Handbook of Discrete-Valued Time Series
Handbooks of Modern Statistical Methods
Richard A. Davis, Scott H. Holan, Robert Lund, Nalini Ravishanker

Detection of Change Points in Discrete-Valued Time Series

Publication details
Claudia Kirch, Joseph Tadjuidje Kamgaing
Published online on: 21 Dec 2015

Accessed on: 05 Dec 2023

PLEASE SCROLL DOWN FOR DOCUMENT

Full terms and conditions of use: https://www.routledgehandbooks.com/legal-notices/terms

This Document PDF may be used for research, teaching and private study purposes. Any substantial or systematic reproductions, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The publisher shall not be liable for an loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
10

Detection of Change Points in Discrete-Valued Time Series

Claudia Kirch and Joseph Tadjuidje Kamgaing

CONTENTS

10.1 Introduction ................................................................................... 219
10.2 General Principles of Retrospective Change Point Analysis ................. 220
10.3 Detection of Changes in Binary Models .......................................... 225
10.4 Detection of Changes in Poisson Autoregressive Models ...................... 227
10.5 Simulation and Data Analysis .......................................................... 230
   10.5.1 Binary Autoregressive Time Series ......................................... 230
       10.5.1.1 Data Analysis: U.S. Recession Data .................................. 231
   10.5.2 Poisson Autoregressive Models .............................................. 232
       10.5.2.1 Data Analysis: Number of Transactions per Minute
             for Ericsson B Stock ......................................................... 233
10.6 Online Procedures ......................................................................... 235
10.7 Technical Appendix ........................................................................ 236
   10.7.1 Regularity Assumptions ............................................................. 236
   10.7.2 Proofs .................................................................................. 239
References .............................................................................................. 243

10.1 Introduction

There has recently been a renewed interest in statistical procedures concerned with the detection of structural breaks in time series, for example, the recent review articles by Aue and Horváth [2] and Horváth and Rice [16]. The literature contains statistics to detect simple mean changes, changes in linear regression, changes in generalized autoregressive conditionally heteroscedastic (GARCH) models; from likelihood ratio to robust M methods (see, e.g., Berkes et al. [3], Davis et al. [6], Hušková and Marušiaková [26], and Robbins et al. [31]). While at first sight, the corresponding statistics appear very different, most of them are derived using the same principles. In this chapter, we shed light on those principles, explaining how corresponding statistics and their respective asymptotic behavior under both the null and alternative hypotheses can be derived. This enables us to give a unified presentation of change point procedures for integer-valued time series. Because the methodology considered in this chapter is by no means limited to these situations, it allows for future extensions in a standardized way.
Hudecová [17] and Fokianos et al. [11] propose change point statistics for binary time series models while Franke et al. [13] and Doukhan and Kegne [7] consider changes in Poisson autoregressive models. Related procedures have also been investigated by Fokianos and Fried [9,10] for integer valued GARCH and log-linear Poisson autoregressive time series, respectively, but with a focus on outlier detection and intervention effects rather than change points.

Section 10.2 explains how change point statistics can be constructed and derives asymptotic properties under both the null and alternative hypotheses, based on regularity conditions, which are summarized in Appendix 10.7.1 to lighten the presentation. This methodology is then applied to binary time series in Section 10.3 and to Poisson autoregressive models in Section 10.4, generalizing the statistics already discussed in the literature. In Section 10.5, some simulations as well as applications to real data illustrate the performance of these procedures. A short review of sequential (also called online) procedures for count time series is given in Section 10.6. Finally, the proofs are given in Appendix 10.7.2.

### 10.2 General Principles of Retrospective Change Point Analysis

Assume that data $Y_1, \ldots, Y_n$ are observed with a possible structural break at the (unknown) change point $k_0$. We will first look at likelihood ratio tests for structural breaks before explaining how to generalize these ideas. To this end, we assume that the data before and after the change can be parameterized by the same likelihood function $L$ but with different (unknown) parameters $\theta_0, \theta_1 \in \Theta \subset \mathbb{R}^d$. A likelihood ratio approach yields the following statistic:

$$
\max_{1 \leq k \leq n} \ell(k) := \max_{1 \leq k \leq n} \left( \ell \left( (Y_1, \ldots, Y_k), \hat{\theta}_k \right) + \ell \left( (Y_{k+1}, \ldots, Y_n), \hat{\theta}_k^0 \right) - \ell \left( (Y_1, \ldots, Y_n), \hat{\theta}_n \right) \right),
$$

where $\ell(Y, \theta)$ is the log-likelihood function and $\hat{\theta}_k$ and $\hat{\theta}_k^0$ are the maximum likelihood estimator based on $Y_1, \ldots, Y_k$ and $Y_{k+1}, \ldots, Y_n$, respectively. The maximum over $k$ is due to the fact that the change point is unknown, so the likelihood ratio statistic maximizes over all possible change points. A similar approach based on some asymptotic Bayes statistic leads to a sum-type statistic, where the sum over $\ell(k)$ is considered (see, e.g., Kirch [19]).

Davis et al. [6] proposed this statistic for linear autoregressive processes of order $p$ with standard normal errors:

$$
Y_t = \beta_0 + \sum_{j=1}^{p} \beta_j Y_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim i.i.d. N(0, 1). \tag{10.1}
$$

In this situation (which includes mean changes as a special case ($p = 0$)), this maximum likelihood statistic does not converge in distribution to a nondegenerate limit but almost surely to infinity (Davis et al. [6]). Nevertheless, asymptotic level $\alpha$ tests based on this maximum likelihood statistic can be constructed using a Darling–Erdös limit theorem as stated in Theorem 10.1b. In small samples, however, the slow convergence of Darling–Erdös limit theorems often leads to some size distortions.
Similarly, one can construct Wald-type statistics based on maxima or sums of quadratic forms of
\[ W(k) := \hat{\theta}_k - \hat{\theta}_k^c, \quad k = 1, \ldots, n. \]

Wald statistics can be generalized to any other estimation procedure for \( \theta \) and are not restricted to maximum likelihood estimators. However, for both maximum likelihood and Wald statistics, the estimators \( \hat{\theta}_k \) and \( \hat{\theta}_k^c \) need to be calculated, which can be problematic in nonlinear situations. In such situations, which are typical for integer-valued time series, these estimators are usually not analytically tractable, but need to be calculated using numerical optimization methods. This can lead to additional computational effort to calculate the statistics or large numerical errors. The latter problems can be reduced by using score-type statistics based on maxima or sums of quadratic forms of
\[ S(k) := S(k, \hat{\theta}_n) = \frac{\partial}{\partial \theta} \ell ((Y_1, \ldots, Y_k), \theta) |_{\theta = \hat{\theta}_n}, \quad k = 1, \ldots, n. \]

In this case, only the estimator based on the full data set \( Y_1, \ldots, Y_n \) needs to be calculated (possibly using numerical methods). The likelihood score statistic for the linear regression model has been investigated in detail by Hušková et al. [27]. Similarly to Wald statistics, score statistics do not need to be likelihood based but can be generalized to different estimators as long as those estimators can be obtained as a solution to

\[ C_k = \sum_{t=1}^{k} Y_{t-1} Y_{t-1}^T, \quad C_k^c = \sum_{t=k+1}^{n} Y_{t-1} Y_{t-1}^T, \quad Y_{t-1} = (1, Y_{t-1}, \ldots, Y_{t-p})^T. \]

As already mentioned, the maximum likelihood statistic (hence the corresponding likelihood Wald and score statistics) does not converge in distribution. Under the null hypothesis, the matrix in the quadratic form of the likelihood score statistic can be approximated asymptotically (as \( k \to \infty \), \( n - k \to \infty \), \( n \to \infty \)) by

\[ C_k^{-1} C_n (C_k^c)^{-1} = \frac{1}{k} \frac{1}{n} \frac{1}{n} C^{-1} + o_P(1), \quad C = EY_{t-1} Y_{t-1}^T. \]
Replacing this term by \( w(k/n)C_n^{-1} \) for a suitable weight function \( w(\cdot) \) leads to a statistic that does converge in distribution. More precisely, we consider

\[
\max_{1 \leq k \leq n} \frac{k}{n} S(k)^T C_n^{-1} S(k),
\]

where \( w : [0, 1] \rightarrow \mathbb{R}_+ \) is a nonnegative continuous weight function fulfilling

\[
\lim_{t \to 0} t^\alpha w(t) < \infty, \quad \lim_{t \to 1} (1-t)^\alpha w(t) < \infty, \quad \text{for some } 0 \leq \alpha < 1
\]

\[
\sup_{\eta \leq t \leq 1-\eta} w(t) < \infty \quad \text{for all } 0 < \eta \leq \frac{1}{2}.
\]

