Asymptotic properties of maximum likelihood estimators in models with multiple change points

HEPING HE[1] and THOMAS A. SEVERINI[2]

[1]*Department of Mathematics, University of Kansas, 1460 Jayhawk Blvd, Lawrence, KS 66045, USA.*
E-mail: hhe@math.ku.edu
[2]*Department of Statistics, Northwestern University, Evanston, IL 60208, USA.*
E-mail: severini@northwestern.edu

Models with multiple change points are used in many fields; however, the theoretical properties of maximum likelihood estimators of such models have received relatively little attention. The goal of this paper is to establish the asymptotic properties of maximum likelihood estimators of the parameters of a multiple change-point model for a general class of models in which the form of the distribution can change from segment to segment and in which, possibly, there are parameters that are common to all segments. Consistency of the maximum likelihood estimators of the change points is established and the rate of convergence is determined; the asymptotic distribution of the maximum likelihood estimators of the parameters of the within-segment distributions is also derived. Since the approach used in single change-point models is not easily extended to multiple change-point models, these results require the introduction of those tools for analyzing the likelihood function in a multiple change-point model.

Keywords: change-point fraction; common parameter; consistency; convergence rate; Kullback–Leibler distance; within-segment parameter

1. Introduction

A change-point model for a sequence of independent random variables $X_1, \ldots, X_n$ is a model in which there exist unknown change points $n_1, \ldots, n_k$, $0 = n_0 < n_1 < \cdots < n_k < n_{k+1} = n$, such that, for each $j = 1, 2, \ldots, k+1$, $X_{n_{j-1}+1}, \ldots, X_{n_j}$ are identically distributed with a distribution that depends on j. Here, we consider parametric change-point models in which the distribution of $X_{n_{j-1}+1}, \ldots, X_{n_j}$ is parametric; however, the form of the distribution can be different for each j. Change-point models are used in many fields. For example, Broemeling and Tsurumi (1987) uses a multiple change-point model for the US demand for money; Lombard (1986) uses a multiple change-point model to model the effect of sudden changes in wind direction on the flight of a projectile; Reed (1998) uses a multiple change-point model in the analysis of forest fire data. A number of authors have used multiple change-point models in the analysis of DNA sequences; see, for example, Braun and Muller (1998), Fu and Curnow (1990a, 1990b) and Halpern (2000). Many further examples are provided in the monographs Chen and Gupta (2000) and Csörgő and Horváth (1997).

The goal of this paper is to establish the asymptotic properties of maximum likelihood estimators of the parameters of a multiple change-point model, under easily verifiable conditions. These

results are based on the following model. Assume that the vectors in the data set $x_1, x_2, \ldots, x_n$ are independently drawn from the parametric model

$$f_j(\psi^0, \theta_j^0; x_i), \qquad n_{j-1}^0 + 1 \leq i \leq n_j^0, j = 1, 2, \ldots, k+1,$$

where $f_j(\psi^0, \theta_j^0; x)$ is a probability density function of a continuous distribution with unknown common parameter ψ^0 for all $j = 1, 2, \ldots, k+1$ and unknown within-segment parameters θ_j^0 for each $j = 1, 2, \ldots, k+1$; $f_j(\psi^0, \theta_j^0; x)$ may have the same functional form for some or all of $j = 1, 2, \ldots, k+1$; ψ^0 may be a vector; θ_j^0 may be a different vector parameter of different dimensions for each $j = 1, 2, \ldots, k+1$. In this model, there are k unknown change points $n_1^0, n_2^0, \ldots, n_k^0$, where the number of change points k is assumed to be known. The parameter ψ^0 is common to all segments.

There are a number of results available on the asymptotic properties of parameter estimators in change-point models. See, for example, Hinkley (1970, 1972), Hinkley and Hinkley (1970), Battacharya (1987), Fu and Curnow (1990a, 1990b), Jandhyala and Fotopoulos (1999, 2001) and Hawkins (2001); the two monographs Chen and Gupta (2000) and Csörgo and Horváth (1997) have detailed bibliographies on this topic.

In particular, Hinkley (1970) considers likelihood-based inference for a single change-point model, obtaining the asymptotic distribution of the maximum likelihood estimator of the change point under the assumption that the other parameters in the model are known. Hinkley (1970) and Hinkley (1972) argue that this asymptotic distribution is also valid when the parameters are unknown.

Unfortunately, there are problems in extending the approach used in Hinkley (1970, 1972) to the setting considered here. The method used in Hinkley (1970, 1972) is based on considering the relative locations of a candidate change point and the true change point. When there is only a single change point, there are only three possibilities: the candidate change point is either greater than, less than or equal to the true change point. However, in models with k change points, the relative positions of the candidate change points and the true change points can become quite complicated and the simplicity and elegance of the single change point argument is lost.

A second problem arises when extending the argument for the case in which the change points are the only parameters in the model to the case in which there are unknown within-segment parameters. The consistency argument used in the former case is extended to the latter case using a "consistency assumption" (Hinkley (1972), Section 4.1); this condition is discussed in Appendix A and examples are given which show that this assumption is a strong one that is not generally satisfied in the class of models considered here.

