Chapter 1

Detection of change points in discrete-valued time series

Claudia Kirch, Joseph Tadjuidje Kamgaing

1.1 Introduction

Fitting reasonable models to time series data is a very important task leading to meaningful statistical inference such as prediction. However, there is always the possibility that the best describing model or the best fitting parameter changes at some unknown point in time rendering any statistical analysis meaningless that does not take this fact into account. Consequently, it is an important question in time series analysis how to detect and estimate such change points.

There are two types of change point procedures: Offline or retrospect procedures which analyse a data set for changes after it has been fully observed on the one hand and sequential or monitoring schemes which analyse data as it arrives on the other hand. In sequential change point analysis classical monitoring charts such as EWMA (exponentially weighted moving average) or CUSUM (cumulative sum) 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 which should be larger than a prespecified value. On the other hand if a change occurs an alarm should be raised as soon as possible afterwards. For first order integer-valued autoregressive processes of Poisson counts CUSUM charts have been investigated by Weiß and Testik [25] as well as Yontay et al. [27] and the EWMA chart by Weiß [24]. A different approach was proposed by Chu et al. [1] in the context of linear regression models which controls asymptotically the size 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 beginning to monitor new incoming observations. Such a data set usually exists in practice as some data needs to be raised before any reasonable model building or estimation for statistical inference can take place. Asymptotic considerations based on the length of this data set are then used to calibrate the procedure. Kirch and Tadjuidje Kamgaing [15] generalize the approach of Chu et al. [1] to using estimating functions in a similar spirit as described in Section 1.2 below for the offline procedure. Examples include integer-valued time series as considered here.

In this paper, we focus on offline change point detection for integer valued time series. This has been considered by Hudeková [11] as well as Fokianos et al. [7] for binary time series models as well as Franke et al. [9] and Doukhan and Kegne [3] for Poisson autoregressive models. Related procedures have also been investigated by Fokianos and Fried [5, 6] for INGARCH respectively log-linear Poisson autoregressive time series but with a focus on outlier detection and intervention effects rather than change points. In order to unify the methodology used in these papers, we will first review standard change point techniques not restricted to the integer valued case in Section 1.2. We will establish standard asymptotic theory under both the null hypothesis as well as alternatives using regularity conditions which to the best of our knowledge has not been done in the literature before. These regularity conditions will then allow us to restate and even generalize the existing methods in the literature both under the null hypothesis as well as under alternatives for binary time series in Section 1.3 as well as Poisson autoregressive models in Section 1.4. In Section 1.5 some simulations as well as applications to real data illustrate the performance of these procedures. Finally, the proofs are given in an appendix.

1.2 General principles of retrospective change point analysis

In this section, we consider retrospective change point tests, where the full data set $Y_1, \ldots, Y_n$ is observed before a decision about a possible change point is made. Under the null hypothesis of no change the likelihood function is completely parametrized by a (partially) unknown parameter $\theta_0 \in \Theta \subset \mathbf{R}^d$. Under the alternative of exactly one change (AMOC - at most one change - situation), there exists $1 < k_0 < n$ such that $Y_1, \ldots, Y_{k_0}$ is parametrized by the same parameter θ_0, hence has analogous likelihood function $L\left((Y_1, \ldots, Y_{k_0}), \theta_0\right)$ but $Y_{k_0+1}, \ldots, Y_n$ is parametrized by a different parameter $\theta_1 \neq \theta_0$ but the same likelihood structure, i.e. with likelihood function $L((Y_{k_0+1}, \ldots, Y_n), \theta_1)$. In this situation k_0 is called the change point and assumed not to be known. If k_0 were known, this would amount to a simpler two-sample situation. We will now first explain general construction principles for change point tests in this situation: To this end, consider first the simpler two-sample situation, where k_0 is known. Then a likelihood ratio approach yields the following statistic

$$\ell(k) := \ell((Y_1, \ldots, Y_n), \widehat{\theta}_n) - \ell((Y_1, \ldots, Y_k), \widehat{\theta}_k) - \ell((Y_{k+1}, \ldots, Y_n), \widehat{\theta}_k^\circ),$$

where $\ell(\mathbf{Y}, \theta)$ is the log-likelihood function, $\widehat{\theta}_k$ and $\widehat{\theta}_k^\circ$ are the maximum likelihood estimator based on $Y_1, \ldots, Y_k$ respectively on $Y_{k+1}, \ldots, Y_n$. In the change point situation, the change-point k_0 is not known, therefore standard statistics either use weighted sums (sum-type statistics) or weighted maxima (max-type-statistics) of $\ell(k)$, $k = 1, \ldots, n$. Similarly, one can construct Wald-type-statistics based on quadratic forms of

$$W(k) := \widehat{\theta}_k - \widehat{\theta}_k^\circ, \qquad k = 1, \ldots, n,$$

or score-type statistics based on quadratic forms of the scores

$$S(k) := S(k, \widehat{\theta}_n) = \frac{\partial}{\partial \theta} \ell\left((Y_1, \ldots, Y_k), \theta\right)|_{\theta=\widehat{\theta}_n}, \qquad k = 1, \ldots, n.$$

Wald and Score-type-statistics can obviously also be constructed based on different estimation methods such as (weighted) least-squares or M-estimators, where reasons to do so can range from computational considerations to robustness issues. In non-linear situations, which are typically encountered for integer-valued time series, estimators such as the maximum-likelihood-estimator are usually not analytically known but need to be calculated using numerical optimization methods leading to both additional computational effort to calculate the statistics as well as an additional error due to the numerical error. The latter problems can be reduced by using score-type-statistics as in this case only the estimator based on the full data set $Y_1, \ldots, Y_n$ needs to be calculated while otherwise $\widehat{\theta}_k$ as well as $\widehat{\theta}_k^\circ$ need to be calculated for every $k = 1, \ldots, n$ with all the problems this entails.

In order to understand where typical weights used in connection with scores come from, it is helpful to take a look at the classical linear autoregressive process of order p with standard normal errors, i.e. where

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

In this case, some tedious calculations show that the following relation holds (exactly not only asymptotically)

$$2\ell(k) = W(k)^T \boldsymbol{C}_k \boldsymbol{C}_n^{-1} \boldsymbol{C}_k^\circ W(k) = S(k)^T \boldsymbol{C}_k^{-1} \boldsymbol{C}_n \left(\boldsymbol{C}_k^\circ\right)^{-1} S(k),$$

$$\text{where} \quad \boldsymbol{C}_k = \sum_{t=1}^{k} \mathbb{Y}_{t-1}\mathbb{Y}_{t-1}^T, \qquad \boldsymbol{C}_k^\circ = \sum_{t=k+1}^{n} \mathbb{Y}_{t-1}\mathbb{Y}_{t-1}^T, \qquad \mathbb{Y}_{t-1} = (1, Y_{t-1}, \ldots, Y_{t-p})^T.$$