Theorem 10.1a shows that this class of statistics converges, under regularity conditions, in distribution to a nondegenerate limit. The following choice of weight function, closely related to the choice of the weights in (10.2), has often been proposed in the literature:

\[
w(t) = (t(1-t))^{-\gamma}, \quad 0 \leq \gamma < 1,
\]

where \( \gamma \) close to 1 detects early or late changes with better power. In the econometrics literature, the following weight functions are often used, which correspond to a truncation of the likelihood ratio statistic and can be viewed as the likelihood ratio statistic under restrictions on the set of admissible change points,

\[
w(t) = (t(1-t))^{-1/2} 1_{\{\epsilon \leq t \leq 1-\epsilon\}}
\]

for some \( \epsilon > 0 \). Similarly, if a priori knowledge of the location of the change point is available, one can increase the power of the designed test statistic for such alternatives by choosing a weight function that is larger near these points (Kirch et al. [20]). Nevertheless, these statistics have asymptotic power one for other change locations (See Theorem 10.2).

Additionally, many change point statistics discussed in the literature do not use the full score function but rather a lower-dimensional projection, where \( C_n \) is replaced by a lower rank matrix. For linear autoregressive models as in (10.1), for example, Kulperger [24] and Horváth [25] use a partial sum process based on estimated residuals, which corresponds to the first component of the likelihood score vector in this example.

For this reason, in the following, we do not require \( S(k, \theta) \) to be the likelihood score (nor even of the same dimension as \( \theta \)), nor do we assume that \( \hat{\theta}_n \) is the maximum likelihood estimator. In fact, we allow for general score-type statistics that are based on partial sum processes of the type

\[
S(k, \hat{\theta}_n) = \sum_{j=1}^k H(X_j, \hat{\theta}_n), \quad \text{with } S(n, \hat{\theta}_n) = 0 \quad \text{and} \quad \hat{\theta}_n \to \theta_0,
\]

where \( \theta_0 \) is typically the correct parameter, \( X_j \) are observations, where, for example, for the autoregressive case of order one, a vector \( X_j = (X_j, X_{j-1})^T \) is used, and \( H \) is some function usually of the type \( AF \) for some (possibly lower rank) matrix \( A \) and an estimating function.
\[ F \text{ that defines the estimator } \hat{\theta}_n \text{ as the unique zero of } \sum_{j=1}^{n} F(X_j, \hat{\theta}_n) = 0. \] Furthermore, it is possible to allow for misspecification, in which case, \( \theta_0 \) becomes the best approximating parameter in the sense of \( EF(X_j, \theta_0) = 0 \). More details on this framework in a sequential context can be found in Kirch and Kamgaing [22].

We are now able to derive the limit distribution of the corresponding score-type change point tests under the null hypothesis under the regularity conditions given in Section 10.7.1. These regularity conditions are implicitly shown in the proofs for change point tests of the types mentioned earlier. Examples for integer-valued time series are given in Sections 10.3 and 10.4.

**Theorem 10.1** We obtain the following null asymptotics:

(a) Let A.1 and A.2 (i) in Section 10.7.1 hold. Assume that the weight function is either a continuous nonnegative and bounded function \( w : [0, 1] \to \mathbb{R}_+ \), or for unbounded functions fulfilling (10.3), let additionally A.2 (ii) in Section 10.7.1 hold. Then:

\[ \begin{align*}
\text{(i) } & \max_{1 \leq k \leq n} \frac{w(k/n)}{n} \mathbb{S}(k, \hat{\theta}_n)^T \Sigma^{-1} \mathbb{S}(k, \hat{\theta}_n) \overset{D}{\longrightarrow} \sup_{0 \leq t \leq 1} w(t) \sum_{j=1}^{d} B_j^2(t), \\
\text{(ii) } & \sum_{1 \leq k \leq n} \frac{w(k/n)}{n} \mathbb{S}(k, \hat{\theta}_n)^T \Sigma^{-1} \mathbb{S}(k, \hat{\theta}_n) \overset{D}{\longrightarrow} \int_0^1 w(t) \sum_{j=1}^{d} B_j^2(t) \, dt,
\end{align*} \]

where \( B_j(\cdot), j = 1, \ldots, d \), are independent Brownian bridges and \( \Sigma \) can be replaced by \( \hat{\Sigma}_n \) if \( \hat{\Sigma}_n - \Sigma = o_P(1) \).

(b) Under A.1 and A.3 in Section 10.7.1 it holds

\[ P \left( a(\log n) \max_{1 \leq k \leq n} \sqrt{\frac{n}{k(n-k)} \mathbb{S}(k, \hat{\theta}_n)^T \Sigma^{-1} \mathbb{S}(k, \hat{\theta}_n) - b_d(\log n) } \leq t \right) \to \exp(-2e^{-t}), \]

where \( a(x) = \sqrt{\Gamma(d)} \log x, b_d(x) = 2 \log x + \frac{d}{2} \log \log x - \log \Gamma(d/2), \Gamma(\cdot) \text{ is the Gamma-function, and } d \text{ is the dimension of the vector } \mathbb{S}(k, \theta). \) Furthermore, \( \Sigma \) can be replaced by an estimator \( \hat{\Sigma}_n \) if \( \| \hat{\Sigma}_n^{-1/2} - \Sigma^{-1/2} \| = o_P((\log \log n)^{-1}). \)

The assumption of continuity of the weight function in (b) can be relaxed to allow for a finite number of points of discontinuity, where \( w \) is either left or right continuous with existing limits from the other side.

Similarly, under alternatives, we provide some regularity conditions, which ensure that the tests mentioned earlier have asymptotic power one. Additionally, we propose a consistent estimator of the change point in rescaled time.

**Theorem 10.2** Under alternatives with a change point of the form

\[ k_0 = \lfloor \lambda n \rfloor, \quad 0 < \lambda < 1, \quad (10.5) \]
we get the following assertions:

(a) If Assumptions B.1, B.3, and B.4(i) in Section 10.7.1 hold, then the Darling–Erdős- and max-type statistics for continuous weight functions fulfilling $w(\lambda) > 0$ have asymptotic power one, that is, for all $x \in \mathbb{R}$ it holds that

(i) $P\left( \max_{1 \leq k \leq n} \frac{w(k/n)}{n} S(k, \hat{\theta}_n) \Sigma^{-1} S(k, \hat{\theta}_n)^T \right) \to 1$,

(ii) $P\left( a(\log n) \max_{1 \leq k \leq n} \sqrt{S(k, \hat{\theta}_n)^T \Sigma^{-1} S(k, \hat{\theta}_n)} - b_d(\log n) \right) \to 1$.

If B.4(i) is replaced by B.4(ii), then the assertion remains true if $\Sigma$ is replaced by $\hat{\Sigma}_n$, a consistent estimator of $\Sigma$.

(b) If additionally B.5 holds, then the sum-type statistics for a continuous weight function $w(\cdot) \neq 0$ fulfilling (10.3) has power one, that is, it holds for all $x \in \mathbb{R}$

$$P\left( \sum_{1 \leq k \leq n} \frac{w(k/n)}{n^2} S(k, \hat{\theta}_n)^T \Sigma^{-1} S(k, \hat{\theta}_n) \right) \to 1.$$

If B.4(i) is replaced by B.4(ii), then the assertion remains true if $\Sigma$ is replaced by $\hat{\Sigma}_n$, a consistent estimator of $\Sigma$.

