Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 6013–6026
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Detecting change-points in Markov chains

Alan M. Polansky*

Division of Statistics, Northern Illinois University De Kalb, IL 60115, USA

Received 17 February 2005; received in revised form 3 August 2006; accepted 30 November 2006
Available online 4 January 2007

Abstract

Markov chains provide a flexible model for dependent random variables with applications in such disciplines as physics, environmental science and economics. In the applied study of Markov chains, it may be of interest to assess whether the transition probability matrix changes during an observed realization of the process. If such changes occur, it would be of interest to estimate the transitions where the changes take place and the probability transition matrix before and after each change. For the case when the number of changes is known, standard likelihood theory is developed to address this problem. The bootstrap is used to aid in the computation of p-values. When the number of changes is unknown, the AIC and BIC measures are used for model selection. The proposed methods are studied empirically and are applied to example sets of data.

Keywords: AIC; BIC; Bootstrap; Dependence; Likelihood; p-Value

1. Introduction

Let $\{X_i\}_{i=0}^{\infty}$ be a sequence of discrete random variables, each with finite support S_c. The set S_c can contain any finite collection of distinct elements, but for simplicity this paper assumes that $S_c = \{1, 2, \ldots, c\}$ where c is a positive finite integer. In the context of stochastic processes, the index i is generally a time index and the value of X_i is called the state of the process at time i. Let $\{P_i\}_{i=0}^{\infty}$ be a sequence of $c \times c$ matrices where P_i has (j, k)th element $0 \leqslant P_i(j, k) \leqslant 1$ and

$$\sum_{k=1}^{c} P_i(j, k) = 1 \tag{1}$$

for each $i \in \mathbb{N}$ and $j \in S_c$. The jth row of P_i is the conditional probability distribution of X_{i+1} given $X_i = j$. Hence $P(X_{i+1} = k | X_i = j) = P_i(j, k)$ for all $i \in \mathbb{N}$, $j \in S_c$ and $k \in S_c$. The processes studied in this paper are assumed to have the Markov property, which implies

$$P(X_{i+1} = k | X_i = j, X_{i-1} = x_{i-1}, \ldots, X_0 = x_0) = P(X_{i+1} = k | X_i = j) = P_i(j, k)$$

for all sequences of constants $\{x_m\}_{m=0}^{i-1}$ where each $x_m \in S_c$.

* Tel.: +1 815 7536847; fax: +1 815 7536776.
E-mail address: polansky@math.niu.edu.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.11.040

If $P_i = P$ for all $i \in \mathbb{N}$, then the process $\{X_i\}_{i=0}^{\infty}$ is known as a time homogeneous Markov chain with finite state space S_c. Such processes are commonly used to model dependent processes observed in such disciplines as biology, computer science, environmental science, geography, social science, physics and economics. See, for example, Clark (1965), Fuh (1993), Geary (1978), Gottschau (1992), Turchin (1986) and Yang (1979). In the applied setting, the transition probability matrix P is unknown and a finite realization $X_0, \ldots, X_n$ of the process is observed in order to estimate and test hypotheses about P. A large amount of research has been devoted to these problems. See Basawa and Rao (1980) and Billingsley (1961a, b) for an overview of much of the relevant research.

A less restrictive model than the Markov chain model described above allows the transition probability matrix to change during the observed realization. Therefore, in this model there is a sequence of transition probability matrices $T_0, \ldots, T_k$ and positive integers $0 = \psi_0 < \psi_1 < \cdots < \psi_k < \psi_{k+1} = n$ such that $P_j = T_i$ for $j = \psi_i, \ldots, (\psi_{i+1} - 1)$ and $i = 1, \ldots, k$. The points $\psi_1, \ldots, \psi_k$ are known as change-points. In the applied setting such a model may be reasonable when the researcher believes that outside interventions may have caused the behavior of the process to change. The change-points $\psi_1, \ldots, \psi_k$ may be known if the time of the interventions is known. In this situation it is useful to estimate $T_1, \ldots, T_k$ as well as test a null hypothesis of the form $\mathrm{H}_0 : T_0 = \cdots = T_k$ versus the alternative hypothesis $\mathrm{H}_1 : T_i \neq T_j$ for at least one $i \neq j$. This allows the researcher to investigate whether the interventions result in significant changes in the process behavior. In other cases the change-points, and perhaps the number of change-points, may also be unknown. In such cases, the estimation of these additional parameters is also of great importance.

Bayesian methods for solving the single change-point problem have been investigated by Maltsev and Silaev (1992), Silaev (1997) and Yakir (1994). Typically these approaches concentrated on optimal stopping strategies for the detection of a single change in the transition probability matrix for real-time observations from an on-line process. This paper emphasizes the detection of change-points for fixed sample sizes. A test for homogeneity for general discrete time Markov sequences, where the sequence of parameters under the alternative hypothesis is a sequence of independent and identically distributed random variables, has been studied by Ramanathan and Rajarshi (1997). In the current paper it is assumed that the parameter changes are deterministic. If multiple samples are available, then the method developed by Anderson and Goodman (1957), and Madansky (1963), can be implemented to detect if at least one change-point occurs in a Markov chain. The methods of the current paper specifically focus on the problem of a single observed realization and ultimately considers the problem of determining the number of change-points in an observed sequence. Some methods in this paper rely on asymptotic likelihood theory. A general review of asymptotic results for change-point problems is given by Csörgö and Horváth (1997).

This paper develops methods for detecting and estimating change-points in discrete-time Markov chains. Three situations are addressed. A likelihood ratio test of the equality of the transition probability matrices is developed for the case when the number and the location of the change-points are known. The p-value for this test is based on asymptotic χ^2 theory. When the number of change-points is known, but the locations are unknown, maximum likelihood estimation is used to estimate the location of the change-points. A likelihood ratio test is developed to test the equality of the transition probability matrices. In this case the p-values must be approximated using the bootstrap. When the number of change-points is unknown, methods based on AIC and BIC are developed to estimate the number of change-points. The paper is organized as follows. Section 2 develops the statistical methodology for the detection of change-points. Section 3 provides empirical evidence that the proposed methods appear to have reasonable behavior. Section 4 applies the proposed methodology to two examples of observed realizations from Markov chains. A general discussion of the methodology is given in Section 5.

2. Change-point detection

Suppose $X_0, \ldots, X_n$ is a realization of length $n+1$ from a Markov chain with state space S_c and transition probability matrix P. Let

$$\Omega_{ij} = \sum_{k=0}^{n-1} \delta(X_{k+1} = j, X_k = i)$$

and

$$\Omega_{i\cdot} = \sum_{k=0}^{n-1} \delta(X_k = i),$$

where δ is the indicator function. Bartlett (1951) has shown that the maximum likelihood estimator of P is given by $\hat{P}$ whose (i, j)th element is

$$\hat{P}(i, j) = \begin{cases} \Omega_{ij}/\Omega_{i\cdot} & \text{if } \Omega_{i\cdot} > 0, \\ \delta_{ij} & \text{if } \Omega_{i\cdot} = 0, \end{cases} \tag{2}$$

where $\delta_{ij} = \delta(i = j)$. The assignment of the elements in row i of $\hat{P}$ are somewhat arbitrary when $\Omega_{i\cdot} = 0$, and other conventions have been suggested than what is used in Eq. (2). For example, another common convention is to set $\hat{P}(i, j) = 0$ for $j = 1, \ldots, c$ when $\Omega_{i\cdot} = 0$. However, the estimates of the transition probability matrices must follow the restriction of Eq. (1) in order to implement the bootstrap algorithms proposed in this paper. Hence, the estimate given in Eq. (2) is used for the remainder of the paper. The properties of the estimate given in Eq. (2) are very closely related to the multinomial distribution, which provides the basis for studying the theoretical properties of the estimator. In particular, $\hat{P}$ is consistent and asymptotically normal. For additional properties, see Bartlett (1951), Billingsley (1961a, b) and Basawa and Rao (1980).

