tilde{k}_{j,n}(\widetilde{\boldsymbol{\theta}})$ throughout this proof.
On the asymptotic 1-set of Theorem 3 and where $\min_{j=1,\ldots,q+1} |k_{j,n} - k_{j-1,n}| > 2G$ (which holds for $n$ large enough by Assumption 3 (b)), it holds for any $j = 1, \ldots, \tilde{q}(\widetilde{\boldsymbol{\theta}})$ with $\boldsymbol{d}_j$ as in (3)

$$\begin{aligned}
&\widehat{k}^{(2)}_{j,n}(\widetilde{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\Psi}}_{j,n}) = \underset{v_{j,n} \leq k \leq w_{j,n}}{\arg\max}\ V^{(j)}_{k,n}(G, \widetilde{\boldsymbol{\theta}}_n), \quad \text{where}\\
&V^{(j)}_{k,n}(G, \widetilde{\boldsymbol{\theta}}_n) = \left\| \widehat{\boldsymbol{\Psi}}^{-1/2}_{j,n} \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k) \right\|^2 - \left\| \widehat{\boldsymbol{\Psi}}^{-1/2}_{j,n} \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k_{j,n}) \right\|^2\\
&\quad = -\left(\boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(\tilde{k}_{j,n}) - \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k)\right) \widehat{\boldsymbol{\Psi}}^{-1}_{j,n} \left(\boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(\tilde{k}_{j,n}) + \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k)\right)\\
&\quad =: -\left(\boldsymbol{E}_1(k, G, \widetilde{\boldsymbol{\theta}}_n) + \boldsymbol{d}_j(\tilde{k}_{j,n} - k)\right) \widehat{\boldsymbol{\Psi}}^{-1}_{j,n} \left(\boldsymbol{E}_2(k, G, \widetilde{\boldsymbol{\theta}}_n) + \boldsymbol{d}_j(2G + k - \tilde{k}_{j,n})\right).
\end{aligned}$$

Denote $\Delta\boldsymbol{H}(\mathbb{X}^{(j)}_i, \widetilde{\boldsymbol{\theta}}_n) = \boldsymbol{H}(\mathbb{X}^{(j)}_i, \widetilde{\boldsymbol{\theta}}_n) - \boldsymbol{H}(\mathbb{X}^{(j)}_i, \widetilde{\boldsymbol{\theta}})$ and $\boldsymbol{H}_0(\mathbb{X}^{(j)}_i, \widetilde{\boldsymbol{\theta}}) = \boldsymbol{H}(\mathbb{X}^{(j)}_i, \widetilde{\boldsymbol{\theta}}) - E\boldsymbol{H}(\mathbb{X}^{(j)}_i, \widetilde{\boldsymbol{\theta}})$.

Then, it holds for $k < \tilde{k}_{j,n}$

$$\begin{aligned}
&\boldsymbol{E}_1(k,G,\widetilde{\boldsymbol{\theta}}_n) = \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(\tilde{k}_{j,n}) - \boldsymbol{M}_{\widetilde{\boldsymbol{\theta}}_n}(k) - \boldsymbol{d}_j(\tilde{k}_{j,n}-k)\\
&= \sum_{i=k+G+1}^{\tilde{k}_{j,n}+G} \Delta\boldsymbol{H}(\mathbb{X}_i^{(j+1)},\widetilde{\boldsymbol{\theta}}_n) + \sum_{i=k-G+1}^{\tilde{k}_{j,n}-G} \Delta\boldsymbol{H}(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}}_n) - 2\sum_{i=k+1}^{\tilde{k}_{j,n}} \Delta\boldsymbol{H}(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}}_n)\\
&\quad + \sum_{i=k+G+1}^{\tilde{k}_{j,n}+G} \boldsymbol{H}_0(\mathbb{X}_i^{(j+1)},\widetilde{\boldsymbol{\theta}}) + \sum_{i=k-G+1}^{\tilde{k}_{j,n}-G} \boldsymbol{H}_0(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}}) - 2\sum_{i=k+1}^{\tilde{k}_{j,n}} \boldsymbol{H}_0(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}}),
\end{aligned}$$

where by Assumption 10 (see also Remark 7 for the situation when $k > \tilde{k}_{j,n}$) and stationarity of the segments it follows

$$\sup_{\xi_n<\tilde{k}_{j,n}-k\leq G} \frac{\|\boldsymbol{E}_1(k,G,\widetilde{\boldsymbol{\theta}}_n)\|}{\tilde{k}_{j,n}-k} = o_P(1). \tag{4}$$