(c) Let the change point be of the form $k_0 = \lfloor \lambda n \rfloor$, $0 < \lambda < 1$, and consider

$$\hat{\lambda}_n = \arg \max_{k \leq n} \frac{S(k, \hat{\theta}_n)^T \Sigma^{-1} S(k, \hat{\theta}_n)}{n}.$$

Under Assumptions B.1, B.3, and B.4(i), $\hat{\lambda}_n$ is a consistent estimator for the change point in rescaled time $\lambda$, that is,

$$\hat{\lambda}_n - \lambda = o_P(1).$$

If B.4(i) is replaced by B.4(iii), then the assertion remains true if $\Sigma$ is replaced by $\hat{\Sigma}_n$, a consistent estimator of $\Sigma$.

Assumption (10.5) is standard in change point analysis but can be weakened for Part (a) of Theorem 10.2. The continuity assumption on the weight function can be relaxed in this situation.

While these test statistics were designed for the situation, where at most one change is expected, they usually also have power against multiple changes. This fact is the underlying principle of binary segmentation procedures (first proposed by Vostrikova [36]), which works as follows: The data set is tested using an at most one change test as given earlier. If that test is significant, the data set is split at the estimated change point and the procedure repeated on both data segments until insignificant. Recently, a randomized version of binary segmentation has been proposed for the simple mean change problem (Fryzlewicz [14]).
The optimal rate of convergence in (b) is usually given by $\hat{\lambda}_n - \lambda = O_P(1/n)$ but requires a much more involved proof (Csörgö and Horváth [5], Theorem 2.8.1, for a proof in a mean change model).

### 10.3 Detection of Changes in Binary Models

Binary time series are important in applications, where one is observing whether a certain event has or has not occurred. Wilks and Wilby [40], for example, observe whether it has been raining on a specific day, and Kauppi and Saikkonen [18] and Startz [34] observe whether a recession has occurred or not in a given month. A common binary time series model is given by

$$Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Z_{t-1}, Z_{t-2}, \ldots \sim \text{Bern}(\pi_t(\beta)),$$

with $g(\pi_t(\beta)) = \beta^T Z_{t-1}$, where $Z_{t-1} = (Z_{t-1}(1), \ldots, Z_{t-1}(p))^T$ is a regressor, which can be purely exogenous (i.e., $\{Z_t\}$ is independent of $\{Y_t\}$), purely autoregressive (i.e., $Z_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})$) or a mixture of both (in particular, the independence assumption does not need to hold). Similar to generalized linear models, the canonical link function $g(x) = \log(x/(1 - x))$ is used and statistical inference is based on the partial likelihood

$$L(\beta) = \prod_{t=1}^{n} \pi_t(\beta)^{y_t}(1 - \pi_t(\beta))^{1-y_t},$$

with corresponding score vector

$$S^{\text{BAR}}(k, \beta_n) = \sum_{t=1}^{k} Z_{t-1}(Y_t - \pi_t(\beta))$$

for the canonical link function.

**Theorem 10.3** We get the following assertions under the null hypothesis:

(a) Let the covariate process $\{Z_t\}$ be strictly stationary and ergodic with finite fourth moments. Then, under the null hypothesis, A.1 and A.3 (i) in Section 10.7.1 are fulfilled for the partial sum process $S^{\text{BAR}}(k, \beta_n)$ and $\hat{\beta}_n$ defined by

$$S^{\text{BAR}}(n, \hat{\beta}_n) = 0.$$  

(b) If $(Y_t, Z_{t-1}, \ldots, Z_{t-p})^T$ is also $\alpha$-mixing with exponential rate, then A.3 (ii) and (iii) in Section 10.7.1 are fulfilled.
In particular, change point statistics based on $S^{\text{BAR}}(k, \hat{\beta}_n)$ have the null asymptotics as stated in Theorem 10.1 with $\Sigma = \text{cov} (Z_{t-1} (Y_t - \pi_t (\beta_0)))$, where $\beta_0$ is the true parameter under the null hypothesis.

**Remark 10.1** For $Z_{t-1} = (Y_{t-1}, \ldots, Y_{t-1})^T$, $Y_t$ is the standard binary autoregressive model (BAR($p$)), for which the assumptions of Theorem 10.3 (b) are fulfilled, see, for example, Wang and Li [37]. However, considering some regularity assumptions on the exogenous process, one can prove that $(Y_t, \ldots, Y_{t-p-1}, Z_t, \ldots, Z_{t-q})$ is a Feller chain, for which Theorem 1 of Feigin and Tweedie [8] implies geometric ergodicity (see Kirch and Tadjuidje Kamgaing [21] for details) implying that it is $\beta$-mixing with exponential rate.

**Remark 10.2** The mixing concept can be regarded as an asymptotic measure of independence in time between the observations. The reader can refer to Tadjuidje et al. [35], Remark 4, for $\alpha$- and $\beta$-mixing definitions, as well as a concise summary of their relationship to geometric ergodicity for a Markov chain.

Instead of using the full vector partial sum process $S^{\text{BAR}}(k, \hat{\beta}_n)$ to construct the test statistics, often lower-dimensional linear combinations are used, such as

$$\tilde{S}^{\text{BAR}}(k, \hat{\beta}_n) = \sum_{t=1}^{k} (Y_t - \pi_t (\hat{\beta}_n)),$$

(10.8)

where $\hat{\beta}_n$ is defined by (10.7). If the assumptions of Theorem 10.3 are fulfilled, then the null asymptotics as in Theorem 10.1 with $\Sigma = \text{cov}(Y_t - \pi_t (\beta_0))$ hold.

The statistic based on $S^{\text{BAR}}(k, \hat{\beta}_n)$ for $w \equiv 1$ has been proposed by Fokianos et al. [11]; a statistic based on $\tilde{S}^{\text{BAR}}(k, \hat{\beta}_n)$ with a somewhat different standardization in a purely autoregressive setup has been considered by Hudecová [17]. Hudecová’s statistic is the score statistic based on the partial likelihood and the restricted alternative of a change only in the intercept.

**BAR-Alternative:** Let the following assumptions hold:

**H1(i)** The change point is of the form $k_0 = \lfloor \lambda n \rfloor$, $0 < \lambda < 1$.

**H1(ii)** The binary time series $\{Y_t\}$ and the covariate process $\{Z_t\}$ before the change fulfills the assumptions of Theorem 10.3a.

**H1(iii)** The time series after the change as well as the covariate process after the change can be written as $Y_t = \tilde{Y}_t + R_1(t)$ and $Z_t = \tilde{Z}_t + R_2(t)$, respectively, $t > \lfloor \lambda n \rfloor$, where $\{\tilde{Y}_t\}$ is bounded, stationary, and ergodic and $\{\tilde{Z}_t\}$ is square integrable as well as stationary and ergodic with remainder terms fulfilling

$$\frac{1}{n} \sum_{j=\lfloor \lambda n \rfloor + 1}^{n} R_1^2(t) = o_P(1), \quad \frac{1}{n} \sum_{j=\lfloor \lambda n \rfloor + 1}^{n} \|R_2(t)\|^2 = o_P(1).$$

**H1(iv)** $\lambda E[1(Y_t - \pi_1(\beta)) + (1 - \lambda)E\tilde{Z}_{n-1}(\tilde{Y}_n - \pi_1(\beta))]$ has a unique zero $\beta_1 \in \Theta$ and $\Theta$ is compact and convex with $\hat{\beta}_n \in \Theta$. 

© 2016 by Taylor & Francis Group, LLC
Assumption (i) is standard in change point analysis but could be relaxed, and assumption (ii) states that the time series before the change fulfills the assumption under the null hypothesis. Assumption (iii) allows for rather general alternatives, including situations where starting values from before the change are used resulting in a nonstationary time series after the change. Assumption (iv) guarantees that the estimator $\hat{\beta}_n$ converges to $\beta_1$. Neither Hudecová [17] nor Fokianos et al. [11] have derived the behavior of their statistics under alternatives.

**Theorem 10.4** Let $H_1(i)$–$H_1(iv)$ hold.

(a) For $S_{BAR}(k, \beta)$ as in (10.6), B.1 and B.2 are fulfilled, which implies B.3. If $k_0 = \lfloor \lambda n \rfloor$, then B.5 is fulfilled with $F_\lambda(\beta) = \lambda \mathbb{E}Z_0(Y_1 - \pi_1(\beta))$.