There are relatively few results available on the asymptotic properties of maximum likelihood estimators in multiple change-point models. Thus, the present paper has done several things. In the general model described above, in which there is a fixed, but arbitrary, number of change points, we show that the maximum likelihood estimators of the change points are consistent and converge to the true change points at the rate $1/n$, under relatively weak regularity conditions. As noted above, a simple extension of the approach used in single change-point models is not available; thus, the second thing achieved by this paper is the introduction of the tools necessary for analyzing the likelihood function in a multiple change-point model. Finally, the asymptotic

distribution of the maximum likelihood estimators of the parameters of the within-segment distributions is derived for the general case described above, in which the form of the distribution can change from segment to segment and in which, possibly, there are parameters that are common to all segments.

The paper is organized as follows. The asymptotic theory of maximum likelihood estimators of a multiple change-point model is described in Section 2. Section 3 contains a numerical example illustrating these results and Section 4 contains some discussion of future research which builds on the results given in this paper. Appendix A discusses the "consistency assumption" used in Hinkley (1972); all technical proofs are given in Appendix B.

2. Asymptotic theory

Consider estimation of the multiple change-point model introduced in Section 1. For any change point configuration $0 = n_0 < n_1 < n_2 < \cdots < n_k < n_{k+1} = n$, the log-likelihood function is given by

$$l \equiv l(n_1, \ldots, n_k, \theta_1, \ldots, \theta_{k+1}, \psi) = \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \log f_j(\psi, \theta_j; x_i).$$

Estimators of all change points, all within-segment parameters and the common parameter are given by

$$(\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_k, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{k+1}, \hat{\psi}) = \underset{0 < n_1 < n_2 < \cdots < n_k < n; \theta_j \in \Theta_j, j=1,2,\ldots,k+1; \psi \in \Psi}{\arg \max} l,$$

where Θ_j, $j = 1, 2, \ldots, k+1$, and Ψ are the parameter spaces of θ_j, $j = 1, \ldots, k+1$, and ψ, respectively.

Let

$$\lambda_j^0 = n_j^0 / n \qquad \text{for } j = 1, 2, \ldots, k,$$
$$\lambda_j = n_j / n \qquad \text{for } j = 1, 2, \ldots, k,$$
$$\lambda^0 = (\lambda_1^0, \lambda_2^0, \ldots, \lambda_k^0),$$
$$\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k);$$
$$\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_{k+1}^0),$$
$$\theta = (\theta_1, \theta_2, \ldots, \theta_{k+1}),$$
$$\phi^0 = (\psi^0, \theta^0) = (\psi^0, \theta_1^0, \theta_2^0, \ldots, \theta_{k+1}^0),$$
$$\phi = (\psi, \theta) = (\psi, \theta_1, \theta_2, \ldots, \theta_{k+1}).$$

Note that λ^0 is taken to be a constant vector as n goes to infinity.

Define

$$\hat{\ell}^{(j)}(\psi, \theta_j) = \sum_{i=\hat{n}_{j-1}+1}^{\hat{n}_j} \log f_j(\psi, \theta_j; x_i), \qquad j = 1, 2, \ldots, k+1,$$

$$\ell^{(j)}(\psi, \theta_j) = \sum_{i=n_{j-1}^0+1}^{n_j^0} \log f_j(\psi, \theta_j; x_i), \qquad j = 1, 2, \ldots, k+1,$$

$$\hat{\ell}(\psi, \theta) = \sum_{j=1}^{k+1} \sum_{i=\hat{n}_{j-1}+1}^{\hat{n}_j} \log f_j(\psi, \theta_j; x_i),$$

$$\ell^0(\psi, \theta) = \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \log f_j(\psi, \theta_j; x_i),$$

$$\ell(\psi, \theta) = \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \log f_j(\psi, \theta_j; x_i).$$

The expected information matrix is given by

$$i(\psi, \theta) = E[-\ell_{\phi\phi}^0(\psi, \theta); \phi] = \begin{pmatrix} E[-\ell_{\psi\psi}^0(\psi, \theta); \phi] & E[-\ell_{\psi\theta}^0(\psi, \theta); \phi] \\ E[-\ell_{\psi\theta}^0(\psi, \theta); \phi]^T & E[-\ell_{\theta\theta}^0(\psi, \theta); \phi] \end{pmatrix},$$

$$E[-\ell_{\psi\theta}^0(\psi, \theta); \phi]$$
$$= \left(E[-\ell_{\psi\theta_1}^{(1)}(\psi, \theta_1); \phi], E[-\ell_{\psi\theta_2}^{(2)}(\psi, \theta_2); \phi], \ldots, E[-\ell_{\psi\theta_{k+1}}^{(k+1)}(\psi, \theta_{k+1}); \phi] \right),$$
$$E[-\ell_{\theta\theta}^0(\psi, \theta); \phi]$$
$$= \text{diag}\left(E[-\ell_{\theta_1\theta_1}^{(1)}(\psi, \theta_1); \phi], E[-\ell_{\theta_2\theta_2}^{(2)}(\psi, \theta_2); \phi], \ldots, E[-\ell_{\theta_{k+1}\theta_{k+1}}^{(k+1)}(\psi, \theta_{k+1}); \phi] \right),$$

where diag$(\cdot)$ denotes a diagonal block matrix whose diagonal blocks are in the bracket, other elements are zeros and the average expected information matrix is given by

$$\bar{i}(\psi, \theta) = \lim_{n \to \infty} \frac{1}{n} i(\psi, \theta).$$

The asymptotic properties of these estimators are based on the following regularity conditions. Other than the parts concerning change points, these conditions are typically similar to those required for the consistency and asymptotic normality of maximum likelihood estimators of parameters in models without change points; see, for example, Wald (1949). Particularly, compactness of parameter spaces is a common assumption in the classical likelihood literature.