It turns out that the maximum over these statistics which corresponds to the maximum likelihood statistic does not converge in distribution to a non degenerate limit but almost surely to infinity. In the simple mean-change-model (where $p = 0$, $S(k) = \sum_{j=1}^{k}(Y_j - \bar{Y}_n)$, $C_k = k$ and $C_k^\circ = n - k$) this follows from the law of iterated logarithm. Nevertheless asymptotic level α tests based on this maximum likelihood statistic can be constructed using a Darling-Erdős limit theorem as stated in Theorem 1.2.1 b) below. In small samples, however, the slow convergence of Darling-Erdőes limit theorems often leads to some size distortions. In order to overcome this problem, the following modified class of statistics has been proposed in the literature

$$\max_{1\leqslant k\leqslant n} \frac{w(k/n)}{n} S(k)^T C_n^{-1} S(k),$$

where $w : [0,1] \to \mathbb{R}_+$ is a non-negative continuous weight function fulfilling

$$\lim_{t\to 0} t^\alpha w(t) < \infty, \quad \lim_{t\to 1}(1-t)^\alpha w(t) < \infty \quad \text{for some } 0 \leqslant \alpha < 1$$
$$\sup_{\eta\leqslant t\leqslant 1-\eta} w(t) < \infty \quad \text{for all } 0 < \eta \leqslant 1/2. \tag{1.2.1}$$

Theorem 1.2.1 a) below shows that this class of statistics converges, under regularity conditions, in distribution to a non-degenerate limit. Under the null hypothesis the weights, implicit in the maximum likelihood statistics, can be approximated asymptotically (as $k \to \infty$,$n-k \to \infty$, $n \to \infty$) by

$$\boldsymbol{C}_k^{-1}\boldsymbol{C}_n(\boldsymbol{C}_k^\circ)^{-1} = \frac{n}{k(n-k)}\boldsymbol{C}^{-1} + o_P(1), \qquad \boldsymbol{C} = \mathbb{E}\mathbb{Y}_{t-1}\mathbb{Y}_{t-1}^T.$$

For this reason, the following choice of weight function has often been proposed in the literature

$$w(t) = (t(1-t))^{-\gamma}, \qquad 0 \leqslant \gamma < 1,$$

where a γ close to 1 detect early or late changes with better power. Similarly, if a priori information about the location of the change point is available one can increase power for such alternatives by choosing a weight function that is maximal around those points while still having asymptotic power one for other change locations (confer Theorem 1.2.2 below for the second statement).

Now, we need certain regularity conditions under which we will derive the asymptotic distribution for score-type change point statistics under the null hypothesis in Theorem 1.2.1 below. Further regularity conditions under alternatives guarantee that the corresponding tests have asymptotic power one. These regularity conditions are neither limited to likelihood scores nor to particular time series models and will be used in the subsequent sections to get the null asymptotics of the partial likelihood score statistics for binary time series in Section 1.3 as well as the least-squares score statistics for Poisson autoregressive models in Section 1.4.

While the techniques used in the proofs are quite common in change point analysis such regularity conditions have never been isolated to the best of our knowledge.

A. 1. *Let*

$$\max_{1\leqslant k\leqslant n} \frac{n}{k(n-k)} \left\| S(k;\widehat{\theta}_n) - (S(k;\theta_0) - \frac{k}{n} S(n;\theta_0)) \right\|^2 = o_P(1)$$

for some θ_0.

This assumptions allows us to replace the estimator in the statistic by a fixed value θ_0 (usually given by the true or best approximating parameter in the given model) in addition to some centering, which is needed as the estimator $\widehat{\theta}_n$ is the zero of $S(n;\widehat{\theta}_n)$. Typically, this assumption can be derived using a Taylor-expansion in addition to the $\sqrt{n}$-consistency of $\widehat{\theta}_n$ for θ_0.

A. 2. *(i) Let* $\{\frac{1}{\sqrt{n}} S(\lfloor nt \rfloor;\theta_0) : 0 \leqslant t \leqslant 1\}$ *fulfill a functional central limit theorem towards a Wiener process* $\{\mathbf{W}(t) : 0 \leqslant t \leqslant 1\}$ *with regular covariance matrix* Σ *as limit.*

(ii) Let both a forward and backward Hájek-Rényi inequality hold true, i.e. for all $0 \leqslant \alpha < 1$ *it holds*

$$\max_{1\leqslant k\leqslant n} \frac{1}{n^{1-\alpha}\, k^{\alpha}} \|S(k;\theta_0)\|^2 = O_P(1), \qquad \max_{1\leqslant k\leqslant n} \frac{1}{n^{1-\alpha}\, (n-k)^{\alpha}} \|S(n;\theta_0) - S(k;\theta_0)\|^2 = O_P(1).$$

Both assumptions are relatively weak and are fulfilled by a large class of time series. Hájek-Rényi-type inequalities as above can for example be obtained from moment conditions (confer Appendix B.1 in Kirch [13]).

For the Darling-Erdős-type asymptotics we need the following stronger assumption.

A. 3. *(i) Let $S(k;\theta_0)$ fulfill a strong invariance principle, i.e. (possibly after changing the probability space) there exists a d-dimensional Wiener process $\mathbf{W}(\cdot)$ with regular covariance matrix Σ such that*

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

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

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

with $\{\tilde{\mathbf{W}}(\cdot)\}\stackrel{\mathcal{D}}{=}\{\mathbf{W}(\cdot)\}$.

(iii) Let

$$\max_{1\leqslant k\leqslant n/\log n}\frac{1}{k}S(k;\theta_0)\Sigma^{-1}S(k;\theta_0)$$
$$\text{and}\quad\max_{n-n/\log n\leqslant k\leqslant n}\frac{1}{n-k}(S(n;\theta_0)-S(k;\theta_0))\Sigma^{-1}(S(n;\theta_0)-S(k;\theta_0))$$

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 (confer Philipps and Stout [18]), 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) (confer e.g. the proof of Theorem 1.3.1 below, where stronger assumptions are needed to get (ii)). 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 also be difficult to prove but follows for mixing time series by definition.

Theorem 1.2.1. *We get the following null asymptotics:*

a) *Let A.1 and A.2 (i) hold. Assume that the weight function is either a continuous non-negative and bounded function* $w : [0, 1] \to \mathbb{R}_+$ *or for unbounded functions fulfilling* (1.2.1) *let additionally A.2 (ii) hold. Then:*

$$(i) \quad \max_{1\leqslant k\leqslant n} \frac{w(k/n)}{n} S(k)^T \Sigma^{-1} S(k) \xrightarrow{\mathcal{D}} \sup_{0\leqslant t\leqslant 1} w(t) \sum_{j=1}^{d} B_j^2(t) = D_1,$$

$$(ii) \quad \sum_{1\leqslant k\leqslant n} \frac{w(k/n)}{n^2} S(k)^T \Sigma^{-1} S(k) \xrightarrow{\mathcal{D}} \int_0^1 w(t) \sum_{j=1}^{d} B_j^2(t)\, dt = D_2,$$

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

b) *Under A. 1 and A. 3 it holds*

$$P\left(a(\log n) \max_{1\leqslant k\leqslant n} \sqrt{\frac{n}{k(n-k)} S(k)^T \Sigma^{-1} S(k)} - b_d(\log n) \leqslant t\right) \to \exp(-2e^{-t}),$$

where $a(x) = \sqrt{2 \log x}$, $b_d(x) = 2 \log x + \frac{d}{2} \log \log x - \log \Gamma(d/2)$ *and* $\Gamma(\cdot)$ *is the Gamma-function. Furthermore,* Σ *can be replaced by an estimator* $\widehat{\Sigma}_n$ *if* $\|\widehat{\Sigma}_n^{-1/2} - \Sigma^{-1/2}\| = o_P((\log \log n)^{-1})$.

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.

We will now state some regularity assumptions under the alternative, for which the above statistics have asymptotic power one. Additionally, we propose a consistent estimator of the change point in rescaled time.