(b) For $\tilde{S}_{BAR}(k, \beta)$ as in (10.8) and if $k_0 = \lfloor \lambda n \rfloor$, then B.5 is fulfilled with $F_\lambda(\beta) = \lambda \mathbb{E}(Y_1 - \pi_1(\beta))$.

B.4 is fulfilled for the full score statistic from Theorem 10.4a if the time series before and after the change are correctly specified binary time series models with different parameters. Otherwise, restrictions apply. Together with Theorem 10.2, this implies that the corresponding change point statistics have power one and the point where the maximum is obtained is a consistent estimator for the change point in rescaled time.

### 10.4 Detection of Changes in Poisson Autoregressive Models

Another popular model for time series of counts is the Poisson autoregression, where we observe $Y_1, \ldots, Y_n$ with

$$Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p} \sim \text{Pois}(\lambda_t), \quad \lambda_t = f_\gamma(Y_{t-1}, \ldots, Y_{t-p}) \tag{10.9}$$

for some $d$-dimensional parameter vector $\gamma \in \Theta$. If $f_\gamma(x)$ is Lipschitz continuous in $x$ for all $\gamma \in \Theta$ with Lipschitz constant strictly smaller than 1, then there exists a stationary ergodic solution of (10.9) that is $\beta$-mixing with exponential rate (Neumann [28]). For a given parametric model $f_0$, this allows us to consider score-type change point statistics based on likelihood equations using the tools of Section 10.2. The mixing condition in connection to some moment conditions typically allows one to derive A.3, while a Taylor expansion in connection with $\sqrt{n}$-consistency of the corresponding maximum likelihood estimator (e.g., derived by Doukhan and Kegne [7], Theorem 3.2) gives A.1 under some additional moment conditions. Related test statistics for independent Poisson data are discussed in Robbins et al. [32]. However, in this chapter, we will concentrate on change point statistics related to those proposed by Franke et al. [13], which are based on least square scores and, as such, do not make use of the Poisson structure of the process. However, the methods described in Section 10.2 can be used to derive change point tests based on the partial likelihood, which can be expected to have higher power if the model is correctly specified.
Consider the least squares estimator \( \hat{\gamma}_n \) defined by

\[
S_{\text{SPAR}}(n, \hat{\gamma}_n) = 0, \quad \text{where } S_{\text{SPAR}}(k, \gamma_n) = \sum_{t=1}^{k} \nabla f_{\gamma}((Y_{t-1}, \ldots, Y_{t-p}))(Y_t - f_{\gamma}(Y_{t-1}, \ldots, Y_{t-p})),
\]

(10.10)

where \( \nabla \) denotes the gradient with respect to \( \gamma \). Under the additional assumption

\[
f_{\gamma}(x) = \gamma_1 + f_{\gamma_2, \ldots, \gamma_d}(x),
\]

that is, if \( \gamma_1 \) is an additive constant in the regression function, this implies in particular that

\[
\sum_{t=1}^{n} (Y_t - f_{\gamma_1}(Y_{t-1}, \ldots, Y_{t-p})) = 0. \quad (10.11)
\]

Assumptions under \( H_0 \):

1. \( \{Y_t\} \) is stationary and \( \alpha \)-mixing with exponential rate such that \( \mathbb{E} \sup_{\gamma \in \Theta} f_{\gamma}^2(Y_t) < \infty \).
2. \( f_{\gamma}(x) = \gamma_1 + f_{\gamma_2, \ldots, \gamma_d}(x) \) is twice continuously differentiable with respect to \( \gamma \) for all \( x \in \mathbb{N}_0^p \) and

\[
\mathbb{E} \sup_{\gamma \in \Theta} \| \nabla f_{\gamma}(Y_t) \nabla^T f_{\gamma}(Y_t) \| < \infty, \quad \mathbb{E} \left( Y_t \sup_{\gamma \in \Theta} \| \nabla^2 f_{\gamma}(Y_{t-1}) \| \right) < \infty.
\]

3. \( e(\gamma) = \mathbb{E}(Y_t - f_{\gamma}(Y_{t-1}))^2 \) has a unique minimizer \( \gamma_0 \) in the interior of some compact set \( \Theta \) such that the Hessian of \( e(\gamma_0) \) is positive definite.

As already mentioned, (i) is fulfilled for a large class of Poisson autoregressive processes under mild conditions. Assumption (ii) means that the autoregressive function used to construct the test statistic is linear in the first component guaranteeing (10.11). Note that this assumption does not need to be fulfilled for the true regression function of \( \{Y_t\} \) (in fact, \( Y_t \) does not even need to be a Poisson autoregressive time series). Assumption (iii) is fulfilled for the true value if \( \{Y_t\} \) is a Poisson autoregressive time series with regression function \( f_{\gamma} \).

**Theorem 10.5** Let under the null hypothesis \( H_0(\text{ii})-(\text{iii}) \) be fulfilled.

(a) If \( \mathbb{E}\|\nabla f_{\gamma_0}(Y_p, \ldots, Y_1)\|^{\nu} < \infty \) for some \( \nu > 2 \), then \( S_{\text{SPAR}}(k, \gamma) = \sum_{j=1}^{k} (Y_t - f_{\gamma_1}(Y_{t-1}, \ldots, Y_{t-p})) \) together with \( \hat{\gamma}_n \) as in (10.10), A.1 and A.3 in Section 10.7.1 hold with \( \Sigma = (\sigma^2) \) the long-run variance of \( Y_t - f_{\gamma_0}(Y_{t-1}, \ldots, Y_{t-p}) \).

(b) If \( f_{\gamma}(x) = \gamma_1 + (\gamma_2, \ldots, \gamma_d)^T x \) \( (p = d - 1) \) and \( \mathbb{E}|Y_0|^\nu < \infty \) for some \( \nu > 4 \), then \( S_{\text{SPAR}}(k, \gamma) = \sum_{j=1}^{k} (Y_{t-1} - \hat{\gamma}_n^T \hat{\gamma}_n) \) where \( \gamma_{t-1} = (1, Y_{t-1}, \ldots, Y_{t-d+1})^T \). Together with \( \hat{\gamma}_n \) as in (10.10), A.1 and A.3 in Section 10.7.1 hold, where \( \Sigma \) is the long-run covariance matrix of \( \{Y_{t-1}(Y_t - \gamma_0^T \hat{\gamma}_n)\} \).

In particular, the change point statistics based on \( \tilde{S}_{\text{SPAR}}(k, \hat{\gamma}_n) \) in (a) and \( S_{\text{SPAR}}(k, \hat{\gamma}_n) \) in (b) have the null asymptotics as stated in Theorem 10.1.
Assumptions under $H_1$: 

$H_1(i)$ The change point is of the form $k_0 = \lfloor \lambda n \rfloor$, $0 < \lambda < 1$.

$H_1(ii)$ For all $\gamma \in \Theta$, $\Theta$ is compact and convex, and $f_\gamma(x)$ is uniformly Lipschitz in $x$ with Lipschitz constant $L_\gamma < 1$.

$H_1(iii)$ The time series before the change is stationary and ergodic such that $\mathbb{E} \sup_{\gamma \in \Theta} \| Y_{t-j} \nabla f_\gamma(Y_{t-1}, \ldots, Y_{t-p}) \| < \infty$, $j = 0, \ldots, p$.

$H_1(iv)$ The time series after the change fulfills $Y_t = \tilde{Y}_t + R_1(t), t > \lfloor \lambda n \rfloor$, where $\{\tilde{Y}_t\}$ is stationary and ergodic such that $\mathbb{E} \sup_{\gamma \in \Theta} \| \tilde{Y}_{t-j} \nabla f_\gamma(\tilde{Y}_{t-1}, \ldots, \tilde{Y}_{t-p}) \| < \infty$, $j = 0, \ldots, p$, with the remainder term fulfilling

$$\frac{1}{n} \sum_{j=\lfloor \lambda n \rfloor + 1}^{n} R_1^2(t) = o_p(1).$$

$H_1(v)$ $\lambda \mathbb{E} \nabla f_\gamma((Y_0, \ldots, Y_{1-p}))(Y_1 - f_\gamma((Y_0, \ldots, Y_{1-p}))) + (1 - \lambda) \mathbb{E} \nabla f_\gamma(\tilde{Y}_0, \ldots, \tilde{Y}_{1-p}) (\tilde{Y}_1 - f_\gamma(\tilde{Y}_0, \ldots, \tilde{Y}_{1-p}))$ has a unique zero $\gamma_1 \in \Theta$ in the strict sense of $B.2$.