To motivate later development, consider the simplest case where the sequence $X_0, \ldots, X_n$ has a single change-point ψ_1. If ψ_1 is known, it can be shown that the maximum likelihood estimators of T_0 and T_1 are computed by applying the Bartlett estimator of Eq. (2) to the sequences of observations $X_0, \ldots, X_{\psi_1}$ and $X_{\psi_1}, \ldots, X_n$, respectively. Let

$$\Omega_{ij}^{(a,b)} = \sum_{k=a}^{b-1} \delta(X_{k+1} = j, X_k = i)$$

and

$$\Omega_{i\cdot}^{(a,b)} = \sum_{k=a}^{b-1} \delta(X_k = i),$$

where $a < b$ are non-negative integers that do not exceed n. Then the maximum likelihood estimator of T_0 is given by $\hat{T}_0$ whose (i, j)th element is

$$\hat{T}_0(i, j) = \begin{cases} \Omega_{ij}^{(0,\psi_1)}/\Omega_{i\cdot}^{(0,\psi_1)} & \text{if } \Omega_{i\cdot}^{(0,\psi_1)} > 0, \\ \delta_{ij} & \text{if } \Omega_{i\cdot}^{(0,\psi_1)} = 0. \end{cases} \tag{3}$$

Similarly, the maximum likelihood estimator of T_1 is given by $\hat{T}_1$ whose (i, j)th element is

$$\hat{T}_1(i, j) = \begin{cases} \Omega_{ij}^{(\psi_1,n)}/\Omega_{i\cdot}^{(\psi_1,n)} & \text{if } \Omega_{i\cdot}^{(\psi_1,n)} > 0, \\ \delta_{ij} & \text{if } \Omega_{i\cdot}^{(\psi_1,n)} = 0. \end{cases} \tag{4}$$

To develop a test of $\mathrm{H}_0 : T_0 = T_1$ versus $\mathrm{H}_1 : T_0 \neq T_1$ let Δ be a metric on the space of $c \times c$ transition probability matrices. Then a size-α test of $\mathrm{H}_0 : T_0 = T_1$ rejects H_0 when $\Delta(\hat{T}_0, \hat{T}_1) > \delta_\alpha$, where δ_α is the solution of the equation

$$\sup_{T \in \mathscr{T}} P(\Delta(\hat{T}_0, \hat{T}_1) > \xi | T_0 = T_1 = T) = \alpha$$

with respect to ξ and $\mathscr{T}$ is the collection of all $c \times c$ transition probability matrices. Some distance measures for probability vectors were studied by Cressie and Read (1984). A convenient distance measure for this case is provided by the log of the likelihood ratio test statistic for testing $\mathrm{H}_0 : T_0 = T_1$ versus $\mathrm{H}_1 : T_0 \neq T_1$, which is given by

$$\Lambda = -2(L(0, \psi_1) + L(\psi_1, n) - L(0, n)), \tag{5}$$

where

$$L(a,b) = \sum_{(i,j)\in\mathscr{I}(a,b)} \Omega_{ij}^{(a,b)} \log(\Omega_{ij}^{(a,b)}/\Omega_{i\cdot}^{(a,b)}),$$

and the collection $\mathscr{I}(a,b)$ is understood to contain all indices (i,j) such that $\Omega_{ij}^{(a,b)} > 0$. See Billingsley (1961b, Chapter 5). We assume that the distribution of the initial state of the Markov chain has all of its mass concentrated at $X_0 = x_0$, the observed initial state. Even if this is not the case, the asymptotic likelihood theory is still valid as the initial distribution adds only a single finite term to the likelihood functions, with Probability 1. See Billingsley (1961b, p. 4) or Basawa and Rao (1980, p. 54). Standard asymptotic test theory can be used to show that $\Lambda \xrightarrow{\mathrm{d}} \chi^2_{c(c-1)}$ as $n \to \infty$ and $\psi_1 \to \infty$ such that $\psi_1/n \to v \in (0,1)$. Hence, a level α test of $\mathrm{H}_0 : T_0 = T_1$ rejects the null hypothesis when $\Lambda > \chi^2_{1-\alpha;c(c-1)}$.

The justification of the asymptotic results for this problem closely follows from the development in Billingsley (1961b, Chapter 4) with some minor modifications. It is assumed that the transition probability matrices T_0 and T_1 correspond to Markov chains that have a single ergodic class and no transient states. For the sake of the asymptotic analysis it is further assumed that ψ_1 is always in the interior of the observed Markov chain as $n \to \infty$. This restriction makes it possible for asymptotic approximations to be applied to the observed sequences both before and after the change-point. Therefore, it is assumed that $\psi_1/n \to v$ as $n \to \infty$ for some $v \in (0,1)$. See Billingsley (1961b, Chapter 5) for further discussion of the assumptions required for this analysis. Under these assumptions, the likelihood function for the change-point problem closely resembles the likelihood function used for testing the homogeneity of two independent observed realizations from Markov chains. The difference turns out to be asymptotically negligible and the results outlined in Billingsley (1961b, p. 20) follow. A crucial element of the proof is the asymptotic normality established in Billingsley (1961b, Equation (4.3)). Because the random variables that occur after the change-point are *not* independent of the random variables that occur before the change-point, the result does not follow as simply as in the case of two independent realizations. However, simple arguments can be used to establish that the partial sums of the likelihood function in the realization with a change-point are a Martingale, from which the asymptotic normality can be established using Theorem 9.1 of Billingsley (1961b). The remainder of the asymptotic theory then follows.