Furthermore,

$$\begin{aligned}
&\boldsymbol{E}_2(k,G,\widetilde{\boldsymbol{\theta}}_n)\\
&= -\boldsymbol{E}_1(k,G,\widetilde{\boldsymbol{\theta}}_n)\\
&\quad + 2\sum_{i=\tilde{k}_{j,n}+1}^{\tilde{k}_{j,n}+G} \Delta\boldsymbol{H}(\mathbb{X}_i^{(j+1)},\widetilde{\boldsymbol{\theta}}_n) - 2\sum_{i=\tilde{k}_{j,n}-G+1}^{\tilde{k}_{j,n}} \Delta\boldsymbol{H}(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}}_n)\\
&\quad + 2\sum_{i=\tilde{k}_{j,n}+1}^{\tilde{k}_{j,n}+G} \boldsymbol{H}_0(\mathbb{X}_i^{(j+1)},\widetilde{\boldsymbol{\theta}}) - 2\sum_{i=\tilde{k}_{j,n}-G+1}^{\tilde{k}_{j,n}} \boldsymbol{H}_0(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}}),
\end{aligned}$$

such that by (4), Assumption 10 (a) and stationarity of the segments in combination with the law of large number (that follows from Assumption 1, see also Remark 7) it holds

$$\sup_{\xi_n<\tilde{k}_{j,n}-k\leq G} \frac{\|\boldsymbol{E}_2(k,G,\widetilde{\boldsymbol{\theta}}_n)\|}{G} = o_P(1).$$

By Assumptions 9 and the consistency of the spectral matrix norm with the Euclidean vector norm we conclude

$$\begin{aligned}
&\sup_{\xi_n<\tilde{k}_{j,n}-k\leq G} V_{k,n}^{(j)}(G,\widetilde{\boldsymbol{\theta}}_n)\\
&\leq (\boldsymbol{d}_j\widehat{\boldsymbol{\Psi}}_{j,n}^{-1}\boldsymbol{d}_j + o_P(1)) \sup_{\xi_n<\tilde{k}_{j,n}-k\leq G} [-(\tilde{k}_{j,n}-k)(2G+k-\tilde{k}_{j,n})].
\end{aligned}$$

Finally,

$$\sup_{\xi_n<\tilde{k}_{j,n}-k\leq G} [-(\tilde{k}_{j,n}-k)(2G+k-\tilde{k}_{j,n})] \leq -\xi_n G < 0,$$

and

$$\boldsymbol{d}_j \widehat{\boldsymbol{\Psi}}_{j,n}^{-1} \boldsymbol{d}_j = \|\widehat{\boldsymbol{\Psi}}_{j,n}^{-1/2} \boldsymbol{d}_j\|^2 \geq \|\widehat{\boldsymbol{\Psi}}_{j,n}^{1/2}\|^{-2} \|\boldsymbol{d}_j\|^2$$