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* $0 \leqslant t \leqslant 1$ *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 [19] if the time series after the change can be approximated by a stationary and ergodic time series and θ comes from a compact parameter set.

B. 2. *Let* $S(n,\widehat{\theta}_n)=0$ *and* θ_1 *the unique zero of* $F_1(\theta)$ *in the strict sense, i.e. for every* $\varepsilon>0$ *there exists a* $\delta>0$ *such that* $|F_1(\theta)|>\delta$ *whenever* $\|\theta-\theta_1\|>\varepsilon$.

The first assumption has implicitly already been contained in A.1 for the centering to make sense. It is fulfilled if the score function used in the change point procedure is the same (or a linear combination) of the one used to estimate the unknown parameters. The second assumption guarantees that this estimator converges to θ_1 under alternatives. The strict uniqueness condition given there is automatically fulfilled if F_1 is continuous in θ and Θ is compact.

Proposition 1.2.1. *Under Assumptions B.1 and B.2 it holds* $\widehat{\theta}_n \xrightarrow{P} \theta_1$.

For the main theorem, we can allow the score function to be different from the one used in the estimation of $\widehat{\theta}_n$, a typical example is a projection into a lower dimensional space such as only the first component of the vector (see Theorem 1.4.2 b) for an example). Assumption B.2 will then typically not be fulfilled as θ_1 can no longer be the unique zero. However, to get the main theorem it can be replaced by:

B. 3. $\widehat{\theta}_n \xrightarrow{P} \theta_1$.

B. 4. *The change point is of the form:* $k_0=\lfloor \lambda n\rfloor$, $0<\lambda<1$.

Assumption **B**.4 is standard in change point analysis but can be weakened for Part a) of Theorem 1.2.2 below.

B. 5. *(i)* $F_\lambda(\theta)\Sigma^{-1}F_\lambda(\theta)\geqslant c>0$ *in an environment of* θ_1.

(ii) $\liminf_{n\to\infty} F_\lambda(\theta)\widehat{\Sigma}_n^{-1}F_\lambda(\theta)\geqslant c>0$ *in an environment of* θ_1.

(iii) $\widehat{\Sigma}_n\to\Sigma_A$ *with* $F_\lambda(\theta)\Sigma_A^{-1}F_\lambda(\theta)\geqslant c>0$ *in an environment of* θ_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 e.g. in Theorem 1.4.2 b) below, 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. Part (ii) and (iii) allow to use estimators of Σ that do not converge or converge to a different limit matrix Σ_A under alternatives than under the null hypothesis. If this limit matrix Σ_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) below.

B. 6. $F_t(\cdot)$ *is continuous in* θ_1 *and* $F_t(\theta_1) = F_\lambda(\theta_1)g(t)$ *with*

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

From these assumptions we now derive the following general theorem:

Theorem 1.2.2. *Under alternatives we get the following assertions:*

a) If Assumptions B.1, B.3, B.4 and B.5(i) hold, then the Darling-Erdős- and max-type statistics for continuous weight functions fulfilling $w(\lambda) > 0$ *have asymptotic power one, i.e. for all* $x \in \mathbb{R}$ *it holds*

$$(i) \quad P\left(\max_{1\leqslant k\leqslant n} \frac{w(k/n)}{n} S(k)^T\Sigma^{-1}S(k) \geqslant x\right) \to 1,$$

$$(ii) \quad P\left(a(\log n) \max_{1\leqslant k\leqslant n} \sqrt{S(k)^T\Sigma^{-1}S(k)} - b_d(\log n) \geqslant x\right) \to 1.$$

If B.5(i) is replaced by B.5(ii), then the assertion remains true if Σ *is replaced by* $\widehat{\Sigma}_n$.

b) If additionally B.5 holds, then the max-type statistics as well as sum-type statistics for a continuous weight function $w(\cdot) \neq 0$ *fulfilling* (1.2.1) *have power one, i.e. it holds for all* $x \in \mathbb{R}$

$$P\left(\sum_{1\leqslant k\leqslant n} \frac{w(k/n)}{n^2} S(k)^T\Sigma^{-1}S(k) \geqslant x\right) \to 1.$$

If B.5(i) is replaced by B.5(ii), then the assertion remains true if Σ *is replaced by* $\widehat{\Sigma}_n$.

c) Let $\widehat{\lambda}_n = \frac{\arg\max S(k)^T\Sigma^{-1}S(k)}{n}$ *be an estimator for the change point in rescaled time* λ. *Under*

Assumptions B.1, B.3, B.4 and B.5(i) it is consistent, i.e.

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

If B.5(i) is replaced by B.5(iii), then the assertion remains true if Σ *is replaced by* $\widehat{\Sigma}_n$.

The continuity assumption on the weight function can also 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 [22]), which work as follows: The data set is tested using an at most one change test as above. If that test is significant, the data set is split at the estimated change point and the procedure repeated on both data segments until not significant anymore.

The optimal rate of convergence in b) is usually given by $\widehat{\lambda}_n - \lambda = O_P(1/n)$ but requires a much more involved proof (confer Csörgő and Horváth [2], Theorem 2.8.1, for a proof in a mean change model).

1.3 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 within a given time frame. Wilks and Wilby [26] for example observe, whether it has been raining on a specific day, Kauppi and Saikkonen [12] and Startz [21] observe whether or not a recession has occurred in a given month. A common binary time series model is given by

$$Y_t \mid Y_{t-1}, Y_{t-2}, \ldots, Z_{t-1}, Z_{t-2}, \ldots \sim \text{Bern}(\pi_t(\boldsymbol{\beta})), \qquad \text{with } g(\pi_t(\boldsymbol{\beta})) = \boldsymbol{\beta}^T \mathbb{Z}_{t-1},$$

for a regressor $\mathbb{Z}_{t-1} = (Z_{t-1}, \ldots, Z_{t-p})^T$, which can be purely exogenous, purely autoregressive or a mixture of both. Typically, the canonical link function $g(x) = \log(x/(1-x))$ is used and statistical inference is based on the partial likelihood

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

with corresponding score-vector

$$S_n(\boldsymbol{\beta}) = \sum_{t=1}^{n} \mathbb{Z}_{t-1}(Y_t - \pi_t(\boldsymbol{\beta})) \tag{1.3.1}$$

for the canonical link function above.

Theorem 1.3.1. *For the partial score vector $S_k(\boldsymbol{\beta})$ and the corresponding maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_n$ we get the following assertions under the null hypothesis:*

a) Let the covariate process $\{\mathbb{Z}_t\}$ be strictly stationary and ergodic with existing fourth moments. Then, under the null hypothesis, A.1 and A.3 (i) are fulfilled.

b) If additionally $(Y_t, Z_{t-1}, \ldots, Z_{t-p})^T$ is mixing with exponential rate, then A.3 (ii) and (iii) are also fulfilled.