When the change-point is unknown, the parameter ψ_1 can be added to the likelihood function as an unknown parameter. The maximum likelihood estimator for ψ_1 will generally not exist in closed form, but can be found algorithmically as

$$\hat{\psi}_1 = \arg\max\{\psi_1 \in \{1,\ldots,n-1\} : L(0,\psi_1) + L(\psi_1,n)\}, \tag{6}$$

where $L(0,\psi_1)+L(\psi_1,n)$ is the maximum observed likelihood, conditional on ψ_1. This estimator is studied empirically in Section 3. When ψ_1 is unknown, the maximum likelihood estimator of T_0 is given by $\tilde{T}_0$ whose (i,j)th element is

$$\tilde{T}_0(i,j) = \begin{cases} \Omega_{ij}^{(0,\hat{\psi}_1)}/\Omega_{i\cdot}^{(0,\hat{\psi}_1)} & \text{if } \Omega_{i\cdot}^{(0,\hat{\psi}_1)} > 0, \\ \delta_{ij} & \text{if } \Omega_{i\cdot}^{(0,\hat{\psi}_1)} = 0, \end{cases}$$

and the maximum likelihood estimator of T_1 is given by $\tilde{T}_1$ whose (i,j)th element is

$$\tilde{T}_1(i,j) = \begin{cases} \Omega_{ij}^{(\hat{\psi}_1,n)}/\Omega_{i\cdot}^{(\hat{\psi}_1,n)} & \text{if } \Omega_{i\cdot}^{(\hat{\psi}_1,n)} > 0, \\ \delta_{ij} & \text{if } \Omega_{i\cdot}^{(\hat{\psi}_1,n)} = 0. \end{cases}$$

To test the null hypothesis $\mathrm{H}_0 : T_0 = T_1$ versus the alternative hypothesis $\mathrm{H}_1 : T_0 \neq T_1$ the likelihood ratio statistic

$$\tilde{\Lambda} = -2(L(0,\hat{\psi}_1) + L(\hat{\psi}_1,n) - L(0,n)) \tag{7}$$

is used. A size-α test of $\mathrm{H}_0 : T_0 = T_1$ rejects H_0 when $\tilde{\Lambda} > \tilde{\lambda}_\alpha$, where $\tilde{\lambda}_\alpha$ is the solution to the equation

$$\sup_{T\in\mathscr{T}} P(\tilde{\Lambda} > \tilde{\lambda} | T_0 = T_1 = T) = \alpha$$

with respect to $\tilde{\lambda}$. Unfortunately, specification of $\tilde{\lambda}_\alpha$ is more difficult in this case because the standard asymptotic theory for maximum likelihood estimators is not valid for parameters that are restricted to be integers. See, for example, Dahiya (1986). In particular, the martingale argument used to establish the asymptotic χ^2 results for the case when the change-point is known is no longer valid for the case when the change-point is unknown. In this case it is convenient to estimate $\tilde{\lambda}_\alpha$ using the bootstrap methodology of Efron (1979).

To motivate the bootstrap method, consider the case where the transition probability matrix under the null hypothesis is known. In this case, a size-α test of $\mathrm{H}_0 : T_0 = T_1 = T$ rejects H_0 when $\tilde{\Lambda} > \bar{\Lambda}_\alpha$, where $\bar{\lambda}_\alpha$ is the solution to the equation

$$P(\tilde{\Lambda} > \bar{\lambda} | T_0 = T_1 = T) = \alpha$$

with respect to $\bar{\lambda}$. The value $\bar{\lambda}_\alpha$ may still be difficult to obtain analytically, but could be approximated using a simulation methodology as follows. Simulate b realizations of length n from a Markov chain with transition probability matrix T and initial state x_0, the observed initial state in the observed sample realization. For each simulated realization, compute the test statistic given in Eq. (7). Denote these as $\tilde{\Lambda}_1^*, \ldots, \tilde{\Lambda}_b^*$. Then $\bar{\lambda}_\alpha$ can the be approximated by the $1-\alpha$ sample percentile of $\tilde{\Lambda}_1^*, \ldots, \tilde{\Lambda}_b^*$. Hence $\bar{\lambda}_\alpha \approx \tilde{\Lambda}^*_{([(1-\alpha)b])}$ where $[x]$ denotes the largest integer less than or equal to x. A p-value for the test can also be approximated with

$$p = \frac{1}{b+1}\left[1 + \sum_{i=1}^{b} \delta(\tilde{\Lambda}_i^* \geqslant \tilde{\Lambda})\right]. \tag{8}$$

See Davison and Hinkley (1997, Section 4.4). In the case where the transition probability matrix T is unknown, the bootstrap methodology uses the same simulation algorithm described above, but uses the estimated transition probability matrix $\hat{T}$ given in Eq. (2) in place of the known transition probability matrix T. This general methodology has been studied theoretically by Athreya and Fuh (1992) who obtained conditions under which the bootstrap method is asymptotically valid. Additional methods for implementing the bootstrap for Markov chains were studied by Fuh (1993). This bootstrap test is compared empirically to the case where the change-point is known in Section 3.

Estimation and testing methods for the case when there are a known number of change-points extend readily from the methods for a single change-point. Suppose that the observed realization $X_0, \ldots, X_n$ has k change-points $\psi_1 < \cdots < \psi_k$. As before, define $\psi_0 = 0$ and $\psi_{k+1} = n$. If $\psi_1 < \cdots < \psi_k$ are known, it can be shown that the maximum likelihood estimator of T_m is computed by applying the Bartlett estimator of Eq. (2) to the sequence of observations $X_{\psi_m}, \ldots, X_{\psi_{m+1}}$. Hence, the maximum likelihood estimator of T_m is given by $\hat{T}_m$ whose (i, j)th element is

$$\hat{T}_m(i,j) = \begin{cases} \Omega_{ij}^{(\psi_m,\psi_{m+1})} / \Omega_{i\cdot}^{(\psi_m,\psi_{m+1})} & \text{if } \Omega_{i\cdot}^{(\psi_m,\psi_{m+1})} > 0, \\ \delta_{ij} & \text{if } \Omega_{i\cdot}^{(\psi_m,\psi_{m+1})} = 0 \end{cases}$$

for $m = 0, \ldots, k$.

To test $\mathrm{H}_0 : T_0 = T_1 = \cdots = T_k$ versus $\mathrm{H}_1 : T_i \neq T_j$ for some $i \neq j$, the test statistic is again based on the log of the likelihood ratio test statistic, given by

$$\Gamma = -2\left(\sum_{m=0}^{k} L(\psi_m, \psi_{m+1}) - L(0, n)\right).$$

Standard asymptotic test theory can be used to show that $\Gamma \xrightarrow{\mathrm{d}} \chi^2_{c(c-1)k}$ as $n \to \infty$ and $\psi_i \to \infty$ such that $|\psi_i - \psi_{i-1}| \to \infty$ for $i = 1, \ldots, k+1$. Hence, a level α test of $\mathrm{H}_0 : T_0 = T_1 = \cdots = T_k$ rejects the null hypothesis when $\Gamma > \chi^2_{1-\alpha; c(c-1)k}$.