The formulation in $H_1(iv)$ allows for certain deviations from stationarity of the time series after the change which can, for example, be caused by starting values from the stationary distribution before the change, while $H_1(v)$ guarantees that $\tilde{Y}_n$ converges to $\gamma_1$ under alternatives.

The following theorem extends the results of Franke et al. [13].

**Theorem 10.6** Let assumptions $H_1(i)$–$H_1(iv)$ be fulfilled.

(a) For $S_{SPAR}^\text{PAR}(k, \gamma)$ as in (10.10), $B.1$ and $B.2$ are fulfilled.

(b) For $\tilde{S}_{SPAR}^\text{PAR}(k, \gamma) = \sum_{j=1}^{k} (Y_t - f_\gamma(Y_{t-1}, \ldots, Y_{t-p}))$ and if $k_0 = \lfloor \lambda n \rfloor$, $B.5$ is fulfilled with $F_\lambda(t) = \mathbb{E}(Y_1 - \gamma_1^T Y_0)$.

(c) For $S_{SPAR}^\text{PAR}(k, \gamma) = (10.10)$ with $f_\gamma(x) = \gamma^T x$ and if $k_0 = \lfloor \lambda n \rfloor$, then $B.5$ is fulfilled with $F_\lambda(t) = \mathbb{E}Y_0(Y_1 - \gamma_1^T Y_0)$.

From this, we can give assumptions under which the corresponding tests have asymptotic power one and the point where the maximum is attained is a consistent estimator for the change point in rescaled time by Theorem 10.2.

$B.4$ is always fulfilled for the full score statistic if the time series before and after the change are correctly specified by the given Poisson autoregressive model. Otherwise, restrictions apply.

Doukhan and Kengne [7] propose to use several Wald-type statistics based on maximum likelihood estimators in Poisson autoregressive models. While their statistics are also designed for the at most one change situation, they explicitly prove consistency under the multiple change point alternative.
10.5 Simulation and Data Analysis

In the previous sections, we have derived the asymptotic limit distribution for various statistics as well as shown that the corresponding tests have asymptotic power one under relatively general conditions. In particular, we have proven the validity of these conditions for two important classes of integer-valued time series: binary autoregressive and Poisson counts. In this section, we give a short simulation study in addition to some data analysis to illustrate the small sample properties of these tests complementing simulations of Hudecová [17] and Fokianos et al. [11]. The critical values are obtained from Monte Carlo experiments of the limit distribution based on 1000 repetitions.

10.5.1 Binary Autoregressive Time Series

In this section, we consider a first-order binary autoregressive time series (BAR(1)) as defined in Section 10.3 with \( Z_{t-1} = (1, Y_{t-1}) \). We consider the statistic

\[
T_n = \max_{1 \leq k \leq n} \frac{1}{n} (s_{\text{BAR}}^k (\check{\beta}_n))^T \check{\Sigma}^{-1} s_{\text{BAR}}^k (\check{\beta}_n),
\]

where \( \check{\Sigma} = \frac{1}{n} \sum_{t=1}^n Z_{t-1} Z_{t-1}^T \pi_t (\check{\beta}_n) (1 - \pi_t (\check{\beta}_n)) \)

and \( s_{\text{BAR}} \) is as in (10.6) and \( \check{\beta}_n \) as in (10.7). Since \( \check{\Sigma} \) consistently estimates \( \Sigma = \mathbb{E} (Z_{t-1} Z_{t-1}^T \pi_t (\beta_0) (1 - \pi_t (\beta_0))) \) under the null hypothesis, by Theorem 10.3 and Theorem 10.1, the asymptotic null distribution of this statistic is given by

\[
\sup_{0 \leq t \leq 1} \left( B_1^2 (t) + B_2^2 (t) \right)
\]

for two independent Brownian bridges \( \{B_1 (\cdot)\} \) and \( \{B_2 (\cdot)\} \) with a simulated 95% quantile of 2.53. Table 10.1 reports the empirical size and power (based on 10,000 repetitions) for various alternatives, where a change always occurred at \( n/2 \). Figure 10.1 shows one sample path for the null hypothesis and each of the alternatives considered there. The size is always conservative and gets closer to the nominal level with increasing sample size as predicted by the asymptotic results. The power is good and increases with the sample size, where some alternatives have better power than others.

<table>
<thead>
<tr>
<th>TABLE 10.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical Size and Power of Binary Autoregressive Model with ( \beta_1 = (2, -2) ) (Parameter before the Change)</td>
</tr>
<tr>
<td>( n )</td>
</tr>
<tr>
<td>200</td>
</tr>
<tr>
<td>500</td>
</tr>
<tr>
<td>1000</td>
</tr>
</tbody>
</table>
Sample paths for the BAR(1) model, $\beta_1 = (2, -2)$, $k_0 = 100$, $n = 200$.

**FIGURE 10.1**

Data Analysis: U.S. Recession Data

We now apply the test statistic mentioned earlier to the quarterly recession data in Figure 10.2 from the United States for the period 1855–2012*. The datum is 1 if there has been a recession in at least one month in the quarter and 0 otherwise. The data have been previously analyzed by different authors; in particular, they have recently been analyzed in a change point context by Hudecová [17].

We find a change in the first quarter of 1933, which corresponds to the end of the great depression that started in 1929 in the United States and leads to a huge unemployment rate in 1932. If we split the time series at that point and repeat the change point procedure, no further significant change points are found. This is consistent with the findings in Hudecová [17], who applied a different statistic based on a binary autoregressive time series of order 3.

* This data set can be downloaded from the National Bureau of Economic Research at http://research.stlouisfed.org/fred2/series/USREC.
10.5.2 Poisson Autoregressive Models

In this section, we consider a Poisson autoregressive model as in (10.9) with $\lambda_t = \gamma_1 + \gamma_2 Y_{t-1}$. For this model, we use the following test statistic based on least squares scores:

$$T_n = \max_{1 \leq k \leq n} \frac{1}{n} S^{\text{PAR}}(k, \hat{\gamma}_n)^T \hat{\Sigma}^{-1} S^{\text{PAR}}(k, \hat{\gamma}_n),$$

where $S^{\text{PAR}}(k, \gamma) = \sum_{t=1}^k Y_{t-1}(Y_t - \lambda_t)$, $Y_{t-1} = (1, Y_{t-1})^T$, and $\hat{\gamma}_n$ as in (10.10) and $\hat{\Sigma}^{-1}$ is the empirical covariance matrix of $\{Y_{t-1}(Y_t - \lambda_t)\}$. By Theorems 10.5b and 10.1, this statistic has the same null asymptotics as in (10.12). Table 10.2 reports the empirical size and power (based on 10,000 repetitions) for various alternatives, where a change always occurred at $n/2$. Figure 10.3 shows one corresponding sample path for each scenario. The test size is always conservative and gets closer to the nominal level with increasing sample size as predicted by the asymptotic results. The power is good.

### TABLE 10.2

Empirical Size and Power of Poisson Autoregressive Model with $\gamma_1 = (1, 0.75)$ (Parameter before the Change)

<table>
<thead>
<tr>
<th></th>
<th>$H_0$</th>
<th>$H_1$: $\gamma_2 = (2, 0.75)$</th>
<th>$H_1$: $\gamma_2 = (2, 0.5)$</th>
<th>$H_1$: $\gamma_2 = (1, 0.5)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>200</td>
<td>500</td>
<td>1000</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>0.028</td>
<td>0.0361</td>
<td>0.967</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>0.531</td>
<td>0.967</td>
<td>0.968</td>
<td>0.999</td>
</tr>
<tr>
<td></td>
<td>0.252</td>
<td>0.683</td>
<td>0.968</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.271</td>
<td>0.895</td>
<td>0.999</td>
<td></td>
</tr>
</tbody>
</table>
and increases with the sample size. Some pilot simulations suggest that using statistics
associated with partial likelihood scores can further increase the power. While a detailed
theoretic analysis can in principle be done based on the results in Section 10.2, it is beyond
the scope of this work.