Consequently,

$$\begin{aligned}
&P\left(\widehat{k}_{j,n}^{(2)}(\widetilde{\boldsymbol{\theta}}_n; \widehat{\boldsymbol{\Psi}}_{j,n}) < \tilde{k}_{j,n} - \xi_n\right) \\
&\leq P\left(\sup_{\tilde{k}_{j,n}-G \leq k < \tilde{k}_{j,n}-\xi_n} V_{k,n}^{(j)}(G, \widetilde{\boldsymbol{\theta}}_n) \geq \sup_{\tilde{k}_{j,n}-\xi_n \leq k \leq \tilde{k}_{j,n}+G} V_{k,n}^{(j)}(G, \widetilde{\boldsymbol{\theta}}_n)\right) + o(1) \\
&\leq P\left(\sup_{\tilde{k}_{j,n}-G \leq k < \tilde{k}_{j,n}-\xi_n} V_{k,n}^{(j)}(G, \widetilde{\boldsymbol{\theta}}_n) \geq 0\right) + o(1) \\
&\leq P\left((\boldsymbol{d}_j \widehat{\boldsymbol{\Psi}}_{j,n}^{-1} \boldsymbol{d}_j + o_P(1)) \sup_{\xi_n < \tilde{k}_{j,n}-k \leq G} [-(\tilde{k}_{j,n} - k)(2G + k - \tilde{k}_{j,n})] \geq 0\right) + o(1) \\
&\leq P\left(|o_P(1)| \geq \boldsymbol{d}_j \widehat{\boldsymbol{\Psi}}_{j,n}^{-1} \boldsymbol{d}_j\right) + o(1) \leq P\left(|o_P(1)| \, \|\widehat{\boldsymbol{\Psi}}_{j,n}^{1/2}\|^2 \geq \|\boldsymbol{d}_j\|^2\right) + o(1) \\
&= o(1),
\end{aligned}$$

where the last line follows from Assumption 9, concluding the proof. □

## F. Proofs of Section 4

*Proof of Theorem 5.* The assertion in (a) follows by Theorem 4 of Kuelbs and Philipp (1980) on noting that the mixing rate of $\boldsymbol{H}(\mathbb{X}_1^{(j)}, \boldsymbol{\theta})$ is at least as good as the one of $\{\mathbb{X}_1^{(j)}\}$ by definition. Because the time series in backward time is also mixing with the same rate, Assumption 10 (b) follow from the invariance principle in backward time (see also Remark 7). □

The remaining assumptions all correspond to well known results in statistics if a global estimator based on estimating functions is used. However, here, it is maximized over an increasing number of windows. To this end, we require versions of uniform laws of large numbers taking these moving windows into account as given in the following lemma.

**Lemma 2.** *Let $\{\mathbb{Y}_t\}$ be p-dimensional random vectors fulfilling Regularity Condition 1 (b) with $\tilde{\nu}$ as below (in (a),(b),(d) and arbitrary in (c)), $\boldsymbol{\Theta} \subset \mathbb{R}^p$ be a compact parameter space and $\boldsymbol{F} = (F_1, \ldots, F_p)^T : (\mathbb{R}^p, \boldsymbol{\Theta}) \to \mathbb{R}^p$ measurable.*

*(a) If $0 < E\,\|\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\|^{2+\tilde{\nu}} < \infty$ for some $\tilde{\nu} > 0$ and some $\boldsymbol{\theta}$, then for the same $\boldsymbol{\theta}$*

$$\sup_{0 \leq k \leq n-G} \left\| \sum_{i=k+1}^{k+G} \left(\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\left(\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right)\right) \right\| = O_P\left(\sqrt{G \log(n/G)}\right),$$

*(b) If for some $\tilde{\nu} > 0$ it holds $0 < E\left\|\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right\|^{2+\tilde{\nu}} < \infty$ for all $\boldsymbol{\theta}$ as well as $E\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\nabla\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right\|^{2+\tilde{\nu}} < \infty$, then*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\max_{0\le k\le n-G}\frac{1}{G}\left\|\sum_{i=k+1}^{k+G}\left(\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\left(\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right)\right)\right\| = o_P(1).$$

*(c) If $E\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right\| < \infty$, then for any sequence $G \to \infty$ it holds*

$$(i) \quad \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\max_{1\le k\le G}\frac{1}{k}\left\|\sum_{i=G-k+1}^{G}\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\right\| = O_P(1),$$

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\max_{1\le k\le G}\frac{1}{k}\left\|\sum_{i=1}^{k}\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\right\| = O_P(1).$$

$$(ii) \quad \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\max_{1\le k\le G}\frac{1}{G}\left\|\sum_{i=1}^{k}(\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\left(\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right))\right\| = o_P(1),$$