Remark 1.3.1. *For $\mathbb{Z}_{t-1} = (Y_{t-1}, \ldots, Y_{t-p})^T, Y_t$ is the standard binary autoregressive model (BAR(p)), for which the assumptions of Theorem 1.3.1, b) are fulfilled, see, e.g. Wang and Li [23]. 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 [4] can be applied to derive its geometric ergodic property (see Kirch and Tadjuidje Kamgaing [14] for details on this issue) implying that it is mixing with exponential rate.*

From Theorem 1.3.1 and Theorem 1.2.1 we immediately get the null asymptotics for the corresponding change point statistics based on the full score vector $S_k(\widehat{\boldsymbol{\beta}}_n)$ as well as on linear combinations of the score vector such as using the first component, i.e. based on $\tilde{S}_k(\widehat{\boldsymbol{\beta}}_n)$ with

$$\tilde{S}_k(\boldsymbol{\beta}) = \sum_{j=1}^{k} (Y_t - \pi_t(\boldsymbol{\beta})), \tag{1.3.2}$$

where in both cases $\widehat{\boldsymbol{\beta}}_n$ is the zero of (1.3.1). The first approach for $w \equiv 1$ has been proposed by Fokianos et al. [7], the second one with a somewhat different standardization in a purely autoregressive setup has been considered by Hudeková [11]. Hudeková'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:

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

H_1(ii) The binary time series $\{Y_t\}$ as well as the covariate process $\{Z_t\}$ before the change fulfill the assumptions of Theorem 1.3.1 (a).

H_1(iii) The time series after the change as well as the covariate process after the change can be written as $Y_t = \tilde{Y}_t + R_1(t)$ respectively $\mathbb{Z}_t = \tilde{\mathbb{Z}}_t + \mathbb{R}_2(t)$, $t > \lfloor n\lambda \rfloor$, where $\{\tilde{Y}_t\}$ is bounded, stationary and ergodic and $\{\tilde{\mathbb{Z}}_t\}$ is square-integrable as well as stationary and ergodic with remainder terms fulfilling

$$\frac{1}{n} \sum_{j=\lfloor \lambda n \rfloor+1}^{n} R_1^2(t) = o_P(1), \qquad \frac{1}{n} \sum_{j=\lfloor \lambda n \rfloor+1}^{n} \|\mathbb{R}_2(t)\|^2 = o_P(1).$$

H_1(iv) $\lambda \mathbb{E}\mathbb{Z}_0(Y_1 - \pi_1(\boldsymbol{\beta})) + (1-\lambda)\mathbb{E}\tilde{\mathbb{Z}}_{n-1}(\tilde{Y}_n - \tilde{\pi}_1(\boldsymbol{\beta}))$ has a unique zero $\boldsymbol{\beta}_1 \in \Theta$ and Θ is compact and convex with $\widehat{\boldsymbol{\beta}}_n \in \Theta$.

The formulation of the alternative allows for rather general alternatives including situations where starting values from before the change are used resulting in a non-stationary time series after the change. Neither Hudeková [11] nor Fokianos et al. [7] have derived the behavior of their statistics under alternatives.

Theorem 1.3.2. *Let H_1(i)- H_1(iv) hold.*

(a) For $S(k,\boldsymbol{\beta}) = S_k(\boldsymbol{\beta})$ as in (1.3.1) B.1 and B.2 are fulfilled, which implies B.3. If $k_0 = \lfloor \lambda n \rfloor$, then B.6 is fulfilled with $F_\lambda(\boldsymbol{\beta}) = \lambda \mathbb{E}\mathbb{Z}_0(Y_1 - \pi_1(\boldsymbol{\beta}))$.

(b) For $S(k,\boldsymbol{\beta}) = \tilde{S}(k)$ as in (1.3.2) and if $k_0 = \lfloor \lambda n \rfloor$, then B.6 is fulfilled with $F_\lambda(\boldsymbol{\beta}) = \lambda \mathbb{E}(Y_1 - \pi_1(\boldsymbol{\beta}))$.

B.5 is fulfilled for the full score statistic from Theorem 1.3.2,(a), if the time series before and after the change are correctly specified by nary time series models with different parameters. Otherwise, restrictions apply. Together with Theorem 1.2.2 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.

1.4 Detection of changes in Poisson autoregressive models

Another popular model for time series of counts is given by the Poisson autoregression, where we observe $Y_{1-p}, \dots, Y_n$ with

$$Y_t \mid Y_{t-1}, Y_{t-2}, \dots, Y_{t-p} \sim \text{Pois}(\lambda_t), \qquad \lambda_t = f_\gamma(Y_{t-1}, \dots, Y_{t-p}). \tag{1.4.1}$$

If $f_\gamma(\mathbf{x})$ is Lipschitz-continuous in $\mathbf{x}$ for all $\gamma \in \Theta$ with Lipschitz constant strictly smaller than 1, then there exists a stationary ergodic solution of the (1.4.1) which is β-mixing with exponential rate (confer Neumann [17]). For a given parametric model f_θ this allows to consider score-type change point statistics based on likelihood equations using the tools of Section 1.2. The mixing condition in connection to some moment conditions typically allow to derive A. 3 while a Taylor-expansion in connection with $\sqrt{n}$-consistency of the corresponding maximum likelihood estimator as e.g. derived by Doukhan and Kegne [3], Theorem 3.2, gives A. 1 under some additional moment conditions. However, in this paper, we will concentrate on change point statistics based on least-squares scores generalizing results of Franke et al. [9]. To this end, consider the least-squares estimator $\widehat{\theta}_n$ defined (in the score version) by $S_n(\widehat{\gamma}_n) = 0$, where

$$S_n(\gamma) = \sum_{t=1}^{n} \nabla f_\gamma((Y_{t-1}, \dots, Y_{t-p}))(Y_t - f_\gamma(Y_{t-1}, \dots, Y_{t-p})), \tag{1.4.2}$$

where ∇ denotes the gradient with respect to γ. Under the additional assumption $f_\gamma(\mathbf{x}) = \gamma_1 + f_{\gamma_2,\ldots,\gamma_d}(\mathbf{x})$ this implies in particular that

$$\sum_{t=1}^{n}(Y_t - f_{\widehat{\gamma}_n}(Y_{t-1},\ldots,Y_{t-p})) = 0.$$

Assumptions under H_0:

H_0(i) Y_t is stationary and mixing with exponential rate such that $\mathbb{E}\sup_{\gamma\in\Theta} f_\gamma^2(Y_t) < \infty$.

H_0(ii) $f_\gamma(\mathbf{x}) = \gamma_1 + f_{\gamma_2,\ldots,\gamma_d}(\mathbf{x})$ is twice continuously differentiable with respect to γ for all $\mathbf{x} \in \mathbb{N}_0^p$ and

$$\mathbb{E}\sup_{\gamma\in\Theta}\|\nabla f_\gamma(\mathbb{Y}_t)\nabla^T f_\gamma(\mathbb{Y}_t)\| < \infty, \qquad \mathbb{E}\left(Y_t \sup_{\gamma\in\Theta}\|\nabla^2 f_\gamma(\mathbb{Y}_{t-1})\|\right) < \infty.$$

H_0(iii) $e(\gamma) = \mathbb{E}(Y_t - f_\gamma(\mathbb{Y}_{t-1}))^2$ has a unique minimizer γ_0 in the interior of some compact set Θ 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) 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) but only for the function used in the statistical procedure. Assumption (iii) is fulfilled for the true value if $\{Y_t\}$ is a Poisson autoregressive time series with regression function f_γ.