10.5.2.1 Data Analysis: Number of Transactions per Minute for Ericsson B Stock

In this section, we use the methods given earlier to analyze the data set that consists of the
number of transactions per minute for the stock Ericsson B during July 3, 2002. The data
set consists of 460 observations instead of 480 for 8 h of transactions because the first 5 min
and last 15 min of transactions are ignored. Fokianos et al. [12] have analyzed the transac-
tions count from the same stock on a different day with a focus on forecasting the number
of transactions. The data and estimated change points (using a binary segmentation pro-
duction as described in Theorem 10.2) are illustrated in Figure 10.4. The red vertical lines

Figure 10.3:
Sample paths for the Poisson autoregressive model, $\gamma_1 = (1, 0.75)$, $k_0 = 250, n = 500$. 

$\gamma_2 = (2, 0.75)$

$\gamma_2 = (2, 0.5)$

$\gamma_2 = (1, 0.5)$
Transactions per minute for the stock Ericsson B on July 3, 2002, where the red vertical lines are the estimated change points.

indicated estimated change points at the 98th (11:12), 148th (12:02), and 300th (14:39) data points, respectively.

In fact, the empirical autocorrelation function of the complete time series in Figure 10.5 decreases very slowly, which can either be taken as evidence of long-range dependence or the presence of change points. Figure 10.6 shows the empirical autocorrelation function of
The empirical autocorrelation of the segmented data, taking into account the estimated change points.

the data in each segment supporting the conjecture that change points, rather than long-range dependence, cause the slow decrease in the empirical autocorrelation function of the full data set.

### 10.6 Online Procedures

In some situations, statistical decisions on whether a change point has occurred need to be made online as the data arrive point by point. In such situations, classical monitoring charts such as exponentially weighted moving average (EWMA) or cumulative sum (CUSUM) check for changes in a parametric model with known parameters. Unlike in classical statistical testing, it is not the size that is controlled under the null hypothesis but the average run length that should be larger than a prespecified value. On the other hand, if a change
occurs an alarm should be raised as soon as possible afterward. For first-order integer-valued autoregressive processes of Poisson counts, CUSUM charts have been investigated by Weiβ and Testik [39] as well as Yontay et al. [41] and the EWMA chart by Weiβ [38]. A different approach was proposed by Chu et al. [4] in the context of linear regression models that controls the size asymptotically and at the same time tests for changes in a specific model with unknown in-control parameters. If a change occurs, this procedure will eventually reject the null hypothesis. In their setting, a historic data set with no change is used to estimate the in-control parameters before the monitoring of new incoming observations starts. Such a data set usually exists in practice as some data are necessary before any reasonable model building or estimation for statistical inference can take place. Asymptotic considerations based on the length of this data set are used to calibrate the procedure. Kirch and Tadjuidje Kamgaing [22] generalize the approach of Chu et al. [4] to estimating functions in a similar spirit as described in Section 10.2 for the off-line procedure. Examples include integer-valued time series as considered here.

10.7 Technical Appendix

10.7.1 Regularity Assumptions

In this appendix, we summarize the regularity conditions needed to obtain the null asymptotics as well as results under alternatives for change point statistics constructed according to the principles in Section 10.2.

While the techniques used in the proofs are common in change point analysis such regularity conditions have never been isolated to the best of our knowledge and could be quite useful in the analysis of change point tests that have not yet been considered in the literature.

We denote the assumptions under the null hypothesis by $A$ and the assumptions under alternatives by $B$.

A.1 Let

$$\max_{1 \leq k \leq n} \frac{n}{k(n - k)} \left\| S(k, \hat{\theta}_n) - (S(k, \theta_0) - \frac{k}{n} S(n, \theta_0)) \right\|^2 = o_P(1)$$

for some $\theta_0$.

This assumption allows us to replace the estimator in the statistic by a fixed value $\theta_0$, which is usually given by the true or best approximating parameter of the model. The centering stems from the fact that the estimator $\hat{\theta}_n$ is the zero of $S(n, \hat{\theta}_n)$. Typically, this assumption can be derived using a Taylor expansion in addition to the $\sqrt{n}$-consistency of $\hat{\theta}_n$ for $\theta_0$.

A.2 (i) Let $\{ \frac{1}{\sqrt{n}} S([nt]; \theta_0) : 0 \leq t \leq 1 \}$ fulfill a functional central limit theorem toward a Wiener process $\{ W(t) : 0 \leq t \leq 1 \}$ with regular covariance matrix $\Sigma$ as a limit.
Detection of Change Points in Discrete-Valued Time Series

(ii) Let both a forward and backward Hájek–Rényi inequality hold true; that is, for all \(0 \leq \alpha < 1\), it holds

\[
\max_{1 \leq k \leq n} \frac{1}{n^{1-\alpha} k^{\alpha}} \|S(k, \theta_0)\|^2 = O_p(1),
\]

\[
\max_{1 \leq k \leq n} \frac{1}{n^{1-\alpha} (n-k)^{\alpha}} \|S(n, \theta_0) - S(k, \theta_0)\|^2 = O_p(1).
\]

Both assumptions are relatively weak and are fulfilled by a large class of time series. Hájek–Rényi-type inequalities as given earlier can, for example, be obtained from moment conditions (Appendix B.1 in Kirch [19]).

For the Darling–Erdös-type asymptotics, we need the following stronger assumption.

A.3  (i) Let \(S(k, \theta_0)\) fulfill a strong invariance principle, that is, (possibly after changing the probability space) there exists a \(d\)-dimensional Wiener process \(W(\cdot)\) with regular covariance matrix \(\Sigma\) such that

\[
\frac{1}{\sqrt{n}} \left( S(n, \theta_0) - W(n) \right) = o \left( (\log \log n)^{-1/2} \right) \quad a.s.
\]

(ii) Let \(\{S(n, \theta_0) - S(k, \theta_0) : k = \frac{n}{2}, \ldots, n-1\} \overset{D}{=} \{\tilde{S}(j, \theta_0) : j = 1, \ldots, n/2\}\) such that

\[
\frac{1}{\sqrt{n}} \left( \tilde{S}(n, \theta_0) - \tilde{W}(n) \right) = o \left( (\log \log n)^{-1/2} \right) \quad a.s.,
\]

with \(\{\tilde{W}(\cdot)\} \overset{D}{=} \{W(\cdot)\}\).

(iii) Let

\[
\max_{1 \leq k \leq n/\log n} \frac{1}{k} S(k, \theta_0) \Sigma^{-1} S(k, \theta_0)
\]

\[
and \quad \max_{n-n/\log n \leq k \leq n} \frac{1}{n-k} (S(n, \theta_0) - S(k, \theta_0)) \Sigma^{-1} (S(n, \theta_0) - S(k, \theta_0))
\]

be asymptotically independent.

Invariance principles as in (i) have been obtained for different kinds of weak dependence concepts such as mixing to state a classic result (Philipps and Stout [29]), where the rate is typically of polynomial order. Since the definition of mixing is symmetric, the backward invariance principle also follows for such time series. Assumption (iii) is fulfilled by the definition of mixing but can otherwise be difficult to prove. Similarly, Assumption (ii) does not necessarily follow by the same methods as (i) (e.g., the proof of Theorem 10.3, where stronger assumptions are needed to get (iii)). Even with the weaker rate \(o(1)\), part (i) implies Assumption A.2 (i) and (ii) for the forward direction (using the Hájek–Rényi inequality for independent random variables) while the backward direction follows from A.3 (ii). Assumption (iii) can be difficult to prove but follows for mixing time series by definition.
We will now state some regularity assumptions under the alternative, for which the statistics given earlier have asymptotic power one.