When the change-points are unknown, the parameters $\psi_1, \ldots, \psi_k$ are added to the likelihood function as unknown parameters. The maximum likelihood estimators for $\psi_1, \ldots, \psi_k$ will generally not exist in closed form, but can be found algorithmically as

$$(\hat{\psi}_1 \cdots \hat{\psi}_k)' = \arg\max \left\{ \psi_1 < \psi_2 < \cdots < \psi_k \in \{1, \ldots, n-1\} : \sum_{m=0}^{k} L(\psi_m, \psi_{m+1}) \right\}, \tag{9}$$

where

$$\sum_{m=0}^{k} L(\psi_m, \psi_{m+1}) \tag{10}$$

is the maximum observed likelihood, conditional on $\psi_1, \ldots, \psi_k$. Hence, when $\psi_1, \ldots, \psi_k$ are unknown, the maximum likelihood estimator of T_m is given by $\tilde{T}_m$ whose (i, j)th element is

$$\tilde{T}_m(i, j) = \begin{cases} \Omega_{ij}^{(\hat{\psi}_m, \hat{\psi}_{m+1})} / \Omega_{i\cdot}^{(\hat{\psi}_m, \hat{\psi}_{m+1})} & \text{if } \Omega_{i\cdot}^{(\hat{\psi}_m, \hat{\psi}_{m+1})} > 0, \\ \delta_{ij} & \text{if } \Omega_{i\cdot}^{(\hat{\psi}_m, \hat{\psi}_{m+1})} = 0 \end{cases} \tag{11}$$

for $m = 0, \ldots, k$.

The test of $\mathrm{H}_0 : T_1 = \cdots = T_k$ versus $\mathrm{H}_1 : T_i \neq T_j$ for some $i \neq j$ for the case when $\psi_1, \ldots, \psi_k$ are unknown is developed in the same manner as with a single unknown change-point. The likelihood ratio test statistic is given by

$$\tilde{\Gamma} = -2 \left(\sum_{m=0}^{k} L(\hat{\psi}_m, \hat{\psi}_{m+1}) - L(0, n) \right), \tag{12}$$

where a size α test will reject H_0 when $\tilde{\Gamma} > \tilde{\gamma}_\alpha$. As with the case of a single change-point, the value $\tilde{\gamma}_\alpha$ is estimated using the bootstrap as follows. Simulate b realizations of length n from a Markov chain with transition probability matrix $\hat{T}$ and initial state x_0, the observed initial state from the observed sample realization. For each simulated realization, compute the test statistic given in Eq. (12). Denote these as $\tilde{\Gamma}_1^*, \ldots, \tilde{\Gamma}_b^*$. Then $\tilde{\gamma}_\alpha$ can be approximated by the $1 - \alpha$ sample percentile of $\tilde{\Gamma}_1^*, \ldots, \tilde{\Gamma}_b^*$. The p-value for the test can be approximated using the same method as in Eq. (8).

Further complications occur when the number of change-points k is also unknown. The most straightforward approach to solving this problem is to treat the parameter k as another unknown parameter in the likelihood function. However, since k controls the number of parameters fit to the observed data, and hence the dimension of the parameter space, it is reasonable to penalize models with more parameters to prevent over-fitting of the data. Otherwise, reasonable parsimonious models may be overlooked. Therefore measures such as AIC (Akaike, 1974) and BIC (Schwarz, 1978) are suggested for this purpose. The AIC objective function is given by

$$\mathrm{AIC}(k) = -2 \sum_{m=1}^{k} L(\hat{\psi}_m, \hat{\psi}_{m+1}) + 2c(c-1)(k+1),$$

where $\hat{\psi}_m$ is the maximum likelihood estimate of ψ_m conditional on k as given in Eq. (9). The AIC estimate of k is given by

$$\hat{k}_{\mathrm{A}} = \arg\min\{k \in \{0, \ldots, n\} : \mathrm{AIC}(k)\}.$$

Similarly, the BIC objective function is given by

$$\mathrm{BIC}(k) = -2 \sum_{m=1}^{k} L(\hat{\psi}_m, \hat{\psi}_{m+1}) + \ln(n)c(c-1)(k+1),$$

so that the BIC estimate of k is given by

$$\hat{k}_{\mathrm{B}} = \arg\min\{k \in \{0, \ldots, n\} : \mathrm{BIC}(k)\}.$$

The criteria are very similar with the main difference being that the BIC objective function uses a greater penalty for adding additional parameters to the model. These estimates are studied empirically in Section 3. Once the value of k is estimated using either the AIC or BIC objective functions, estimates of the change-points and the corresponding transition probability matrices can be computed using the methods given in Eqs. (9) and (11), using the value of $\hat{k}$ in place of k.

3. Empirical studies

In this section, the estimators and the tests developed in Section 2 are investigated empirically using computer based simulations. In the first study, the performance of the maximum likelihood estimator for a single change-point given in Eq. (6) is considered. To study this estimator, 1000 realizations from a Markov chain with change-point ψ_1 and

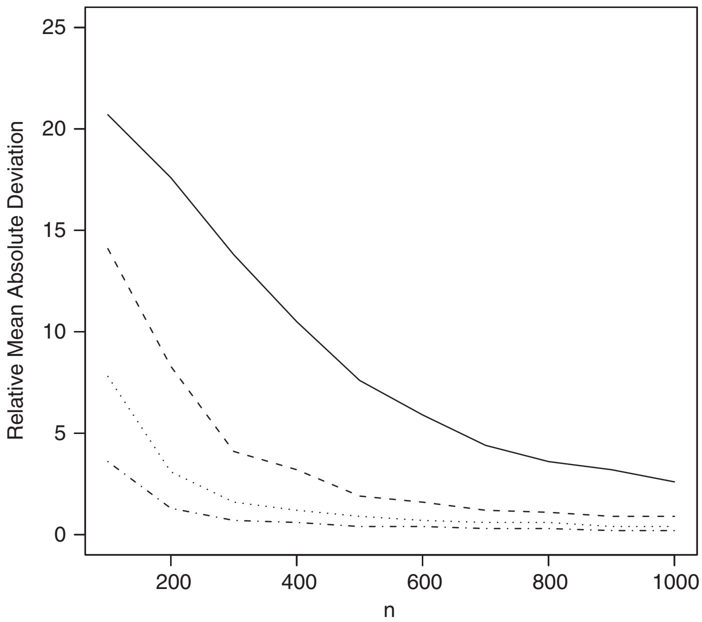

Fig. 1. Estimated relative mean absolute deviation of the estimator of a single change-point for $q_1 = 0.20$ (solid line), $q_1 = 0.15$ (dashed line), $q_1 = 0.10$ (dotted line) and $q_1 = 0.05$ (dot-dash line).

probability transition matrices T_0 and T_1 are simulated. Realizations of length $n = 100, 200, \ldots, 1000$ are considered. The change-point is located at $n/2$. Note that this is the best possible situation, as moving the change-point near either end of the observed realization essentially decreases the sample size. To simplify the study, the $c \times c$ probability transition matrices used to simulate the realizations all have the same general form with (i, j)th element

$$T(i, j) = \begin{cases} 1 - (c-1)q & \text{if } i = j, \\ q & \text{if } i \neq j, \end{cases} \tag{13}$$

where $0 < q < (c-1)^{-1}$ is a specified parameter. Denote the value of q used for the transition probability matrix T_i as q_i. For this study $c = 3$ is used and the transition probability matrix T_0 has q_0 held at 0.3 throughout the simulation while the values $0.05, 0.10, \ldots, 0.20$ are used for the value of q_1 in the transition probability matrix T_1. For each combination of simulation parameters discussed above, the change-point estimator is computed on each of the 1000 simulated realizations. The absolute difference between the estimated change-point and the true location of the change-point is then averaged over the 1000 simulated realizations to estimate $E|\hat{\psi}_1 - \psi_1|$. Because the parameter space of the change-point increases with n, the mean absolute deviation estimates are reported as a percentage of the realization length n to provide a measure of performance relative to the size of the parameter space. That is

$$\text{relative mean absolute deviation} = \frac{\widehat{E|\hat{\psi}_1 - \psi_1|}}{n} \times 100\%.$$

It has been suggested by Dahiya (1986) that the probability of correctly estimating the parameter value is a better measure of estimator performance in the case of a discrete parameter space. This measure is not used in the current investigation because the necessary conditions under which this probability is related to the mean square error of the estimator may not hold for all of the cases studied.