Theorem 1.4.1. *Let under the null hypothesis H_0(i) - (iii).*

a) If $\mathbb{E}\|\nabla f_{\gamma_0}(Y_p,\ldots,Y_1)\|^\nu < \infty$ for some $\nu > 2$, then $S(k,\gamma) = \sum_{j=1}^{k}(Y_t - f_{\widehat{\gamma}_n}(Y_{t-1},\ldots,Y_{t-p}))$ fulfills A.1 and A.3 with $\Sigma = (\sigma^2)$ the long-run variance of $Y_t - f_{\gamma_0}((Y_{t-1},\ldots,Y_{t-p}))$.

b) If $f_\gamma(\mathbf{x}) = \gamma_1 + (\gamma_2,\ldots,\gamma_d)^T\mathbf{x}$ and $\mathbb{E}|Y_0|^\nu < \infty$ for some $\nu > 4$, then $S(k,\gamma) = \sum_{j=1}^{k}\mathbb{Y}_{t-1}(Y_t - \widehat{\gamma}_n^T\mathbb{Y}_{t-1})$, where $\mathbb{Y}_{t-1} = (Y_{t-1},\ldots,Y_{t-p})^T$ fulfills A.1 and A.3, where Σ is the long-run covariance matrix of $\{\mathbb{Y}_{t-1}(Y_t - \gamma_0^T\mathbb{Y}_{t-1})\}$.

Together with Theorem 1.2.1 this implies the null asymptotics for a large class of change point statistics.

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 $\gamma \in \Theta$, Θ is compact and convex, and $f_\gamma(\mathbf{x})$ is uniformly Lipschitz in $\mathbf{x}$ 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.

This formulation allows for certain deviations from stationarity of the time series after the change which can e.g. be caused by starting values from the stationary distribution before the change.

The following theorem extends results of Franke et al. [9].

Theorem 1.4.2. *Let assumptions H_1(i) – H_1(iv) be fulfilled.*

a) For $S(k, \gamma) = S_k(\gamma)$ as in (1.4.2) B.1 and B.2 are fulfilled.

b) For $S(k, \gamma) = \sum_{j=1}^{k} (Y_t - f_\gamma(Y_{t-1}, \ldots, Y_{t-p}))$ and if $k_0 = \lfloor \lambda n \rfloor$, B.6 is fulfilled with $F_\lambda(t) = \mathbb{E}(Y_1 - \gamma_1^T \mathbb{Y}_0)$..

c) For $S(k, \gamma) = S_k(\gamma)$ as in(1.4.2) with $f_\gamma(\mathbf{x}) = \gamma^T \mathbf{x}$ and if $k_0 = \lfloor \lambda n \rfloor$, then B.6 is fulfilled with $F_\lambda(t) = \mathbb{E}\mathbb{Y}_0(Y_1 - \gamma_1^T \mathbb{Y}_0)$.

B.5 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 [3] 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.

1.5 Simulation and data analysis

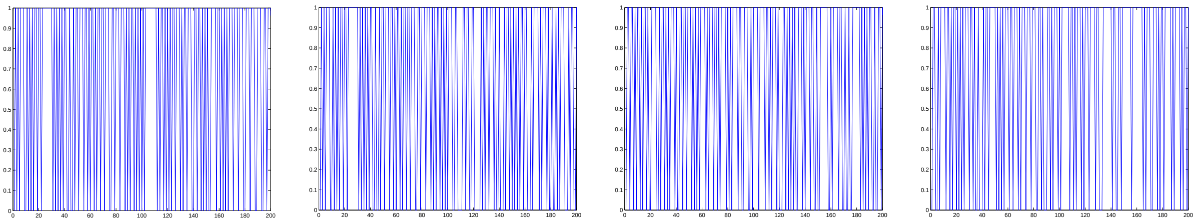
In the previous sections we have derived the asymptotic limit distribution for various statistics as well as shown that the corresponding tests have asymptotic power one under relatively general conditions. In particular, we have proven the validity of these conditions for two important classes of integer valued time series: Binary autoregressive as well as 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 Hudeková [11] and Fokianos et al. [7].

1.5.1 Binary autoregressive time series

In this section we consider a first order binary autoregressive time series (BAR(1)) as defined in Section 1.3 with $\mathbb{Z}_{t-1} = (1, Y_{t-1})$. We consider the statistic

$$T_n = \max_{1\leqslant k\leqslant n} \frac{1}{n} S_k(\boldsymbol{\beta}))^T \widehat{\boldsymbol{\Sigma}}^{-1} S_k(\boldsymbol{\beta}),$$
$$\text{where } \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{Z}_{t-1}\mathbb{Z}_{t-1}^T \pi_t(\widehat{\boldsymbol{\beta}}_0)(1-\pi_t(\widehat{\boldsymbol{\beta}}_0))$$

and $S_k(\boldsymbol{\beta})$ is as in (1.3.1). Since $\widehat{\boldsymbol{\Sigma}}$ consistently estimates $\boldsymbol{\Sigma} = \mathbb{E}(\mathbb{Z}_{t-1}\mathbb{Z}_{t-1}^T\pi_t(\boldsymbol{\beta}_0)(1-\pi_t(\boldsymbol{\beta}_0))^T$ under the null hypothesis, by Theorem 1.3.1 as well as Theorem 1.2.1 the asymptotic null



Null hypothesis | Alternative: $\boldsymbol{\beta}_2 = (1, -2)$ | Alternative: $\boldsymbol{\beta}_2 = (1, -1)$ | Alternative: $\boldsymbol{\beta}_2 = (2, -1)$

Figure 1.1: Sample paths for the BAR(1)-model, $\boldsymbol{\beta}_1 = (2, -2)$, $k_0 = 250$, $n = 500$.

distribution of this statistic is given by

$$\sup_{0 \leqslant t \leqslant 1} \left(B_1^2(t) + B_2^2(t)\right) \tag{1.5.1}$$

for two independent Brownian bridges $\{B_1(\cdot)\}$ and $\{B_2(\cdot)\}$ with a simulated 95% quantile of 2.53. Table 1.1 reports the empirical size and power (based on 10 000 repetitions) for various alternatives, where a change always occurred at $n/2$. Figure 1.5.1 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. Similarly, the power is quite good, further increasing with increasing sample size, where as always some alternatives are easier detected than others.

Table 1.1: Empirical size and power of binary autoregressive model with $\boldsymbol{\beta}_1 = (2, -2)$ (parameter before the change)

	H_0			$H_1 : \boldsymbol{\beta}_2 = (1, -2)$			$H_1 : \boldsymbol{\beta}_2 = (1, -1)$			$H_1 : \boldsymbol{\beta}_2 = (2, -1)$		
n	200	500	1000	200	500	1000	200	500	1000	200	500	1000
	0.032	0.040	0.044	0.650	0.985	1.00	0.176	0.520	0.871	0.573	0.961	0.999

Data analysis: US recession data

We now apply the above test statistic to the quarterly recession data (confer Figure 1.2) from the USA for the period 1855–2012[1]. The data 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

[1]This data set can be downloaded from the National Bureau of Economic Research at `http://research.stlouisfed.org/fred2/series/USREC`.

by different authors in particular it has recently been analyzed in a change point context by Hudeková [11].

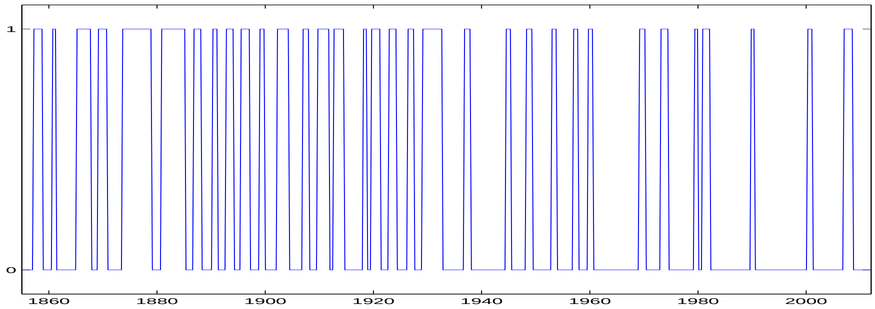


Figure 1.2: Quarterly US Recession data (1855–2012)

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 US and lead 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 result can be obtained. This is consistent with the findings in Hudeková [11], who applied a different statistic based on a binary autoregressive time series of order 3.