**B.1** Let \( \sup_{\theta \in \Theta} \left\| \frac{1}{n} S(\lfloor nt \rfloor, \theta) - F_t(\theta) \right\| = o_p(1) \) uniformly in \( 0 \leq t \leq 1 \) for some function \( F_t(\theta) \).

This assumption can be obtained under weak moment assumptions from a strong uniform law of large numbers such as Theorem 6.5 in Ranga Rao [30] if the time series after the change can be approximated by a stationary and ergodic time series and \( \theta \) comes from a compact parameter set.

**B.2** Let \( S(n, \hat{\theta}_n) = 0 \) and \( \theta_1 \) the unique zero of \( F_1(\theta) \) in the strict sense, that is, for every \( \epsilon > 0 \) there exists a \( \delta > 0 \) such that \( |F_1(\theta)| > \delta \) whenever \( \|\theta - \theta_1\| > \epsilon \).

The first assumption is related to the centering in Assumption A.1 and is usually fulfilled by construction. The second assumption guarantees that this estimator converges to \( \theta_1 \) under alternatives. The strict uniqueness condition given there is automatically fulfilled if \( F_1 \) is continuous in \( \theta \) and \( \Theta \) is compact.

**Proposition 10.1** Under Assumptions B.1 and B.2, it holds \( \hat{\theta}_n \xrightarrow{p} \theta_1 \).

For the main theorem, we can allow the score function to be different from the one used in the estimation of \( \hat{\theta}_n \). A typical example is a projection into a lower-dimensional space such as only the first component of the vector (see Theorem 10.6b for an example). Assumption B.2 will then typically not be fulfilled as \( \theta_1 \) can no longer be the unique zero. However, to get the main theorem, it can be replaced by the following:

**B.3** \( \hat{\theta}_n \xrightarrow{p} \theta_1 \).

**B.4**

(i) \( F_\lambda(\theta) \Sigma^{-1} F_\lambda(\theta) \geq c > 0 \) in an environment of \( \theta_1 \).

(ii) \( \liminf_{n \to \infty} F_\lambda(\theta) \hat{\Sigma}_n^{-1} F_\lambda(\theta) \geq c > 0 \) in an environment of \( \theta_1 \).

(iii) \( \hat{\Sigma}_n \to \Sigma_\lambda \) with \( F_\lambda(\theta) \Sigma_\lambda^{-1} F_\lambda(\theta) \geq c > 0 \) in an environment of \( \theta_1 \).

This assumption is crucial in understanding which alternatives we can detect. Typically, in the correctly specified model before and after the change, and if the same score vector is used for the change point test as for the parameter estimation, (i) will always be fulfilled. However, if only part of the score vector is used as, for example, in Theorem 10.6b, then this condition describes which alternatives are still detectable. Typically, the power to detect those will be larger than for the full test at the cost of not having power for different alternatives at all. Parts (ii) and (iii) allow one to use estimators of \( \Sigma \) that do not converge or converge to a different limit matrix \( \Sigma_\lambda \) under alternatives than under the null hypothesis.
If this limit matrix $\Sigma_A$ is positive definite, then this is no additional restriction on which alternatives can be detected. Obviously, (iii) implies (ii). The following additional assumption is also often fulfilled and yields the additional assertions in (b) and (c).

**B.5** $F_t(\cdot)$ is continuous in $\theta_1$ and $F_t(\theta_1) = F_\lambda(\theta_1)g(t)$ with

$$g(t) = \begin{cases} \frac{1}{\lambda} t, & t \leq \lambda, \\ \frac{1}{1-\lambda}(1-t), & t \geq \lambda. \end{cases}$$

### 10.7.2 Proofs

**Proof of Theorem 10.1** By Assumption A.1, we can replace $S(k, \hat{\theta}_n)$ in all three statistics by $S(k, \theta_0)$ without changing the asymptotic distribution, where—for the statistics in (a)—one needs to note that by (10.3)

$$\sup_{1 \leq k \leq n} w\left(\frac{k}{n}\right) \frac{k - n}{n} = O(1).$$

For a bounded and continuous weight function, the assertion then immediately follows by the functional central limit theorem. For an unbounded weight function, note that for any $0 < \tau < 1/2$, it follows from the Hájek–Rényi inequality and (10.3) that

$$\max_{1 \leq k \leq \tau n} w(k/n) \frac{k - n}{n} \sum^{-1/2}_1 \max_{1 \leq k \leq n 1 - \alpha} \|S(k, \theta_0)\|^2 \rightarrow 0$$

as $\tau \rightarrow 0$ uniformly in $n$. By the backward inequality and the fact that $S(k, \theta_0) - \frac{k}{n} S(n, \theta_0) = -((S(n, \theta_0) - S(k, \theta_0)) - \frac{n-k}{n} S(n, \theta_0))$, an analogous assertion holds for $\max_{1 \leq k \leq n 1 - \tau}$ as well as for the corresponding maxima over the limit Brownian bridges. Since the functional central limit theorem implies the claimed distributional convergence for $\max_{1 \leq k \leq n 1 - \tau}$, careful arguments yield the assertion. The result for estimated $\Sigma$ is immediate.

To prove (b), first note that by the invariance principle in A.2 and the law of iterated logarithm, we get

$$\max_{1 \leq k \leq \log n} \frac{n}{k(n - k)} \left(S(k, \theta_0) - \frac{k}{n} S(n, \theta_0)\right)^T \sum^{-1} \left(S(k, \theta_0) - \frac{k}{n} S(n, \theta_0)\right) = o_P \left(\left(\frac{b_d(\log n)}{a(\log n)}\right)^2\right).$$

By the invariance principle and Theorem 2.1.4 in Schmitz [33] (the theorem used is for the univariate case but the rates immediately carry over to the multivariate situation here), we also get
The invariance principle in combination with Horvath [15], Lemma 2.2 (in addition to analogous arguments as given earlier), implies that

\[
P \left( a(\log n) \max_{\log n \leq k \leq \log n} \sqrt{\frac{n}{k(n-k)}} \left( S(k, \theta_0) - \frac{k}{n} S(n, \theta_0) \right) ^T \Sigma^{-1} \left( S(k, \theta_0) - \frac{k}{n} S(n, \theta_0) \right) \right) \leq \exp(-t^2).
\]

By Assumption A.2, the exact same arguments lead to analogous assertions for \( k \geq n/2 \), which imply the assertion by the asymptotic independence guaranteed by Assumption A.3. From this, we also get that

\[
\sqrt{\log n} \left\| \sqrt{\frac{n}{k(n-k)}} \left( S(k, \theta_0) - \frac{k}{n} S(n, \theta_0) \right) ^T \Sigma^{-1} \left( S(k, \theta_0) - \frac{k}{n} S(n, \theta_0) \right) \right\| \leq \sqrt{\log n} \left\| \left( \Sigma^{-1/2} - \hat{\Sigma}^{-1/2} \right) \sqrt{\frac{n}{k(n-k)}} \left( S(k, \theta_0) - \frac{k}{n} S(n, \theta_0) \right) \right\| = O_p(\log n) \left\| \Sigma^{-1/2} - \hat{\Sigma}^{-1/2} \right\| = o_p(1),
\]

showing that the statistic with estimated covariance matrix has the same asymptotics.