These conditions are different from those required by Ferger (2001) and Döring (2007), who consider estimation of change points in a nonparametric setting in which nothing is assumed about the within-segment distributions, using a type of nonparametric M-estimator based on empirical processes. Thus, these authors do not require conditions on the within-segment likelihood functions; on the other hand, their method does not provide estimators of within-segment parameters.

Assumption 2.1. *It is assumed that for $j = 1, 2, \ldots, k$, $f_{j+1}(\psi^0, \theta^0_{j+1}; x) \neq f_j(\psi^0, \theta^0_j; x)$ on a set of non-zero measure.*

This assumption guarantees that the distributions in two neighboring segments are different; clearly, this is required for the change points to be well defined.

Assumption 2.2. *It is assumed that*:

1. *for $j = 1, 2, \ldots, k + 1$, θ_j and θ^0_j are contained in Θ_j, where Θ_j is a compact subset of $\mathcal{R}^{d_j}$; ψ and ψ^0 are contained in Ψ where Ψ is a compact subset of $\mathcal{R}^d$; here, $d, d_1, \ldots, d_{k+1}$ are non-negative integers*;
2. *$\ell(\psi, \theta)$ is third-order continuously differentiable with respect to ψ, θ*;
3. *the expectations of the first and second order derivatives of $\ell^0(\psi, \theta)$ with respect to ϕ exist for ϕ in its parameter space.*

Compactness of the parameter space is used to establish the consistency of the maximum likelihood estimators of $n_1/n, \ldots, n_k/n, \theta_1, \ldots, \theta_{k+1}, \psi$; see, for example, Bahadur (1971) for further discussion of this condition and its necessity in general models. If we assume further conditions on models, the compactness of the parameter space may be avoided. But this appears to be a substantial task for future work. Differentiability of the log-likelihood function is used to justify certain Taylor series expansions. Both parts of Assumption 2.2 are relatively weak and are essentially the same as conditions used in parametric models without change points; see, for example, Schervish (1995), Section 7.3. Part 3 is very weak and is used in the proof of Theorem 2.3.

Assumption 2.3. *It is assumed that*:

1. *for any $j = 1, 2, \ldots, k + 1$ and any integers s, t satisfying $0 \leq s < t \leq n$,*

$$E\left\{ \max_{\psi \in \Psi, \theta_j \in \Theta_j} \left(\sum_{i=s+1}^t \{\log f_j(\psi, \theta_j; X_i) - E[\log f_j(\psi, \theta_j; X_i)]\} \right)^2 \right\} \leq C(t - s)^r,$$

where $r < 2$ and C is a constant;
2. *for any $j = 1, 2, \ldots, k + 1$ and any integers s, t satisfying $n^0_{j-1} \leq s < t \leq n^0_j$,*

$$E\left\{ \max_{\psi \in \Psi, \theta_j \in \Theta_j} \left(\sum_{i=s+1}^t \{[\log f_j(\psi, \theta_j; X_i) - \log f_j(\psi^0, \theta^0_j; X_i)] - v(\psi, \theta_j; \psi^0, \theta^0_j)\} \right)^2 \right\}$$
$$\leq D(t - s)^r,$$

where $v(\psi, \theta_j; \psi^0, \theta_j^0)$ is introduced in equation (2), *$r < 2$ and D is a constant.*

Parts 1 and 2 of Assumption 2.3 are technical requirements on the behavior of the log-likelihood function between and within segments, respectively. This condition is used to ensure that the information regarding the within- and between-segment parameters grows quickly enough to establish consistency and asymptotic normality of the parameter estimators. These conditions are relatively weak; it is easy to check that they are satisfied by at least all distributions in the exponential family. Consider a probability density function of exponential family form:

$$f(\eta, x) = h(x)c(\eta) \exp\left(\sum_{i=1}^{m} w_i(\eta)t_i(x)\right).$$

It is then straightforward that the Schwarz inequality gives

$$\left(\sum_{i=s+1}^{t} \{\log f(\eta, X_i) - E[\log f(\eta, X_i)]\}\right)^2$$

$$\leq \left[1 + \sum_{q=1}^{m} w_q(\eta)^2\right]$$

$$\times \left\{\left[\sum_{i=s+1}^{t} \left(\log h(X_i) - E(\log h(X_i))\right)\right]^2 + \sum_{q=1}^{m}\left[\sum_{i=s+1}^{t} \left(t_q(X_i) - E(t_q(X_i))\right)\right]^2\right\}.$$

Therefore, Part 1 of Assumption 2.3 is satisfied with $r = 1$ because the function $w_q(\eta)$ assumed to be continuous can achieve its maximum on the compact parameter space. Similarly, Part 2 of Assumption 2.3 is also satisfied with $r = 1$.

The main results of this paper are given in the following three theorems.