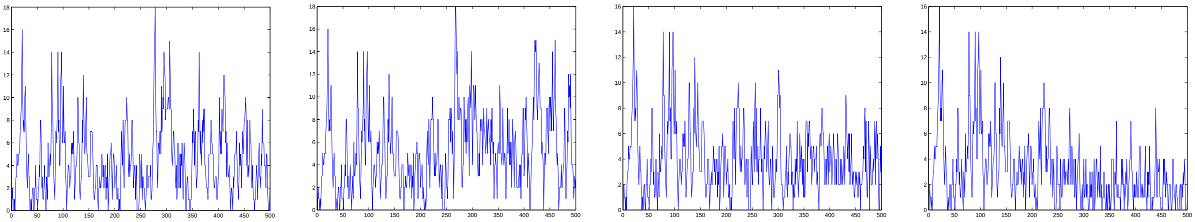
1.5.2 Poisson autoregressive models

In this section we consider a Poisson autoregressive model as in (1.4.1) with $\lambda_t = \gamma_1 + \gamma_2 Y_{t-1}$. For this model, we use the following test statistic based on least-squares scores

$$T_n = \max_{1 \leqslant k \leqslant n} \frac{1}{n} S_k(\gamma)^T \widehat{\boldsymbol{\Sigma}}^{-1} S_k(\gamma),$$

$$\text{where } S_k(\gamma) = \sum_{t=1}^{k} \mathbb{Y}_{t-1}(Y_t - \lambda_t), \quad \mathbb{Y}_{t-1} = (1, Y_{t-1})^T$$

and $\widehat{\boldsymbol{\Sigma}}^{-1}$ is the empirical covariance matrix of $\{\mathbb{Y}_{t-1}(Y_t - \lambda_t)\}$. By Theorem 1.4.1 b) and 1.2.1 this statistic has the same null asymptotics as in (1.5.1). Table 1.2 reports the empirical size and power (based on 10 000 repetitions) for various alternatives, where a change always occurred at $n/2$. Figure 1.5.2 shows one corresponding sample path for each scenario. The size is always conservative and gets closer to the nominal level with increasing sample size as predicted by the asymptotic results. Similarly, the power is quite good, further increasing



Null hypothesis   Alternative: $\gamma_2 = (2, 0.75)$ Alternative: $\gamma_2 = (2, 0.5)$ Alternative: $\gamma_2 = (1, 0.5)$

Figure 1.3: Sample paths for the Poisson autoregressive model, $\gamma_1 = (1, 0.75)$, $k_0 = 100$, $n = 200$.

with increasing sample size, where as always some alternatives are easier detected than others. Some first simulations suggest that using the statistic associated with the maximum-likelihood-scores can further increase the power. While a detailed theoretic analysis can in principle be done based on the results in Section 1.2, it is beyond the scope of this work.

Table 1.2: Empirical size and power of Poisson autoregressive model with $\gamma_1 = (1, 0.75)$ (parameter before the change)

	H_0			$H_1 : \gamma_2 = (2, 0.75)$			$H_1 : \gamma_2 = (2, 0.5)$			$H_1 : \gamma_2 = (1, 0.5)$		
n	200	500	1000	200	500	1000	200	500	1000	200	500	1000
	0.028	0.0361	0.036	0.531	0.967	0.999	0.252	0.683	0.968	0.271	0.895	0.999

Data analysis: Number of transactions per minute for Ericsson B stock

In this section we use the above methods to analyze the data set that consists of the number of transactions per minute for the stock Ericsson B during July 3rd 2002. The data set consists of 460 observations instead of the 480 for the eight hours of transactions because the first five minutes and last 15 minutes of transactions are not taken into account. Fokianos et al. [8] have analyzed the transactions 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 procedure as described below Theorem 1.2.2) are illustrated in Figure 1.4.

In fact, the empirical autocorrelation function of the complete time series in Figure 1.5 decreases very slowly, which can either be taken as evidence of long-range dependence or the presence of change points. Figure 1.6 shows the empirical autocorrelation function of the data in each segment supporting the conjecture that change points rather than long range dependence cause the slow decrease of the empirical autocorrelation function of the full data set.

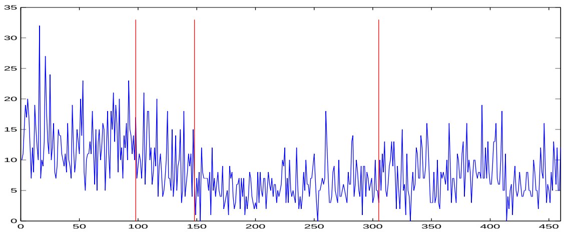


Figure 1.4: Transactions per minute for the stock Ericsson B on July 3rd, 2002, where the red vertical lines are the estimated change points

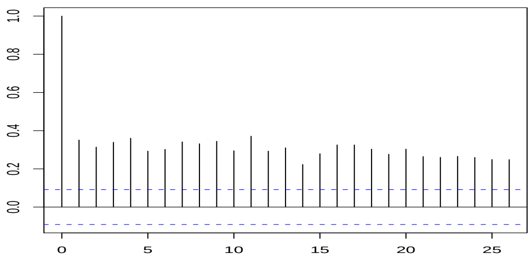
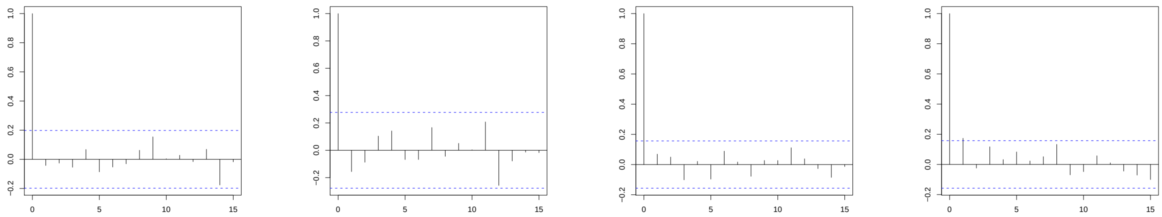


Figure 1.5: Sample autocorrelation function



1st Segment 2nd Segment 3rd Segment 4th Segment

Figure 1.6: The empirical autocorrelation of the segmented data, taking into account the estimated change points.

1.6 Appendix: Proofs

Proof of Theorem 1.2.1. By Assumption A.1 we can replace $S(k)$ 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 (1.2.1)

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

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 (1.2.1) that

$$\max_{1\leqslant k\leqslant \tau n} \frac{w(k/n)}{n}(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0))\Sigma^{-1}(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0))$$
$$\leqslant \sup_{1-\tau\leqslant t<1} w(t)t^\alpha \|\Sigma^{-1/2}\| \max_{1\leqslant k\leqslant n} \frac{1}{n^{1-\alpha}\, k^\alpha}\|S(k;\theta_0)\|^2 \overset{P}{to} 0$$

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

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

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

By the invariance principle in addition to Theorem 2.1.4 in Schmitz [20] (the theorem there is for the univariate case but the rates immediately carry over to the multivariate situation here) we also get

$$\max_{n/\log n\leqslant k\leqslant n/2} \frac{n}{k(n-k)} \left(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)\right)^T \Sigma^{-1} \left(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)\right) = o_P\left(\left(\frac{b_d(\log n)}{a(\log n)}\right)^2\right).$$

The invariance principle in combination with Horvath [10], Lemma 2.2 (in addition to analogous arguments as above), imply that

$$P\left(a(\log n) \max_{\log n\leqslant k\leqslant n/\log n} \sqrt{\frac{n}{k(n-k)}\, S(k)^T\Sigma^{-1}S(k)} - b_d(\log n) \leqslant t\right) \to \exp(-e^{-t}).$$

By Assumption A.2 the exact same arguments lead to analogous assertions for $k \geqslant n/2$, which imply the assertion by the asymptotic independence guaranteed by Assumption A.3.