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\max_{1\le k\le G}\frac{1}{G}\left\|\sum_{i=k}^{G}(\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\left(\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right))\right\| = o_P(1).$$

*(d) If $E\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right\|^{2+\tilde{\nu}} < \infty$, then*

$$\sup_{0\le k\le n-G}\sum_{i=k+1}^{k+G}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\right\| = O_P\left(G\right).$$

*Proof.* Analogously to Theorem 5 $\{\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\}$ fulfills an invariance principle from which assertion (a) follows by similar arguments as in the proof of Theorem 1 (see Reckrühm (2019), Theorem E.2.12 for details).

The proof technique for (b) is well known (and we only use a basic version thereof). Thus, we only sketch the proof. First note that by the compactness assumption on $\boldsymbol{\Theta}$ for each $\delta > 0$ there exist $M = M(\delta) \ge 1$ and $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M \in \boldsymbol{\Theta}$ such that for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ there is an $m = 1, \ldots, M$ with $\|\boldsymbol{\theta} - \boldsymbol{\xi}_m\| < \delta$. We get for any $\boldsymbol{\xi}, \boldsymbol{\theta}$

$$\begin{aligned}&\max_{0\le k\le n-G}\frac{1}{G}\sum_{i=k+1}^{k+G}\left\|\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - (\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}) - E\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}))\right\| \\ &\le \left(2E\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\nabla\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\right\| + o_P(1)\right)\|\boldsymbol{\theta} - \boldsymbol{\xi}\| = \|\boldsymbol{\theta} - \boldsymbol{\xi}\|\, O_P(1),\end{aligned}$$

where the last line follows from a first order Taylor expansion in addition to a moving law of large numbers as in (a) applied to the time series $\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\nabla\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\right\|\}$ (see also (d)).

For given $\eta_1, \eta_2 > 0$ we can now choose $\delta = \delta(\eta_1, \eta_2) > 0$ such that

$$P\Big( \sup_{\|\boldsymbol{\theta}-\boldsymbol{\xi}\|<\delta} \max_{0\le k\le n-G} \frac{1}{G} \sum_{i=k+1}^{k+G} \|\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) \\ - (\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}) - E\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}))\| \ge \eta_1 \Big) \le \eta_2$$

for all $n \ge n_0(\eta_1, \eta_2)$. For the (to $\delta$) corresponding $\boldsymbol{\xi_1}, \ldots, \boldsymbol{\xi_M}$ it holds by another application of (a)

$$P\left( \max_{m=1,\ldots,M} \max_{0\le k\le n-G} \frac{1}{G} \left\| \sum_{i=k+1}^{k+G} (\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}_m) - E\left(\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\xi}_m)\right)) \right\| \ge \eta_1 \right) \le \eta_2.$$

for all $n \ge n_1(\eta_1, \eta_2)$. Combining these arguments yields (b) on noting that for any $\delta$ and corresponding $\boldsymbol{\xi}_m$, $m = 1, \ldots, M$, it holds

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{0\le k\le n-G} \frac{1}{G} \left\| \sum_{i=k+1}^{k+G} (\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\left(\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\theta})\right)) \right\| \\ \le \max_{m=1,\ldots,M} \max_{0\le k\le n-G} \frac{1}{G} \left\| \sum_{i=k+1}^{k+G} (\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}_m) - E\left(\boldsymbol{F}(\mathbb{Y}_1, \boldsymbol{\xi}_m)\right)) \right\| \\ + \sup_{\|\boldsymbol{\theta}-\boldsymbol{\xi}\|<\delta} \max_{0\le k\le n-G} \frac{1}{G} \sum_{i=k+1}^{k+G} \|\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) - E\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta}) \\ - (\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}) - E\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\xi}))\|.$$

By Rao (1962), Theorem 6.5, a uniform (in $\boldsymbol{\theta}$) strong law of large numbers holds for $\{\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\}$ because stationarity and mixing implies ergodicity (both forward and backward). By the almost sure convergence standard arguments give the assertions in (c). The proof of (d) follows along the same lines as the proof of (a) but applied to the function $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{F}(\mathbb{Y}_i, \boldsymbol{\theta})\|$. The necessary centering is of the order $G$. □