Theorem 2.1 (Consistency). *Under Assumption* 2.1, *Part* 1 *of Assumption* 2.2 *and Part* 1 *of Assumption* 2.3, *$\hat{\lambda}_i \to_p \lambda_i^0, \hat{\theta}_j \to_p \theta_j^0$ and $\hat{\psi} \to_p \psi^0$ as $n \to +\infty$, that is, $\hat{\lambda}_i - \lambda_i^0 = o_p(1), \hat{\theta}_j - \theta_j^0 = o_p(1)$ and $\hat{\psi} - \psi^0 = o_p(1)$, where $\hat{\lambda}_i = \hat{n}_i/n$ for $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k + 1$.*

Note that $\hat{n}_i, i = 1, 2, \ldots, k$, are not consistent (Hinkley (1970)); it is the estimators of the change-point fractions $\hat{\lambda}_i, i = 1, 2, \ldots, k$, that are consistent. The consistency of $\hat{\theta}_j, j = 1, 2, \ldots, k + 1$, and $\hat{\psi}$ is the same as the corresponding result in classical likelihood theory for independent, identically distributed data.

Theorem 2.2 (Convergence rate). *Under Assumptions* 2.1–2.3, *we have*

$$\lim_{\delta \to \infty} \lim_{n \to \infty} P_r(n\|\hat{\lambda} - \lambda^0\|_\infty \geq \delta) = 0,$$

where $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_k)$, $\|\hat{\lambda} - \lambda^0\|_\infty = \max_{1 \leq j \leq k} |\hat{\lambda}_j - \lambda_j^0|$. That is, $\hat{\lambda}_i - \lambda_i^0 = O_p(n^{-1})$ for $i = 1, 2, \ldots, k$.

We now consider the asymptotic distribution of $\hat{\phi}$, where $\hat{\phi} = (\hat{\psi}, \hat{\theta})$.

Theorem 2.3 (Limiting distributions). *Under Assumptions* 2.1–2.3,

$$\sqrt{n}(\hat{\phi} - \phi^0) \xrightarrow{\mathcal{D}} N_{d+d_1+d_2+\cdots+d_{k+1}}(0, \bar{i}(\psi^0, \theta^0)^{-1}),$$

where $N_{d+d_1+d_2+\cdots+d_{k+1}}(0, \bar{i}(\psi^0, \theta^0)^{-1})$ is the $d+d_1+d_2+\cdots+d_{k+1}$-dimensional multivariate normal distribution with mean vector zero and covariance matrix $\bar{i}(\psi^0, \theta^0)^{-1}$.

The proofs of Theorems 2.1–2.3 are based on the following approach.
Define a function J by

$$
\begin{aligned}
J = & \sum_{j=1}^{k+1} \sum_{i=1}^{k+1} \frac{n_{ji}}{n} \left\{ \left[\int_{-\infty}^{+\infty} [\log f_j(\psi, \theta_j; x) - \log f_i(\psi^0, \theta_i^0; x)] f_i(\psi^0, \theta_i^0; x) \, dx \right. \right\} \\
& + \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \{ \log f_j(\psi, \theta_j; x_i) - E[\log f_j(\psi, \theta_j; X_i)] \} \\
& - \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \{ \log f_j(\psi^0, \theta_j^0; x_i) - E[\log f_j(\psi^0, \theta_j^0; X_i)] \},
\end{aligned}
\tag{1}
$$

where n_{ji} is the number of observations in the set $[n_{j-1} + 1, n_j] \cap [n_{i-1}^0 + 1, n_i^0]$ for $i, j = 1, 2, \ldots, k + 1$. We obviously have that

$$
\underset{0 < n_1 < n_2 < \cdots < n_k < n; \theta_j \in \Theta_j, 1 \leq j \leq k+1; \psi \in \Psi}{\arg \max} l = \underset{0 < n_1 < n_2 < \cdots < n_k < n; \theta_j \in \Theta_j, 1 \leq j \leq k+1; \psi \in \Psi}{\arg \max} J;
$$

thus, the maximum likelihood estimators may be defined as the maximizers of J rather than as the maximizers of l.

Let $v(\psi, \theta_j; \psi^0, \theta_i^0)$ be defined by

$$
v(\psi, \theta_j; \psi^0, \theta_i^0) = \int_{-\infty}^{+\infty} \left[\log \frac{f_j(\psi, \theta_j; x)}{f_i(\psi^0, \theta_i^0; x)} \right] f_i(\psi^0, \theta_i^0, x) \, dx
$$

for $i, j = 1, 2, \ldots, k + 1$.
$$\tag{2}$$

Note that J may be written $J = J_1 + J_2$, where

$$
J_1 = \sum_{j=1}^{k+1} \sum_{i=1}^{k+1} \frac{n_{ji}}{n} v(\psi, \theta_j; \psi^0, \theta_i^0)
\tag{3}
$$

and

$$J_2 = \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}+1}^{n_j} \{\log f_j(\psi, \theta_j; x_i) - E[\log f_j(\psi, \theta_j; X_i)]\}$$

$$- \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=n_{j-1}^0+1}^{n_j^0} \{\log f_j(\psi^0, \theta_j^0; x_i) - E[\log f_j(\psi^0, \theta_j^0; X_i)]\}. \tag{4}$$

Alternatively, we may write

$$J_2 = \frac{1}{n} \sum_{j=1}^{k+1} \sum_{i=1}^{k+1} \left\{ \sum_{t \in \tilde{n}_{ji}} [\log f_j(\psi, \theta_j; x_t) - E(\log f_j(\psi, \theta_j; X_t))] \right.$$

$$\left. - \sum_{t \in \tilde{n}_{ji}} [\log f_i(\psi^0, \theta_i^0; x_t) - E(\log f_i(\psi^0, \theta_i^0, X_t))] \right\}, \tag{5}$$

where $\tilde{n}_{ji} = [n_{j-1} + 1, n_j] \cap [n_{i-1}^0 + 1, n_i^0]$.