The results of the simulation are presented in Fig. 1. One can observe from Fig. 1 that the relative mean absolute deviation of the change-point estimator decreases as the sample size increases and as the difference between T_0 and T_1 becomes larger. This indicates that the asymptotic behavior of the change-point estimator appears to be reasonable. Note that if we used an estimator of the change-point that simply randomly picked one of the time points between 0 and n as the estimate of the change-point, independent of the data, then we would expect a relative mean absolute

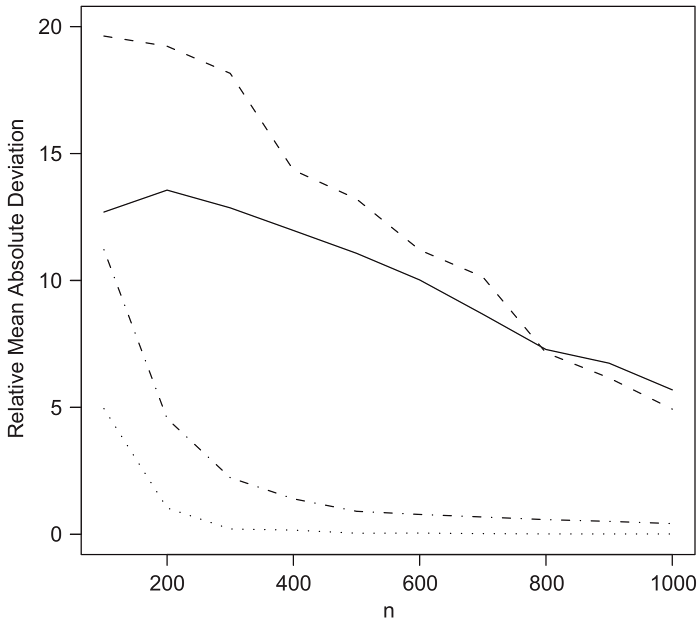

Fig. 2. Estimated relative mean absolute deviation of the estimators of two change-points for the smaller change-point with $q_0 = 0.3$, $q_1 = 0.2$ and $q_2 = 0.1$ (solid line), the larger change-point with $q_0 = 0.3$, $q_1 = 0.2$ and $q_2 = 0.1$ (dashed line), the smaller change-point with $q_0 = 0.5$, $q_1 = 0.3$ and $q_2 = 0.1$ (dotted line), the larger change-point with $q_0 = 0.5$, $q_1 = 0.3$ and $q_2 = 0.1$ (dot-dash line).

deviation of 25%. This result can aid in determining which of the results in Fig. 1 show that the proposed estimator is essentially not useful. For example, when $q_2 = 0.20$ it is apparent that a sample size of at least 200 is required before that method produces a meaningful estimator.

This study also considered processes with two change-points, located at $n/4$ and $n/2$ with corresponding transition probability matrices T_0, T_1 and T_2. Two sets of transition probability matrices are considered: $q_0 = 0.3$, $q_1 = 0.2$ and $q_2 = 0.1$ and $q_0 = 0.5$, $q_1 = 0.3$ and $q_2 = 0.1$. The results of the simulation are presented in Fig. 2. The same general behavior is observed as in the case of a single-change-point except that the estimated relative mean absolute deviations are smaller than in the single change-point case. This is due to the fact that the two change-points are being fit within the same parameter space and there is consequently less room for variation. One can also note that the estimates of the change-point located at $n/4$ have lower relative mean absolute deviations than the estimates of the change-point located at $n/2$. Again, there is less room in the parameter space for the lower change-point to be fit in below the upper change-point.

The empirical power of the tests for the null hypothesis $\mathrm{H}_0 : T_0 = T_1$ versus the alternative hypothesis $\mathrm{H}_1 : T_0 \neq T_1$ under both the conditions where ψ_1 is known and unknown are also studied. The parameters used are the same as those used in the study of the change-point estimator, except that the condition $q_0 = q_1$ is added so that the achieved significance level of the tests can be studied. The observed empirical power for the test when ψ_1 is known is given in Fig. 3 and the observed power for the test when ψ_1 is unknown is given in Fig. 5. One can observe from Figs. 3 and 5 that the power functions for both tests behave in a reasonable manner. When the null hypothesis is true ($q_1 = 0.3$), the number of rejections is close to what would be expected for a significance level of $\alpha = 0.10$. When the null hypothesis is not true, the power of both tests increases as the sample size and the difference between the null and alternative hypothesis, given by $|q_0 - q_1|$, increases. One can also observe that the test for the case when the location of the change-point is unknown is less powerful than for the case when the change-point is known. This difference decreases as the sample size and $|q_0 - q_1|$ increase.

This study also included two change-points, located at $n/4$ and $n/2$ with corresponding transition probability matrices T_0, T_1 and T_2. Four sets of transition probability matrices are considered: $q_0 = q_1 = q_2 = 0.10$, $q_0 = q_1 = 0.10$ and $q_2 = 0.15$, $q_0 = 0.10$, $q_1 = 0.15$ and $q_2 = 0.20$, and $q_0 = 0.10$, $q_1 = 0.20$ and $q_2 = 0.30$. The observed empirical power for the test when ψ_0 and ψ_1 are known is given in Fig. 4 , and the observed power for the test when ψ_0 and ψ_1 are

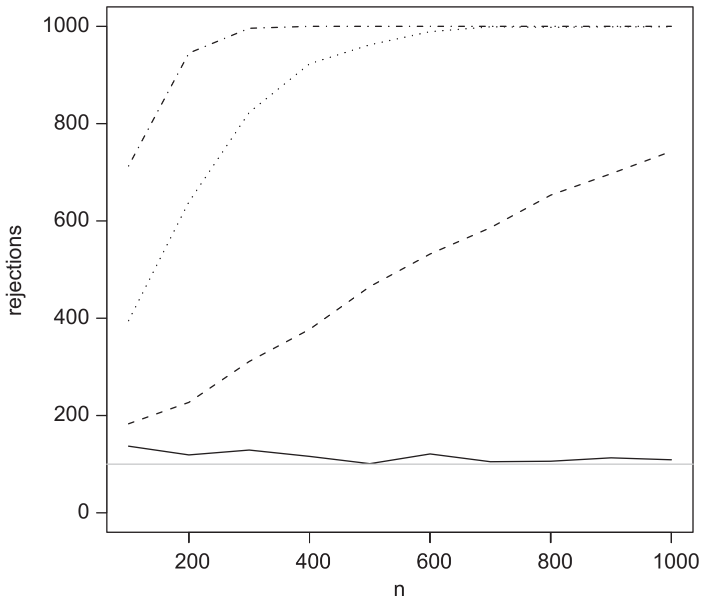

Fig. 3. The observed number of rejections in 1000 simulated realizations for the test of $H_0 : T_0 = T_1$ when ψ_1 is known and $q_1 = 0.30$ (solid line), $q_1 = 0.25$ (dashed line), $q_1 = 0.20$ (dotted line) and $q_1 = 0.15$ (dot-dash line). The specified significance level is $\alpha = 0.10$.