*Proof of Theorem 6.* For $k_{j-1,n} < k \le k_{j,n} - G$ a Taylor expansion in $\widehat{\boldsymbol{\theta}}_{k+1,k+G}$ around $\boldsymbol{\theta}_j$ yields that there exists $\left\|\boldsymbol{\xi}_{k,n}^{(j)} - \boldsymbol{\theta}_j\right\| \le \left\|\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right\|$ such that

$$-\frac{1}{\sqrt{G}} \sum_{i=k+1}^{k+G} \boldsymbol{H}(\mathbb{X}_i^{(j)}, \boldsymbol{\theta_j}) \\ = \left( \frac{1}{G} \sum_{i=k+1}^{k+G} \nabla \boldsymbol{H}(\mathbb{X}_i^{(j)}, \boldsymbol{\xi}_{k,n}^{(j)}) \right)^T \sqrt{G} \left( \widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j \right)$$

$$= \left(o_P(1) + \boldsymbol{V}_{(j)}(\boldsymbol{\xi}_{k,n}^{(j)})\right) \sqrt{G} \left(\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right) \text{ uniformly in } k,$$

where the last line follows by Regularity Conditions 2 (a) in combination with Lemma 2 (b). By Lemma 2 (a), Regularity Condition 1 and the definition of $\boldsymbol{\theta}_j$ we get

$$\sup_{1\leq k\leq n-G} \frac{1}{\sqrt{G}} \left\| \sum_{i=k+1}^{k+G} \boldsymbol{H}(\mathbb{X}_i^{(j)}, \boldsymbol{\theta_j}) \right\| = O_P\left(\sqrt{\log(n/G)}\right).$$

In combination with Regularity Conditions 2(b) this yields

$$\max_{j=1,\ldots,q+1} \max_{k_{j-1,n}<k\leq k_{j,n}-G} \sqrt{G} \left\|\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right\| = O_P\left(\sqrt{\log(n/G)}\right). \tag{5}$$

In particular, this shows the validity of Assumption 6 (a).
Moreover, for each $l = 1, \ldots, p$ and $k_{j-1,n} < k \leq k_{j,n} - G$ a second order Taylor expansion yields the existence of $\left\|\boldsymbol{\xi}_{l,n,k}^{(j)} - \boldsymbol{\theta}_j\right\| \leq \left\|\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right\|$ with

$$\begin{aligned}
&- \sum_{i=k+1}^{k+G} H_l(\mathbb{X}_i^{(j)}, \boldsymbol{\theta}_j) \\
&= \left( \sum_{i=k+1}^{k+G} \left(\nabla H_l(\mathbb{X}_i^{(j)}, \boldsymbol{\theta}_j)\right)\right)^T \left(\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right) \\
&\quad + \frac{1}{2} \left(\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right)^T \left( \sum_{i=k+1}^{k+G} \nabla^2 H_l(\mathbb{X}_i^{(j)}, \boldsymbol{\xi}_{l,n,k}^{(j)})\right) \left(\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right) \\
&= \left( E\left(\nabla H_l(\mathbb{X}_1^{(j)}, \boldsymbol{\theta}_j)\right) + O_P\left(\sqrt{\frac{\log(n/G)}{G}}\right)\right)^T G\left(\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right) \\
&\quad + O_P\left(G\right) \|\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\|^2 \quad \text{uniformly in } k,
\end{aligned}$$

where the last line follows by an application of Lemma 2 both (a) and (d). Thus, an application of (5) yields uniformly in $k$

$$\begin{aligned}
&- \frac{1}{\sqrt{2G}} \sum_{i=k+1}^{k+G} H_l(\mathbb{X}_i^{(j)}, \boldsymbol{\theta}_j) \\
&= E\left(\nabla H_l(\mathbb{X}_1^{(j)}, \boldsymbol{\theta}_j)\right)^T \sqrt{\frac{G}{2}} \left(\widehat{\boldsymbol{\theta}}_{k+1,k+G} - \boldsymbol{\theta}_j\right) + o_P\left((\log n/G)^{-1/2}\right),
\end{aligned}$$

showing the validity of Assumption 2, concluding the proof of (a).