Note that J_1 is a weighted sum of the negative Kullback–Leibler distances; it will be shown that J_2 approaches 0 as $n \to \infty$. Also, $v(\psi, \theta_j; \psi^0, \theta_i^0) \leq 0$ with equality if and only if $f_j(\psi, \theta_j; x) = f_i(\psi^0, \theta_i^0; x)$ almost everywhere (Kullback and Leibler (1951)).

Lemma 2.1 gives a bound for J_1.

Lemma 2.1. *Under Assumption* 2.1 *and Part* 1 *of Assumption* 2.2, *there exist two positive constants $C_1 > 0$ and $C_2 > 0$ such that, for any λ and ϕ,*

$$J_1 \leq -\max\{C_1 \|\lambda - \lambda^0\|_\infty, C_2 \rho(\phi, \phi^0)\},$$

where $\|\lambda - \lambda^0\|_\infty = \max_j |\lambda_j - \lambda_j^0|$ and $\rho(\phi, \phi^0) = \max_j |v(\psi, \theta_j; \psi^0, \theta_j^0)|$.

Lemma 2.2 describes between-segment properties and within-segment properties of this model.

Lemma 2.2. *Under Part* 1 *of Assumption* 2.2, *the following two results follow from Parts* 1 *and* 2 *of Assumption* 2.3 *respectively*:

(I) *for any $j = 1, 2, \ldots, k + 1$, any $0 \leq m_1 < m_2 \leq n$ and any positive number $\varepsilon > 0$, there exist a constant A_j, independent of ε, and a constant $r < 2$, such that*

$$P_r \left(\max_{m_1 \leq s < t \leq m_2, \theta_j \in \Theta_j, \psi \in \Psi} \left| \sum_{i=s+1}^{t} \{\log f_j(\psi, \theta_j; X_i) - E[\log f_j(\psi, \theta_j; X_i)]\} \right| > \varepsilon \right)$$

$$\leq A_j \frac{(m_2 - m_1)^r}{\varepsilon^2}. \tag{6}$$

(II) *for any* $j = 1, 2, \ldots, k+1$ *and any positive number* $\varepsilon > 0$, *there exist a constant* B_j, *independent of* ε, *and a constant* $r < 2$, *such that*

$$P_r \left(\max_{n_{j-1}^0 \le s < t \le n_j^0, \psi \in \Psi, \theta_j \in \Theta_j} \sum_{i=s+1}^{t} \{[\log f_j(\psi, \theta_j; X_i) - \log f_j(\psi^0, \theta_j^0; X_i)] \right.$$
$$\left. - v(\psi, \theta_j; \psi^0, \theta_j^0)\} > \varepsilon \right) \le B_j \frac{(n_j^0 - n_{j-1}^0)^r}{\varepsilon^2}. \tag{7}$$

In practical applications, it is useful to have an estimator of $\bar{i}(\psi^0, \theta^0)$. Let

$$\hat{i}(\hat{\psi}, \hat{\theta}) = \begin{pmatrix} \hat{E}[-\hat{\ell}_{\psi\psi}(\hat{\psi}, \hat{\theta}); \hat{\phi}] & \hat{E}[-\hat{\ell}_{\psi\theta}(\hat{\psi}, \hat{\theta}); \hat{\phi}] \\ \hat{E}[-\hat{\ell}_{\psi\theta}(\hat{\psi}, \hat{\theta}); \hat{\phi}]^T & \hat{E}[-\hat{\ell}_{\theta\theta}(\hat{\psi}, \hat{\theta}); \hat{\phi}] \end{pmatrix},$$

$$\hat{E}[-\hat{\ell}_{\psi\psi}(\hat{\psi}, \hat{\theta}); \hat{\phi}] = \sum_{j=1}^{k+1} \sum_{i=\hat{n}_{j-1}+1}^{\hat{n}_j} \frac{1}{f_j^2(\hat{\psi}, \hat{\theta}_j; x_i)} f_{j\psi}(\hat{\psi}, \hat{\theta}_j; x_i) f_{j\psi}^T(\hat{\psi}, \hat{\theta}_j; x_i),$$

$$\hat{E}[-\hat{\ell}_{\psi\theta_j}(\hat{\psi}, \hat{\theta}); \hat{\phi}] = \sum_{i=\hat{n}_{j-1}+1}^{\hat{n}_j} \frac{1}{f_j^2(\hat{\psi}, \hat{\theta}_j; x_i)} f_{j\psi}(\hat{\psi}, \hat{\theta}_j; x_i) f_{j\theta_j}^T(\hat{\psi}, \hat{\theta}_j; x_i),$$

$$\hat{E}[-\hat{\ell}_{\theta_j\theta_j}(\hat{\psi}, \hat{\theta}); \hat{\phi}] = \sum_{i=\hat{n}_{j-1}+1}^{\hat{n}_j} \frac{1}{f_j^2(\hat{\psi}, \hat{\theta}_j; x_i)} f_{j\theta_j}(\hat{\psi}, \hat{\theta}_j; x_i) f_{j\theta_j}^T(\hat{\psi}, \hat{\theta}_j; x_i)$$

for $j = 1, 2, \ldots, k+1$. Then $\hat{i}(\hat{\psi}, \hat{\theta})/n$ is a consistent estimator of $\bar{i}(\psi^0, \theta^0)$.