\( \square \)

**Proof of Proposition 10.1** Using the subsequence principle, it suffices to prove the following deterministic result: Let \( \sup_x \| G_n(x) - G(x) \| \to 0 \) (as \( n \to \infty \)). Then it holds for \( G_n(x_n) = 0 \) and \( x_1 \) is the unique zero of \( G(x) \) in the strict sense of B.2 that \( x_n \to x_1 \). To this end, assume that this is not the case. Then there exists \( \varepsilon > 0 \) and a subsequence \( \alpha(n) \) such that \( |x_{\alpha(n)} - x_1| \geq \varepsilon \). But then since \( x_1 \) is a unique zero in the strict sense, \( \| G(x_{\alpha(n)}) \| \geq \delta \) for some \( \delta > 0 \). However, this is a contradiction as

\[
\| G(x_{\alpha(n)}) \| = \| G(x_{\alpha(n)}) - G(x_{\alpha(n)}) (x_{\alpha(n)}) \| \leq \sup_x \| G(x_{\alpha(n)}) (x) - G(x) \| \to 0.
\]

\( \square \)

**Proof of Theorem 10.2** B.1, B.3, (10.5), and B.5 (i) imply

\[
\frac{1}{n^2} S(k_0, \hat{\theta}_n) \Sigma^{-1} S(k_0, \hat{\theta}_n) \geq c + o_p(1);
\]
hence,
\[
\max_{1 \leq k \leq n} \frac{w(k/n)}{n} S(k, \hat{\theta}_n)^T \Sigma^{-1} S(k, \hat{\theta}_n) \geq n w(\lambda)(c + o_P(1)) \to \infty,
\]
\[
\frac{a(\log n)}{b_d(\log n)} \max_{1 \leq k \leq n} \frac{n}{k(n-k)} S(k, \hat{\theta}_n)^T \Sigma^{-1} S(k, \hat{\theta}_n) \geq \frac{a(\log n)}{b_d(\log n)} n \frac{1}{\lambda(1-\lambda)} (c + o_P(1)) \to \infty,
\]
which implies assertion (a). If additionally, B.5 holds, we get the assertion for the maximum-type statistic analogously if we replace \( k_0 \) by \([\delta n] \) with \( w(\delta) > 0 \). For the sum-type statistics, we similarly get
\[
\frac{1}{n} \sum_{j=1}^{n} w(k/n) \frac{1}{n} S(k, \hat{\theta}_n)^T \Sigma^{-1} S(k, \hat{\theta}_n) = n c \left( \int_0^1 w(t)g^2(t) \, dt + o_P(1) \right) \to \infty,
\]
since the assumptions on \( w(\cdot) \) guarantee the existence of \( \int_0^1 w(t)g^2(t) \, dt \) and \( \int_0^1 w(t)g^2(t) \, dt \neq 0 \). The second assertion follows by standard arguments since \( \lambda \) is the unique maximizer of the continuous function \( g \) and by Assumptions B.1 and B.5, it holds
\[
\sup_{0 \leq t \leq 1} \left| \frac{1}{n^2} S([nt]_j, \hat{\theta}_n)^T \Sigma^{-1} S([nt]_j, \hat{\theta}_n) - F_{\lambda}(\theta_1) \Sigma^{-1} F_{\lambda}(\theta_1) g^2(t) \right| \to 0,
\]
where B.4 guarantees that the limit is not zero. The proofs show that the assertions remain true if \( \Sigma \) is replaced by \( \hat{\Sigma}_n \) under the stated assumptions. \( \square \)

**Proof of Theorem 10.3** Assumption A.1 can be obtained by a Taylor expansion, the ergodic theorem, and the \( \sqrt{n} \)-consistency of the estimator \( \hat{\theta}_n \). The arguments are given in detail in Fokianos et al. [11] (Proof of Proposition 3), where by the stationarity of \( \{Z_t\} \) their arguments go through in our slightly more general situation for \( k \leq n/2 \). For \( k > n/2 \), analogous arguments give the assertion on noting that (with the notation of Fokianos et al. [11])
\[
\sum_{i=1}^{k} Z^{(i)}_{t-i} Z^{(j)}_{t-i-1} \tau_t(\beta)(1 - \tau_t(\beta)) - \frac{k}{n} \sum_{i=1}^{n} Z^{(i)}_{t-i} Z^{(j)}_{t-i-1} \tau_t(\beta)(1 - \tau_t(\beta))
\]
\[
= - \sum_{i=k+1}^{n} Z^{(i)}_{t-i} Z^{(j)}_{t-i-1} \tau_t(\beta)(1 - \tau_t(\beta)) + \frac{n-k}{n} \sum_{i=1}^{n} Z^{(i)}_{t-i} Z^{(j)}_{t-i-1} \tau_t(\beta)(1 - \tau_t(\beta)).
\]
Assumption A.3 (i) follows from the strong invariance principle in Proposition 2 of Fokianos et al. [11]. Assumption A.3 (ii) does not follow by the same proof techniques as an autoregressive process in reverse order has different distributional properties than an autoregressive process. However, if the covariate \( \{Y_t, Z_{t-1}, \ldots, Z_{t-p}\}^T \) is \( \alpha \)-mixing, the same holds true for the summands of the score process (with the same rate). Since the mixing property also transfers to the time-inverse process, the strong invariance principle follows from the invariance principle for mixing processes given by Kuelbs and Philipp [23], Theorem 4. The mixing assumption then also implies A.3 (iii). \( \square \)
Proof of Theorem 10.4  First note that
\[
\sup_{\theta \in \Theta} \| S^{\text{BAR}}(k, \beta) - E S^{\text{BAR}}(k, \beta) \| = o_p(n) \tag{10.13}
\]
uniformly in \( k \leq k_0 = \lfloor n \lambda \rfloor \) by the uniform ergodic theorem of Ranga Rao [30], Theorem 6.5. For \( k > k_0 \), it holds
\[
Z_k Y_k = \tilde{Z}_k \tilde{Y}_k + \tilde{Z}_k R_1(t) + \tilde{Y}_k R_2(t) + R_1(t) R_2(t),
\]
\[
Z_k \pi_k(\beta) = \tilde{Z}_k \pi_k(\beta) + \tilde{R}_1(k) \pi_k(\beta) = \tilde{Z}_k \tilde{\pi}_k(\beta) + O(\tilde{Z}_k \beta^T R_1(k)) + O(R_1(k)),
\]
where \( g(\tilde{\pi}_k(\beta)) = \beta^T \tilde{Z}_{t-1} \) and the last line follows from the mean value theorem. An application of the Cauchy–Schwarz inequality together with (iii) and the compactness of \( \Theta \) shows that the remainder terms are asymptotically negligible, implying that
\[
\sup_{\beta} \| S^{\text{BAR}}(k, \beta) - S^{\text{BAR}}(k_0, \beta) - E S^{\text{BAR}}(k, \beta) - E S^{\text{BAR}}(k_0, \beta) \| = o_p(n)
\]
uniformly in \( k > k_0 \), where
\[
\hat{S}^{\text{BAR}}(k, \beta) = \sum_{t=1}^{\min(k_0, k)} Z_{t-1}(Y_t - \pi_t(\beta)) + \sum_{t=k_0+1}^{k} \tilde{Z}_{t-1}(\tilde{Y}_t - \tilde{\pi}_t(\beta))
\]
Together with (10.13), this implies B.1 with
\[
F_1(\lambda) = \min(t, \lambda) E Z_0(Y_1 - \pi_1(\beta)) + (t - \lambda) E \tilde{Z}_{n-1}(\tilde{Y}_n - \tilde{\pi}_n(\beta)).
\]
Since \( F_1(\beta) \) is continuous in \( \beta \), B.2 follows from (iv). B.5 follows since by definition of \( \beta_1 \) it holds \( E \tilde{Z}_{n-1}(\tilde{Y}_n - \tilde{\pi}_n(\beta_1)) = -\lambda/(1 - \lambda) E Z_0(Y_1 - \pi_1(\beta_1)) \).

Proof of Theorem 10.5  It suffices to show that the assumptions of Theorem 10.1 are fulfilled. Assumption A.1 follows for (a) and (b) analogously to the proof of Lemma 1 in Franke et al. [13]. The invariance principles in A.2 and A.3 then follow from the strong mixing assumption and the invariance principle of Kuelbs and Philipp [23]. The asymptotic independence of A.3 also follows from the mixing condition.

Proof of Theorem 10.6  This is analogous to the proof of Theorem 10.4, where the mean value theorem is replaced by the Lipschitz assumption, which also implies that \(|f_\gamma(x)| \leq |f_\gamma(0) + x|\).

Acknowledgments
Most of this work was done while the first author was at Karlsruhe Institute of Technology (KIT), where her position was financed by the Stifterverband für die Deutsche Wissenschaft
by funds of the Claussen-Simon-trust. The second author was partly supported by the DFG grant SA 1883/4-1. This work was partly supported by DFG grant KI 1443/2-2.

References