From this, we also get that $\left\| \sqrt{\frac{n}{k(n-k)}}(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)) \right\| = O_P(\sqrt{\log\log n})$, which implies

$$\begin{aligned}
&\sqrt{\log\log n}\sqrt{\frac{n}{k(n-k)}\left(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)\right)^T \Sigma^{-1}\left(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)\right)} \\
&- \sqrt{\log\log n}\sqrt{\frac{n}{k(n-k)}\left(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)\right)^T \widehat{\Sigma}^{-1}\left(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)\right)} \\
&\leqslant \sqrt{\log\log n}\left\|(\Sigma^{-1/2} - \widehat{\Sigma}^{-1/2})\sqrt{\frac{n}{k(n-k)}}\left(S(k;\theta_0) - \frac{k}{n}S(n;\theta_0)\right)\right\| \\
&= O_P(\log\log n)\left\|\Sigma^{-1/2} - \widehat{\Sigma}^{-1/2}\right\| = o_P(1),
\end{aligned}$$

showing that the statistic with estimated covariance matrix yields the same asymptotics. □

Proof of Proposition 1.2.1. 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 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 $|x_{\alpha(n)} - x_1| \geqslant \varepsilon$. But then since x_1 is a unique zero in the strict sense it holds $\|G(x_{\alpha(n)})\| \geqslant \delta$ for some $\delta > 0$. However, this is a contradiction as

$$\|G(x_{\alpha(n)})\| = \|G(x_{\alpha(n)}) - G_{\alpha(n)}(x_{\alpha(n)})\| \leqslant \sup_x \|G_{\alpha(n)}(x) - G(x)\| \to 0.$$

□

Proof of Theorem 1.2.2. B.1, B.3, B.4 and B.5 (i) imply

$$\frac{1}{n^2}S(k_0)\Sigma^{-1}S(k_0) \geqslant c + o_P(1),$$

hence

$$\begin{aligned}
&\max_{1\leqslant k\leqslant n} \frac{w(k/n)}{n} S(k)^T\Sigma^{-1}S(k) \geqslant nw(\lambda)(c + o_P(1)) \to \infty, \\
&\frac{a(\log n)}{b_d(\log n)} \max_{1\leqslant k\leqslant n} \frac{n}{k(n-k)} S(k)^T\Sigma^{-1}S(k) \geqslant \frac{a(\log n)}{b_d(\log n)} n \frac{1}{\lambda(1-\lambda)}(c + o_P(1)) \to \infty,
\end{aligned}$$

which implies assertion a). If additionally, B.6 holds, we get the assertion for the maximum-type statistic analogously if we replace k_0 above by $\lfloor \vartheta n \rfloor$ with $w(\vartheta) > 0$. For the sum-type statistics we get similarly

$$\frac{1}{n}\sum_{j=1}^{n} w(k/n)\frac{1}{n}S(k)^T\Sigma^{-1}S(k) = nc\left(\int_0^1 w(t)g^2(t)\,dt + o_P(1)\right) \to \infty,$$

since the assumptions on $w(\cdot)$ guarantee the existence of $\int_0^1 w(t)g^2(t)\,dt$ and it holds $\int_0^1 w(t)g^2(t)\,dt \neq 0$. The second assertion follows by standard arguments since λ is the unique maximizer of the continuous function g and by Assumptions B.1 and B.6 it holds

$$\sup_{0\leqslant t\leqslant 1}\left|\frac{1}{n^2}S(\lfloor nt\rfloor)^T\Sigma^{-1}S(\lfloor nt\rfloor) - F_\lambda(\theta_1)\Sigma^{-1}F_\lambda(\theta_1)\,g^2(t)\right| \to 0,$$

where B.5 guarantees that the limit is not zero. The proofs show that the assertions remain true if Σ is replaced by $\widehat{\Sigma}_n$ under the stated assumptions. □

Proof of Theorem 1.3.1. Assumption A.1 can be obtained by a Taylor expansion, the ergodic theorem in addition to the $\sqrt{n}$-consistency of the estimator $\widehat{\boldsymbol{\beta}}_n$. The arguments are given in detail in Fokianos et al. [7] (Proof of Proposition 3), where by the stationarity of $\{Z_t\}$ their arguments go through in our slightly more general situation for $k \leqslant n/2$, for $k > n/2$ analogous arguments give the assertion on noting that (with the notation of Fokianos et al. [7])

$$\begin{aligned}
&\sum_{t=1}^{k} Z_{t-i}^{(i)}Z_{t-1}^{(j)}\pi_t(\beta)(1-\pi_t(\beta)) - \frac{k}{n}\sum_{t=1}^{n} Z_{t-i}^{(i)}Z_{t-1}^{(j)}\pi_t(\beta)(1-\pi_t(\beta)) \\
&= -\sum_{t=k+1}^{n} Z_{t-i}^{(i)}Z_{t-1}^{(j)}\pi_t(\beta)(1-\pi_t(\beta)) + \frac{n-k}{n}\sum_{t=1}^{n} Z_{t-i}^{(i)}Z_{t-1}^{(j)}\pi_t(\beta)(1-\pi_t(\beta)).
\end{aligned}$$

Assumption A.3 (i) follows from the strong invariance principle in Proposition 2 of Fokianos et al. [7]. Assumption A. 3 (ii) does not follow by the same proof technics 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 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 [16], Theorem 4. The mixing assumption then also implies A.3 (iii). □

Proof of Theorem 1.3.2. First note that

$$\sup_{\theta\in\Theta} \|S_k(\boldsymbol{\beta}) - \mathbb{E}S_k(\boldsymbol{\beta})\| = o_p(n) \tag{1.6.1}$$

uniformly in $k \leqslant k_0 = \lfloor n\lambda \rfloor$ by the uniform ergodic theorem of Ranga Rao [19], Theorem 6.5. For $k > k_0$ it holds

$$\begin{aligned}
&\mathbb{Z}_k Y_k = \tilde{\mathbb{Z}}_k \tilde{Y}_k + \tilde{\mathbb{Z}}_k R_1(t) + \tilde{Y}_k \mathbb{R}_2(t) + R_1(t)\mathbb{R}_2(t),\\
&\mathbb{Z}_k \pi_k(\boldsymbol{\beta}) = \tilde{\mathbb{Z}}_k \pi_k(\boldsymbol{\beta}) + \mathbb{R}_1(k)\pi_k(\boldsymbol{\beta}) = \tilde{\mathbb{Z}}_k \tilde{\pi}_k(\boldsymbol{\beta}) + O(\tilde{\mathbb{Z}}_k \boldsymbol{\beta}^T \mathbb{R}_1(k)) + O(\mathbb{R}_1(k)),
\end{aligned}$$