3. An example

Consider the problem of analyzing the mineral content of a core sample, which is extensively studied in Chen and Gupta (2000), Chernoff (1973) and Srivastava and Worsley (1986). In particular, we consider the data in Chernoff (1973) on the mineral content of 12 minerals in a core sample measured at $N = 53$ equally spaced points. Since some of the minerals have a very low assay, we follow Chen and Gupta (2000) and Srivastava and Worsley (1986) in analyzing only the $p = 5$ variables Z_1, Z_8, Z_9, Z_{10} and Z_{12} with the highest assays. Thus, we assume that $(Z_1, Z_8, Z_9, Z_{10}, Z_{12})$ has a 5-variate normal distribution with a within-segment mean parameter vector and a variance-covariance matrix that is common to all segments. The analyses of Chen and Gupta (2000), Chernoff (1973) and Srivastava and Worsley (1986) suggest that there are 5 change points of the mean vector and, hence, we make that assumption here.

The estimates of 5 change points, within-segment parameters of mean vectors and common parameter of variance-covariance matrix were computed using maximum likelihood. The estimated change points are 7, 20, 24, 32 and 41, which are different from those estimated change

points by Chen and Gupta (2000), Chernoff (1973) and Srivastava and Worsley (1986), and are more reasonable. This is because Chen and Gupta (2000), Chernoff (1973) and Srivastava and Worsley (1986) use the binary segmentation procedures which detect multiple change points one by one, not simultaneously, whereas the method in this paper simultaneously estimates multiple change points. The estimated six within-segment mean vectors are in the following. They are arranged according to the order of from left to right. For example, the two vectors on the first line are, respectively, the first and second within-segment mean vectors.

$$(287.14, 58.57, 25.71, 240.00, 422.86), \qquad (277.31, 144.61, 24.69, 306.15, 274.62),$$

$$(321.25, 502.50, 150.00, 620.00, 217.50), \qquad (397.50, 635.00, 428.75, 625.00, 4.38),$$

$$(470.00, 188.89, 214.44, 255.56, 108.89), \qquad (425.0, 155.92, 183.42, 320.0, 333.33).$$

The estimated common variance-covariance matrix is

$$\begin{pmatrix} 1485.71 & -966.03 & 569.41 & -421.41 & -590.87 \\ -966.03 & 8523.65 & 4649.95 & 5982.95 & 1054.22 \\ 569.41 & 4649.95 & 8767.11 & 4434.76 & 736.33 \\ -421.41 & 5982.95 & 4434.76 & 8768.49 & 780.03 \\ -590.87 & 1054.22 & 736.33 & 780.03 & 3193.37 \end{pmatrix}.$$

4. Discussion

This paper establishes the consistency of maximum likelihood estimators of the parameters of a general class of multiple change-point models and gives the asymptotic distribution of the parameters of the within-segment distributions. The required regularity conditions are relatively weak and are generally satisfied by exponential family distributions.

Some important problems in the analysis of multiple change-point models were not considered here. One is that the asymptotic distribution of the maximum likelihood estimator of the vector of change points was not considered. The reason for this is that the methods used to determine this asymptotic distribution are quite different from the methods used to establish the consistency of the maximum likelihood estimator; see, for example, Hinkley (1970) for a treatment of this problem in a single change-point model. Thus, this is essentially a separate research topic. However, the asymptotic properties obtained in this paper are necessary for the establishment of the asymptotic distribution of the maximum likelihood estimator of the vector of change points in this model. This will be a subject of future work.

Another important problem is to extend the results of this paper to the case in which the number of change points is not known and must be determined from the data. Clearly, a likelihood-based approach to this problem will require an understanding of the properties of maximum likelihood estimators in the model in which the number of change points is known. Thus, the results of the present paper can be considered as a first step toward the development of a likelihood-based methodology that can be used to determine simultaneously the number and location of the change points. This is also a topic of future research.

Appendix A: The consistency assumption of Hinkley (1972)

Consider a change-point model with a single change point, n_1^0, and suppose that there are no common parameters in the model. In Hinkley (1972), it is shown that $\hat{n}_1$, the maximum likelihood estimator of n_1^0, satisfies $\hat{n}_1 = n_1^0 + O_p(1)$ under the condition

$$\sup_{\theta_1} \sum_{i=n_1^0+1}^{n_1^0+m} \{\log f_1(X_i; \theta_1) - \log f_2(X_i; \theta_2^0)\} \to -\infty \tag{A.1}$$

with probability 1 as $m \to \infty$, which was described as a "consistency assumption". Note that the random variables in the sum $X_{n_1^0+1}, \ldots, X_{n_1^0+m}$ are drawn from the distribution with density f_2.

Suppose that

$$\frac{1}{m} \sum_{i=n_1^0+1}^{n_1^0+m} \{\log f_1(X_i; \theta_1) - \log f_2(X_i; \theta_2^0)\}$$

converges to

$$E\left\{\log \frac{f_1(X; \theta_1)}{f_2(X; \theta_2^0)}\right\}$$

as $m \to \infty$, uniformly in θ_1, where X is distributed according to the distribution with density $f_2(\cdot; \theta_2^0)$. Equation (A.1) then holds, provided that

$$\sup_{\theta_1} E\left\{\log \frac{f_1(X; \theta_1)}{f_2(X; \theta_2^0)}\right\} < 0;$$

note that, by properties of the Kullback–Leibler distance and Assumption 2.1,

$$E\left\{\log \frac{f_1(X; \theta_1)}{f_2(X; \theta_2^0)}\right\} < 0$$

for each θ_1.