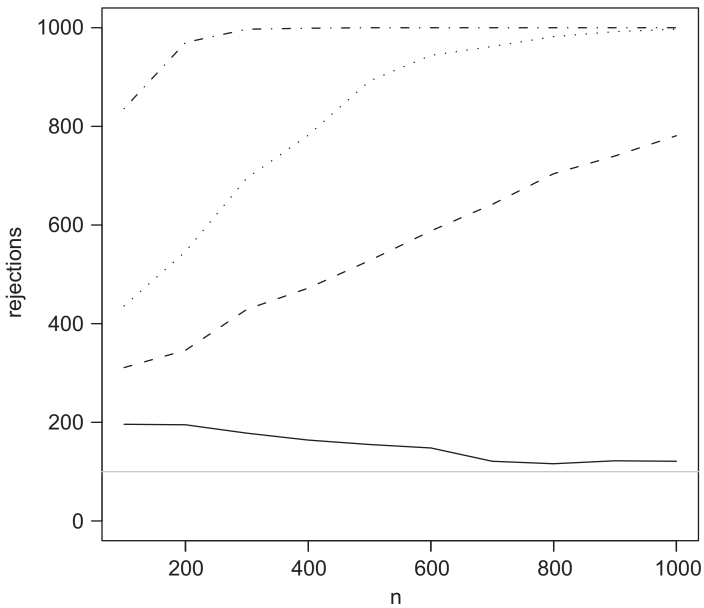

Fig. 4. The observed number of rejections in 1000 simulated realizations for the test of $H_0 : T_0 = T_1 = T_2$ when ψ_1 and ψ_2 are known and $q_0 = q_1 = q_2 = 0.10$ (solid line), $q_0 = q_1 = 0.10$ and $q_2 = 0.15$ (dashed line), $q_0 = 0.10$, $q_1 = 0.15$ and $q_2 = 0.20$ (dotted line) and $q_0 = 0.10$, $q_1 = 0.20$ and $q_2 = 0.30$ (dot-dash line). The specified significance level is $\alpha = 0.10$.

unknown is given in Fig. 6. The same general trends as with the single change-point case are observed in these results with the exception that the achieved significance level is larger than would be expected when $\alpha = 0.10$ for smaller sample sizes (Figs. 5 and 6).

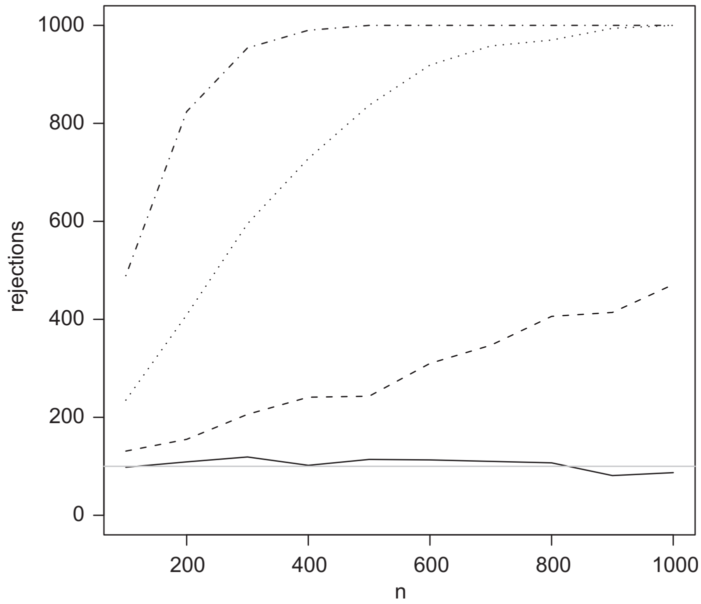

Fig. 5. The observed number of rejections in 1000 simulated realizations for the test of $H_0: T_0 = T_1$ when ψ_1 is unknown and $q_1 = 0.30$ (solid line), $q_1 = 0.25$ (dashed line), $q_1 = 0.20$ (dotted line) and $q_1 = 0.15$ (dot-dash line). The specified significance level is $\alpha = 0.10$.

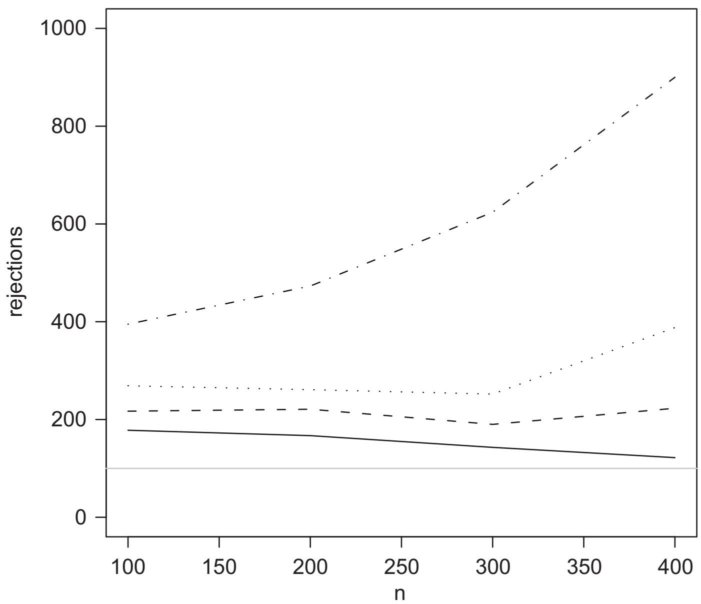

Fig. 6. The observed number of rejections in 1000 simulated realizations for the test of $H_0: T_0 = T_1 = T_2$ when ψ_1 and ψ_2 are unknown and $q_0 = q_1 = q_2 = 0.10$ (solid line), $q_0 = q_1 = 0.10$ and $q_2 = 0.15$ (dashed line), $q_0 = 0.10$, $q_1 = 0.15$ and $q_2 = 0.20$ (dotted line) and $q_0 = 0.10$, $q_1 = 0.20$ and $q_2 = 0.30$ (dot-dash line). The specified significance level is $\alpha = 0.10$.