where $g(\tilde{\pi}_k(\boldsymbol{\beta})) = \boldsymbol{\beta}^T \tilde{\mathbb{Z}}_{t-1}$ and the last line follows from the mean value theorem. An application of the Cauchy-Schwarz inequality together with (iii) and the compactness of Θ shows that the remainder terms are asymptotically negligible, implying

$$\sup_{\boldsymbol{\beta}} \|S_k(\boldsymbol{\beta}) - S_{k_0}(\boldsymbol{\beta}) - \mathbb{E}\tilde{S}_k(\boldsymbol{\beta}) - \mathbb{E}S_{k_0}(\boldsymbol{\beta})\| = o_p(n)$$

uniformly in $k > k_0$, where

$$\tilde{S}_k(\boldsymbol{\beta}) = \sum_{t=1}^{\min(k_0,k)} \mathbb{Z}_{t-1}(Y_t - \pi_t(\boldsymbol{\beta})) + \sum_{t=k_0+1}^{k} \tilde{\mathbb{Z}}_{t-1}(\tilde{Y}_t - \tilde{\pi}_t(\boldsymbol{\beta}))$$

Together with (1.6.1) this implies B.1 with

$$F_t(\lambda) = \min(t,\lambda)\, \mathbb{E}\mathbb{Z}_0(Y_1 - \pi_1(\boldsymbol{\beta})) + (t-\lambda)_+\, \mathbb{E}\tilde{\mathbb{Z}}_{n-1}(\tilde{Y}_n - \tilde{\pi}_n(\boldsymbol{\beta})).$$

Since $F_t(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$, B.2 follows from (iv). B.6 follows since by definition of $\boldsymbol{\beta}_1$ it holds $\mathbb{E}\tilde{\mathbb{Z}}_{n-1}(\tilde{Y}_n - \tilde{\pi}_n(\boldsymbol{\beta}_1)) = -\lambda/(1-\lambda)\mathbb{E}\mathbb{Z}_0(Y_1 - \pi_1(\boldsymbol{\beta}_1))$. □

Proof of Theorem 1.4.1. It suffices to show that the assumptions of Theorem 1.2.1 are fulfilled. Assumption A.1 follows for a) as well as b) analogously to the proof of Lemma 1 in Franke et al. [9]. The invariance principles in A.2 and A.3 then follow from the strong mixing assumption and the invariance principle of Kuelbs and Philipp [16]. The asymptotic independence of A.3 also follows from the mixing condition. □

Proof of Theorem 1.4.2. This is analogous to the proof of Theorem 1.3.2 where the mean value theorem is replaced by the Lipschitz assumption, which also implies $|f_\gamma(\mathbf{x})| \leqslant |f_\gamma(0) + \mathbf{x}|$. □

Acknowledgments

This position of the first author was financed by the Stifterverband für die Deutsche Wissenschaft by funds of the Claussen-Simon-trust. The work was partly supported by DFG grant KI 1443/2-2.

Bibliography

[1] Chu, C.-S.J., Stinchcombe, M., and White, H. Monitoring structural change. *Econometrica*, 64:1045–1065, 1996.

[2] Csörgő, M. and Horváth, L. *Limit Theorems in Change-Point Analysis.* Wiley, Chichester, 1997.

[3] Doukhan, P. and Kegne, W. Inference and testing for structural change in time series of counts model. *arXiv:1305.1751*, 2013.

[4] Feigin, P.D. and Tweedie, R.L. Random coefficient autoregressive processes: A Markov chain analysis of stationarity and finiteness of moments. *J. Time Ser. Anal.*, 6:1–14, 1985.

[5] Fokianos, K. and Fried, R. Interventions in INGARCH processes. *J. Time Ser. Anal.*, 31:210–225, 2010.

[6] Fokianos, K. and Fried, R. Interventions in log-linear poisson autoregression. *Stat. Model.*, 12:299–322, 2012.

[7] Fokianos, K., Gombay, E., and Hussein, A. Retrospective change detection for binary time series models. *J. Statist. Plann. Inf*, 2014. To appear.

[8] Fokianos, K., Rahbek, A., and Tjostheim, D. Poisson autoregression. *J. Amer. Statist. Assoc.*, 104:1430–1439, 2009.

[9] Franke, J., Kirch, C., and Tadjuidje Kamgaing, J. Changepoints in times series of counts. *J. Time Ser. Anal.*, 33:757–770, 2012.

[10] Horváth , L. Change in autoregressive processes. *Stochastic Process. Appl.*, 44:221–242, 1993.

[11] Hudeková, S. Structural changes in autoregressive models for binary time series. *J. Statist. Plann. Inf.*, 143(10), 2013.

[12] Kauppi, H., and Saikkonen, P. Predicting US recessions with dynamic binary response models. *Review of Economics and Statistics*, 90:777–791, 2008.

[13] Kirch, C. *Resampling Methods for the Change Analysis of Dependent Data.* PhD thesis, University of Cologne, Cologne, 2006. http://kups.ub.uni-koeln.de/volltexte/2006/1795/.

[14] Kirch, C. and Tadjuidje Kamgaing, J. Geometric ergodicity of binary autoregressive models with exogenous variables. 2013. http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:hbz:386-kluedo-36475. Preprint.

[15] Kirch, C. and Tadjuidje Kamgaing, J. Monitoring time series based on estimating functions. 2014+. In Preparation.

[16] Kuelbs, J., and Philipp, W. Almost sure invariance principles for partial sums of mixing B-valued random variables. *Ann. Probab.*, 8:1003–1036, 1980.

[17] Neumann, M. H. Absolute regularity and ergodicity of Poisson count processes. *Bernoulli*, 17(4):1268–1284, 2011.

[18] Philipp, W. and Stout, W. *Almost sure invariance principles for partial sums of weakly dependent random variables.* Memoirs of the American Mathematical Society 161. American Mathematical Society, Providence, RI, 1975.

[19] Ranga Rao, R. Relation between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, 33:659–680, 1962.

[20] Schmitz, A. *Limit Theorems in Change-Point Analysis for Dependent Data.* PhD thesis, University of Cologne, Germany, 2011. http://kups.ub.uni-koeln.de/4224/.

[21] Startz, R. Binomial autoregressive moving average models with an application to US recession. *J. Bus. Econom. Statist.*, 26:1–8, 2008.

[22] Vostrikova, L.Y. Detection of 'disorder' in multidimensional random processes. *Sov. Math. Dokl.*, 24:55–59, 1981.

[23] Wang, C. and Li, W. K. On the autopersistence functions and the autopersistence graphs of binary autoregressive time series. *J. Time Ser. Anal.*, 32(6):639–646, 2011.

[24] Weiss, C.H. Detecting mean increases in poisson INAR(1) processes with EWMA control charts. *J. Applied Statist.*, 38:383–398, 2011.

[25] Weiß, C.H. and Testik, M.C. The Poisson INAR(1) CUSUM chart under overdispersion and estimation error. *IIE Transactions*, 43(11):805–818, 2011.

[26] Wilks, D., and Wilby, R. The weather generation gamea review of stochastic weather models. *Progress in Physical Geography*, 23:329–357, 1999.

[27] Yontay, P., Weiß, C.H., Testik, M.C., and Bayindir, Z.P. A two-sided cumulative sum chart for first-order integer-valued autoregressive processes of Poisson counts. *Qual. Reliab. Engng. Int.*, 29:33–42, 2012.