Thus, condition (A.1) fails whenever the distribution corresponding to the density $f_2(\cdot; \theta_2^0)$ is in the closure of the set of distributions corresponding to densities of the form $f_1(\cdot; \theta_1)$, in a certain sense.

One such case occurs if f_1 and f_2 have the same parametric form with parameters θ_1^0, θ_2^0, respectively, satisfying $\theta_1^0 \neq \theta_2^0$. For instance, suppose that the random variables in the first segment are normally distributed with mean θ_1^0 and standard deviation 1 and the random variables in the second segment are normally distributed with mean θ_2^0 and standard deviation 1. Then

$$\sup_{\theta_1} \sum_{i=n_1^0+1}^{n_1^0+m} \{\log f_1(X_i; \theta_1) - \log f_2(X_i; \theta_2^0)\} = \frac{m}{2} (\bar{X}_m - \theta_2^0)^2,$$

where

$$\bar{X}_m = \frac{1}{m} \sum_{i=n_1^0+1}^{n_1^0+m} X_i$$

is normally distributed with mean θ_2^0 and variance $1/m$. Clearly, (A.1) does not hold in this case.

A similar situation occurs when the distribution with density $f_2(\cdot; \theta_2^0)$ can be viewed as a limit of the distributions with densities $f_1(\cdot; \theta_1)$. For instance, suppose that f_1 is the density of a Weibull distribution with rate parameter β and shape parameter α, $\theta_1 = (\alpha, \beta)$, $\beta \neq 1$, and f_2 is the density of an exponential distribution with rate parameter θ_2.

In this appendix, we show that this is a strong assumption that is not generally satisfied by otherwise well-behaved models. For instance, suppose that f_1 and f_2 have the same functional form and that the difference between the two distributions is due to the fact that $\theta_1^0 \neq \theta_2^0$. Again, (A.1) will not hold.

Thus, the consistency condition used in Hinkley (1972) is too strong for the general model considered here.

Appendix B: Technical details

Proof of Lemma 2.1. We first need to prepare some results which are to be used in this proof. For $i = 1, 2, \ldots, k$, let us define

$$g_i(\alpha, \phi^0) = \sup_{1 \leq j \leq k+1} \sup_{\theta_j \in \Theta_j} \sup_{\psi \in \Psi} [\alpha v(\psi, \theta_j; \psi^0, \theta_{i+1}^0) + (1-\alpha)v(\psi, \theta_j; \psi^0, \theta_i^0)],$$

where $0 \leq \alpha \leq 1$. We then have that $g_i(0, \phi^0) = g_i(1, \phi^0) = 0$ for $i = 1, 2, \ldots, k$. It is straightforward to show that $g_i(\alpha, \phi^0)$ is a convex function with respect to α for any $i = 1, 2, \ldots, k$.

Let $G_i(\phi^0) = 2g_i(1/2, \phi^0)$. Because $\alpha = 2\alpha(1/2) + (1-2\alpha)0$ for $0 \leq \alpha \leq 1/2$, convexity of $g_i(\alpha, \phi^0)$ gives that

$$g_i(\alpha, \phi^0) \leq 2\alpha g_i(1/2, \phi^0) = \alpha G_i(\phi^0) \qquad \text{for } i = 1, 2, \ldots, k.$$

Noting that

$$g_i(1/2, \phi^0) = \frac{1}{2} \sup_{1 \leq j \leq k+1} \sup_{\theta_j \in \Theta_j} \sup_{\psi \in \Psi} [v(\psi, \theta_j; \psi^0, \theta_{i+1}^0) + v(\psi, \theta_j; \psi^0, \theta_i^0)],$$

it follows from Assumption 2.1 that $G_i(\phi^0) < 0$. If we let $\bar{G}(\phi^0) = \max_{1 \leq i \leq k} G_i(\phi^0)$, then $\bar{G}(\phi^0) < 0$.

Let $\Delta_\lambda^0 = \min_{1 \leq j \leq k-1} |\lambda_{j+1}^0 - \lambda_j^0|$. Consider a change-point fraction configuration λ such that $\|\lambda - \lambda^0\|_\infty \leq \Delta_\lambda^0/4$. For any $j = 1, 2, \ldots, k$, there are two cases: a candidate change-point fraction λ_j may be on the left or the right of the true change-point fraction λ_j^0.