The use of the AIC and BIC methods for estimating the number of change-points is also investigated. As in the previous studies, realizations from a Markov chain with $c = 3$ and a single change point located at $n/2$ are simulated. Two cases are considered. The first case uses $n = 200$, $q_0 = 0.3$ and $q_1 = 0.1$, and the second case uses $n = 300$, $q_0 = 0.3$ and $q_1 = 0.05$. For each case, 50 realizations are simulated. The AIC and BIC criteria are then used to estimate

Table 1
Frequency for each of the estimated values of k from 50 simulated realizations

Model	Method	Estimated k			
		0	1	2	3 or more
$n = 200, q_0 = 0.30, q_1 = 0.10$	AIC	0	3	6	41
$n = 200, q_0 = 0.30, q_1 = 0.10$	BIC	13	37	0	0
$n = 300, q_0 = 0.30, q_1 = 0.05$	AIC	0	0	2	48
$n = 300, q_0 = 0.30, q_1 = 0.05$	BIC	0	50	0	0

The true value of k in each case is 1.

the value of k. Because of the computational cost of the estimation algorithms, only the values $\hat{k} \in \{0, 1, 2, 3\}$ are considered. Since values larger than $k = 3$ are not considered, when $\hat{k} = 3$ the result is classified as $\hat{k} \geqslant 3$. The results of the simulations are given in Table 1. One can observe from Table 1 that the BIC method is clearly better suited for the problems studied here, while the AIC criterion appears to over-fit the model in almost every case. In fact, as the sample size and the difference between q_0 and q_1 increases, the performance of the AIC method becomes worse. This indicates that the AIC method produces a possibly inconsistent estimator of k. The estimate of k based on the BIC method becomes more reliable as the sample size and the difference between q_0 and q_1 increase.

4. Examples

4.1. Industrial machine usage

Consider an industrial machine that has three states of usage: not in use (State 1), normal use (State 2) and heavy use (State 3). Suppose that the usage state of the machine is observed once an hour, and that the sequence of hourly state observations follows a discrete time irreducible ergodic Markov chain. The management of the manufacturing plant that utilizes the machine is interested in avoiding State 3, which increases the maintenance and repair costs of the machine. A new work flow policy is proposed by the management at the plant. In order to test whether the new work flow policy is effective, the policy is implemented halfway through an observed 60 h manufacturing cycle. Table 2 presents a simulated sequence of hourly state observations from such a process. The observation at hour 0 indicates the initial state of the machine.

The change in work flow policy is implemented mid-cycle so that transition from the 30th to the 31st observation is the first transition observed under the new policy. Of interest is evaluating the transition probability matrix before, and after, the work flow policy change and testing whether the difference between the estimated transition probability matrices is statistically significant. The maximum likelihood estimators of T_0 and T_1 calculated from the observed sequence in Table 2 are

$$\hat{T}_0 = \begin{bmatrix} 0.36 & 0.21 & 0.43 \\ 0.67 & 0.33 & 0.00 \\ 0.40 & 0.10 & 0.50 \end{bmatrix}$$

and

$$\hat{T}_1 = \begin{bmatrix} 0.17 & 0.66 & 0.17 \\ 0.18 & 0.44 & 0.38 \\ 0.33 & 0.56 & 0.11 \end{bmatrix}.$$

Both estimated transition probability matrices correspond to irreducible ergodic Markov chains so that the limiting probabilities of each of the states are easy to compute. For a Markov chain with transition probability matrix given by T_0, the limiting probabilities of States 1, 2 and 3 are 0.43, 0.19 and 0.22, respectively. Similarly, for a Markov chain with transition probability matrix given by T_1, the limiting probabilities of States 1, 2 and 3 are 0.22, 0.52 and 0.26, respectively. It appears that the new work flow policy is successful in that the proportion of time in States 1 and 3 is reduced, providing more efficient use of the machine and lower maintenance costs. To determine if the difference between $\hat{T}_0$ and $\hat{T}_1$ is significant, we use the likelihood statistics given in Eq. (5), where the change-point is known to

Table 2
The initial state (hour 0) and 60 hourly observed state observations for the machine usage process described in the example

Hour	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
State	1	2	2	1	3	3	2	2	1	1	3	3	3	1	3	1
Hour		16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
State		3	1	1	1	2	1	1	1	3	3	3	1	2	1	3
Hour		31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
State		2	1	2	3	1	3	2	2	2	2	2	1	2	2	2
Hour		46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
State		3	2	2	3	3	2	3	1	1	2	1	2	3	2	3

The new work flow policy is implemented at hour 30, so that the first transition under the new policy is from hour 30 to hour 31.

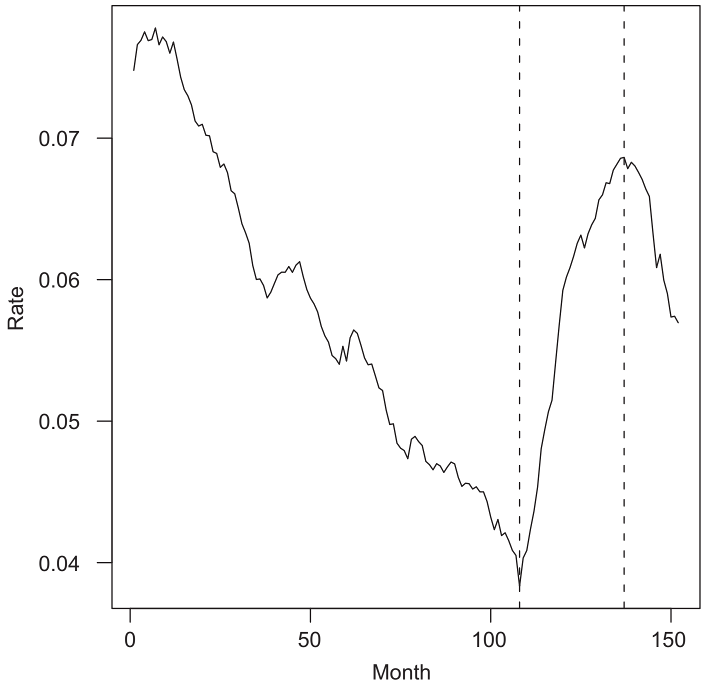

Fig. 7. Seasonally adjusted monthly unemployment rate for Texas between January 1992 and August 2004. The vertical dashed lines indicate the locations of the estimated change-points in the series.

be 30. Analysis of the observed sequence in Table 2 yields $\Lambda = 16.3247$, a χ^2-statistic with 6 degrees of freedom. The corresponding p-value for the test of equality between T_0 and T_1 is 0.01211. Therefore, the change in work flow policy has a significant effect on machine use at the 5% significance level.