By the strong law of large numbers (and similar arguments as in the proof of Lemma 2 (c)) and by definition of $\boldsymbol{\theta}_j$ it holds

$$\max_{k_{j-1,n}-G\le k\le k_{j-1,n}-\varepsilon\, G}\left\|\frac{1}{G}\sum_{i=k+1}^{k+G}\boldsymbol{H}(\mathbb{X}_i,\boldsymbol{\theta}_j)-\frac{k_{j-1,n}-k}{G}E\boldsymbol{H}(\mathbb{X}_i^{(j-1)},\boldsymbol{\theta}_j)\right\|$$

$$=\max_{k_{j-1,n}-G\le k\le k_{j-1,n}-\varepsilon\, G}\left\|\frac{1}{G}\sum_{i=k+1}^{k_{j-1,n}}\left(\boldsymbol{H}(\mathbb{X}_i^{(j-1)},\boldsymbol{\theta}_j)-E\boldsymbol{H}(\mathbb{X}_i^{(j-1)},\boldsymbol{\theta}_j)\right)\right.$$

$$\left.+\frac{1}{G}\sum_{i=k_{j-1,n}+1}^{k+G}\boldsymbol{H}(\mathbb{X}_i^{(j)},\boldsymbol{\theta}_j)\right\|=o_P(1).$$

By the identifiable uniqueness of $\boldsymbol{\theta}_j$ it holds $E\boldsymbol{H}(\mathbb{X}_i^{(j-1)},\boldsymbol{\theta}_j)\neq 0$, such that

$$\sqrt{\frac{G}{\log(n/G)}}\min_{k_{j-1,n}-G\le k\le k_{j-1,n}-\varepsilon\, G}\left\|\frac{1}{G}\sum_{i=k+1}^{k+G}\boldsymbol{H}(\mathbb{X}_i,\boldsymbol{\theta}_j)\right\|$$

$$\geq\sqrt{\frac{G}{\log(n/G)}}\left(\varepsilon\left\|E\boldsymbol{H}(\mathbb{X}_i^{(j-1)},\boldsymbol{\theta}_j)\right\|+o_P(1)\right)\xrightarrow{P}\infty.$$

By Lemma 2 (c) it holds

$$\sup_{\boldsymbol{\theta}\in\Theta}\max_{k_{j-1,n}-G\le k\le k_{j-1,n}-\varepsilon\, G}\left\|\frac{1}{G}\sum_{i=k+1}^{k+G}\nabla\boldsymbol{H}(\mathbb{X}_i,\boldsymbol{\theta})\right\|_F=O_P(1),$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm.
Furthermore, a Taylor expansion of $\widehat{\boldsymbol{\theta}}_{k+1,k+G}$ around $\boldsymbol{\theta}_j$ yields for some $\boldsymbol{\xi}_{k,n}$

$$\sqrt{\frac{G}{\log(n/G)}}\left\|\frac{1}{G}\sum_{i=k+1}^{k+G}\boldsymbol{H}(\mathbb{X}_i,\boldsymbol{\theta}_j)\right\|$$

$$=\left\|\left(\frac{1}{G}\sum_{i=k+1}^{k+G}\nabla\boldsymbol{H}(\mathbb{X}_i,\boldsymbol{\xi}_{k,n})\right)^T\sqrt{\frac{G}{\log(n/G)}}\left(\widehat{\boldsymbol{\theta}}_{k+1,k+G}-\boldsymbol{\theta}_j\right)\right\|$$

$$\leq\left\|\frac{1}{G}\sum_{i=k+1}^{k+G}\nabla\boldsymbol{H}(\mathbb{X}_i,\boldsymbol{\xi}_{k,n})\right\|_F\sqrt{\frac{G}{\log(n/G)}}\|\widehat{\boldsymbol{\theta}}_{k+1,k+G}-\boldsymbol{\theta}_j\|,$$