4.2. Texas unemployment data

As an example we consider the monthly data on state unemployment rates provided by the US Department of Labor. Specifically, we consider the unemployment rate for the state of Texas between January 1992 and August 2004. The data are seasonally adjusted and are readily available at the Department of Labor website (http://stats.bls.gov). A plot of the observed unemployment rate is given in Fig. 7. Let R_i be the observed unemployment rate for the ith month for $i = 0, \ldots, n+1$. Define a simple two-state stochastic process from this observed process as $X_i = \delta(R_i < R_{i+1}) + 1$ for $i = 1, \ldots, n$. We will assume that this series can be modeled using a Markov chain with possible change-points. The results of the empirical studies of Section 3 tend to indicate that the BIC method is the more appropriate for

Table 3
Value of BIC for $k = 0, 1, \ldots, 5$ estimated change-points in the Texas unemployment rate series

k	0	1	2	3	4	5
BIC(k)	1.34638	1.30543	1.22506	1.24246	1.27579	1.62787

estimating the number of change-points in this model. Therefore the BIC method is computed for fitting $k = 0, 1, \ldots, 5$ change-points for the series. The values of the BIC are given in Table 3. One can clearly observe from Table 3 that the BIC measure is minimized for the model that contains two change-points. The corresponding maximum likelihood estimates of the change-points are $\hat{\psi}_1 = 106$ and $\hat{\psi}_2 = 135$. These months correspond to December 2000 and May 2003. The locations of the beginning points of these transitions are plotted in Fig. 7. One can observe from the figure that the behavior of the series changes near both of these points. The estimated transition probability matrices are

$$\hat{T}_0 = \begin{bmatrix} 0.7333 & 0.2667 \\ 0.6774 & 0.3226 \end{bmatrix},$$

$$\hat{T}_1 = \begin{bmatrix} 0.0000 & 1.0000 \\ 0.0769 & 0.9231 \end{bmatrix}$$

and

$$\hat{T}_2 = \begin{bmatrix} 0.7273 & 0.2727 \\ 1.0000 & 0.0000 \end{bmatrix}.$$

The estimates of the transition probability matrices indicate that while the behavior of the series is more erratic prior to the first change-point, one can observe that between the first and second change-points the series has a clear affinity for State 2, where the rate increases from the previous month. After the second change-point there is a clear affinity for State 1, where the rate decreases from the previous month. The general behavior is clearly indicated in the original series given in Fig. 7.

5. Discussion

This paper presents an array of methods that are designed to detect and estimate change-points in observed realizations from Markov chains. As was indicated, if the locations of the change-points are known, then the testing theory follows directly from the standard likelihood and asymptotic theory that has been previously developed for Markov chains. In the case where the locations of the change-points are unknown, the likelihood ratio is still used for testing, but the asymptotic theory does not hold. Therefore, the bootstrap is used to estimate p-values in this case. Empirical evidence from a small simulation study reveals that the bootstrap has the potential to be an effective tool for this problem. In the most difficult case where the number and location of the change-points are unknown, the AIC and BIC measures were suggested to aid with model selection. A small simulation study revealed that the AIC method is apparently unsuited for this purpose, but that the BIC method is potentially a useful method.

While the results of this work are encouraging, there are many potential avenues for additional research. In the case where the number, but not the location, of the change-points is known, it would be helpful to theoretically investigate the asymptotic null distribution of the likelihood ratio statistic. A comparison in the performance of the test based on using the asymptotic distribution and the bootstrap method would then be possible. In the case where the number of change-points is not known, it would be helpful to investigate the possibility of fine-tuning the penalty term in the BIC measure in order to provide the most efficient estimate of k. Additionally, other theoretical properties of the estimate of k, such as the bias and mean squared error, need to be investigated as well.

Estimation of multiple change-points in a long observed realization can be computationally burdensome. Note that for a series of length $n+1$, there are $\binom{n-1}{k}$ possible locations for the k change-points $\psi_1, \ldots, \psi_k$. For example, if $n = 500$ and $k = 5$, then the function given in Eq. (10) must be evaluated more than 252 billion times. Further, to evaluate AIC(k) or BIC(k) for $k \in \{0, \ldots, v\}$, the function given in Eq. (10) must be evaluated

$$\sum_{k=0}^{v} \binom{n-1}{k}$$

times. Therefore, to solve large problems using these methods, the development of computational methods that would ease this burden would be important.

Finally, it should be noted that the methodology studied in this paper can be easily modified to work with the case where there are several independent realizations instead of just a single observed realization. However, in this case it would be essential to assume that the sequence of probability transition matrices for the Markov chain associated with each observed realization is the same. Another possible generalization of this methodology would be to the case where S_c is countable.

References

Akaike, H., 1974. A new look at statistical model identification. IEEE Trans. Automat. Control 19, 716–723.

Anderson, T.W., Goodman, L.A., 1957. Statistical inference about Markov chains. Ann. Math. Statist. 28, 89–109.

Athreya, K.B., Fuh, C.D., 1992. Bootstrapping Markov chains. In: LePage, R., Billard, L. (Eds.), Exploring the Limits of Bootstrap. Wiley, New York, pp. 49–64.

Bartlett, M.S., 1951. The frequency goodness of fit test for probability chains. Proc. Cambridge Philos. Soc. 47, 89–110.

Basawa, I.V., Rao, B.L.S.P., 1980. Statistical Inference for Stochastic Processes. Academic Press, New York.

Billingsley, P., 1961a. Statistical methods in Markov chains. Ann. Math. Statist. 32, 12–40.

Billingsley, P., 1961b. Statistical Inference for Markov Processes. The University of Chicago Press, Chicago, IL.

Clark, W.A.V., 1965. Markov chain analysis in Geography: an application to the movement of rental housing areas. Ann. Assoc. Amer. Geographers 55, 351–359.

Cressie, N., Read, T.R.C., 1984. Multinomial goodness-of-fit tests. J. Roy. Statist. Soc., Ser. B 46, 440–464.

Csörgö, M., Horváth, L., 1997. Limit Theorems in Change-Point Analysis. Wiley, New York.

Dahiya, R.C., 1986. Integer-parameter estimation in discrete distributions. Comm. Statist. Theory Methods 15, 709–725.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press, Cambridge.

Efron, B., 1979. The bootstrap: another look at the jackknife. Ann. Statist. 7, 1–26.

Fuh, C.D., 1993. Statistical inquiry for Markov chains by bootstrap method. Statist. Sinica 3, 53–66.

Geary, K., 1978. Indicators of educational progress—a Markov chain approach applied to Switzerland. J. Modern African Stud. 16, 141–151.

Gottschau, A., 1992. Exchangeability in multivariate Markov chain models. Biometrika 58, 751–763.

Madansky, A., 1963. Tests of homogeneity for correlated samples. J. Amer. Statist. Assoc. 58, 97–119.

Maltsev, A.A., Silaev, A.M., 1992. Optimal estimation of the time of change in the characteristics of a Markov chain. Automat. Remote Control 53, 52–58.

Ramanathan, T.V., Rajarshi, M.B., 1997. Testing homogeneity over time of a parameter of a Markov sequence. Comm. Statist. Theory Methods 26, 317–330.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Silaev, A.M., 1997. Optimal estimation of the parameters of Markovian sequences that change their characteristics at random time instants. Automat. Remote Control 58, 1601–1610.

Turchin, P.B., 1986. Modeling the effect of host patch size on Mexican bean beetle emigration. Ecology 67, 124–132.

Yakir, B., 1994. Optimal detection of a change in distribution when the observations form a Markov chain with finite state space. In: Carlstein, E. Müller, H.-G., Siegmund, D., (Eds.), Change-Point Problems, IMS Lecture Notes—Monograph Series volume 23, pp. 346–358.

Yang, M.C.K., 1979. Some results on size and utilization distribution of home range. In: Ord, J.K., Patil, G.P., Tailie, C. (Eds.), Statistical Distributions in Ecological Work, vol. 4. International Co-operative Publishing House, Fairland, Maryland, pp. 429–449.