where we used the consistency of the Frobenius matrix norm with the Euclidean vector norm in the last step. Thus, we have shown that the left hand side diverges to infinity stochastically, while the first term on the right hand side is stochastically bounded – both in an appropriate uniform sense. By standard arguments this shows that indeed

$$\sqrt{\frac{G}{\log(n/G)}}\min_{k_{j-1,n}-G\le k\le k_{j-1,n}-\varepsilon\, G}\|\widehat{\boldsymbol{\theta}}_{k+1,k+G}-\boldsymbol{\theta}_j\|\xrightarrow{P}\infty.$$

The assertion for $k_{j,n}-(1-\varepsilon)\,G\leq k\leq k_{j,n}$ follows analogously concluding the proof. □

*Proof of Theorem 7.* The proof is analogous to the proof of (5) where the sum $-\frac{1}{\sqrt{n}}\sum_{i=a}^{b}\boldsymbol{H}(\mathbb{X}_i,\widetilde{\boldsymbol{\theta}}_{\gamma_a,\gamma_b})$ is considered instead. The better rate compared to (5) is due to the fact that a piecewise application of the central limit theorem (which follows from the mixing condition) to each regime yields

$$\frac{1}{\sqrt{n}}\left\|\sum_{i=a}^{b}\boldsymbol{H}(\mathbb{X}_i,\widetilde{\boldsymbol{\theta}}_{\gamma_a,\gamma_b})\right\| = O_P\left(1\right).$$

The convergence of $\frac{1}{n}\sum_{i=a}^{b}\nabla\boldsymbol{H}(\mathbb{X}_i,\boldsymbol{\xi}_n^{(\gamma_a,\gamma_b)})$ under these regularity conditions follows also e.g. from a piecewise application of Lemma 2 (c)(ii) (with $G$ replaced by $n$). □

*Proof of Theorem 8.* A first order Taylor expansion yields

$$\begin{aligned}&\sum_{i=k+1}^{k+G}\boldsymbol{H}(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}}_n)-\sum_{i=k+1}^{k+G}\boldsymbol{H}(\mathbb{X}_i^{(j)},\widetilde{\boldsymbol{\theta}})\\&=\left(\sum_{i=k+1}^{k+G}\nabla\boldsymbol{H}(\mathbb{X}_i^{(j)},\boldsymbol{\xi}_{k,n}^{(j)})\right)^T\left(\widetilde{\boldsymbol{\theta}}_n-\widetilde{\boldsymbol{\theta}}\right)=O_P(G/\sqrt{n})=o_P\left(\sqrt{\frac{G}{\log(n/G)}}\right),\end{aligned}$$

where the last line follows uniformly in $k$ by Lemma 2 (b). This shows the validity of Assumption 4 (i). The proof of (ii) and Assumption 10 (a) are analogous by using Lemma 2(c) instead. □

## References

Eichinger, B., & Kirch, C. (2018). A mosum procedure for the estimation of multiple random change points. *Bernoulli*, *24*(1), 526–564.

Kuelbs, J., & Philipp, W. (1980). Almost sure invariance principles for partial sums of mixing b-valued random variables. *The Annals of Probability*, *8*(6), 1003–1036.

Liboschik, T., Fokianos, K., & Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, *82*(5), 1–51. doi: 10.18637/jss.v082.i05

Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics*, *33*(2), 659–680.

Reckrühm, K. (2019). *Estimating multiple structural breaks in time series: A generalized mosum approach based on estimating functions* (Doctoral dissertation, Otto-von-Guericke-Universität Magdeburg, Faculty of Mathematics). Retrieved from `http://dx.doi.org/10.25673/13832`

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

Weiß, C. H. (2010). The inarch (1) model for overdispersed time series of counts. *Communications in Statistics-Simulation and Computation*, *39*(6), 1269–1291.

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 infrastructure for regular and irregular time series. *Journal of Statistical Software*, *14*(6), 1–27.