For any j with λ_j on the right of λ_j^0, we have that $\lambda_{j-1} \le \lambda_j^0 \le \lambda_j$. Then

$$J_1 \le \frac{n_{j,j+1}}{n} v(\psi, \theta_j; \psi^0, \theta_{j+1}^0) + \frac{n_{jj}}{n} v(\psi, \theta_j; \psi^0, \theta_j^0).$$

If we define $\alpha_{j,j+1} = n_{j,j+1}/(n_{j,j+1} + n_{jj})$, then the case $\|\lambda - \lambda^0\|_\infty \le \Delta_\lambda^0/4$ gives that $\alpha_{j,j+1} \le \frac{1}{2}$ and

$$J_1 \le \frac{n_{j,j+1} + n_{jj}}{n} [\alpha_{j,j+1} v(\psi, \theta_j; \psi^0, \theta_{j+1}^0) + (1 - \alpha_{j,j+1}) v(\psi, \theta_j; \psi^0, \theta_j^0)]$$

$$\le \frac{n_{j,j+1}}{n} G_j(\phi^0) \le (\lambda_j - \lambda_j^0) \bar{G}(\phi^0).$$

For any j with λ_j on the left of λ_j^0, we have that $\lambda_j \le \lambda_j^0 \le \lambda_{j+1}$. Similarly, we define $\alpha_{j,j-1} = n_{j,j-1}/(n_{j,j-1} + n_{jj})$. Using the fact that $\alpha_{j,j-1} \le \frac{1}{2}$, it similarly gives that $J_1 \le (\lambda_j^0 - \lambda_j) \bar{G}(\phi^0)$.

Therefore, if $\|\lambda - \lambda^0\|_\infty \le \Delta_\lambda^0/4$, then we obtain that $J_1 \le \|\lambda - \lambda^0\|_\infty \bar{G}(\phi^0)$. On the other hand,

$$J_1 \le \min_{1 \le j \le k+1} v(\psi, \theta_j; \psi^0, \theta_j^0) \frac{n_{jj}}{n} = - \max_{1 \le j \le k+1} |v(\psi, \theta_j; \psi^0, \theta_j^0)| \frac{n_{jj}}{n}.$$

We have $n_{jj}/n \ge \Delta_\lambda^0/2$ for any j, so

$$J_1 \le -\frac{1}{2} \Delta_\lambda^0 \sup_{1 \le j \le k+1} |v(\psi, \theta_j; \psi^0, \theta_j^0)| = -\frac{1}{2} \Delta_\lambda^0 \rho(\phi, \phi^0).$$

Now, consider the other case of a change-point fraction configuration λ, where $\|\lambda - \lambda^0\|_\infty > \Delta_\lambda^0/4$. It is clear that there exists a pair of integers (i, j) such that $n_{ij} \ge n\Delta_\lambda^0/4$, $n_{i,j+1} \ge n\Delta_\lambda^0/4$ and $n_{ij} \ge n_{i,j+1}$. Let $\alpha_{i,j+1} = n_{i,j+1}/(n_{i,j+1} + n_{ij})$. For any ϕ, we have that

$$J_1 \le \frac{n_{i,j+1} + n_{ij}}{n} [\alpha_{i,j+1} v(\psi, \theta_i; \psi^0, \theta_{j+1}^0) + (1 - \alpha_{i,j+1}) v(\psi, \theta_i; \psi^0, \theta_j^0)]$$

$$\le \frac{n_{i,j+1} + n_{ij}}{n} \min(\alpha_{i,j+1}, 1 - \alpha_{i,j+1}) \bar{G}(\phi^0)$$

$$\le \frac{\Delta_\lambda^0}{2} \min\left(\frac{n_{i,j+1}}{n}, \frac{n_{ij}}{n}\right) \bar{G}(\phi^0)$$

$$\le \frac{1}{2} \left(\frac{\Delta_\lambda^0}{2}\right)^2 \bar{G}(\phi^0).$$

Combining the results from the two cases of $\|\lambda - \lambda^0\|_\infty \le \Delta_\lambda^0/4$ and $\|\lambda - \lambda^0\|_\infty > \Delta_\lambda^0/4$, it follows that

$$J_1 \le \bar{G}(\phi^0) \min\left(\frac{1}{2}\left(\frac{\Delta_\lambda^0}{2}\right)^2, \|\lambda - \lambda^0\|_\infty\right) \le \frac{1}{2}\left(\frac{\Delta_\lambda^0}{2}\right)^2 \bar{G}(\phi^0) \|\lambda - \lambda^0\|_\infty$$

and

$$J_1 \le \frac{\Delta_\lambda^0}{2} \max\left[-\rho(\phi, \phi^0), \frac{\Delta_\lambda^0}{4} \bar{G}(\phi^0)\right] \le -\frac{\Delta_\lambda^0}{2} \min\left[\rho(\phi, \phi^0), -\frac{\Delta_\lambda^0}{4} \bar{G}(\phi^0)\right]. \qquad (B.1)$$

Note that (B.1) can be simplified. If we define

$$\varrho(\phi, \phi^0) = \max_{1 \le j \le k+1} \sup_{\theta_j \in \Theta_j} \sup_{\psi \in \Psi} |v(\psi, \theta_j; \psi^0, \theta_j^0)|,$$

then we have that $\rho(\phi, \phi^0)/\varrho(\phi, \phi^0) \le 1$. It follows from inequality (B.1) that

$$J_1 \le -\frac{\Delta_\lambda^0}{2} \varrho(\phi, \phi^0) \min\left[\frac{\rho(\phi, \phi^0)}{\varrho(\phi, \phi^0)}, -\frac{\Delta_\lambda^0}{4} \bar{G}(\phi^0)/\varrho(\phi, \phi^0)\right].$$

If $-(\Delta_\lambda^0/4)\bar{G}(\phi^0)/\varrho(\phi, \phi^0) \le 1$, then we have that

$$J_1 \le (\Delta_\lambda^0/2)^2\left(\rho(\phi, \phi^0)/\varrho(\phi, \phi^0)\right)\left(\bar{G}(\phi^0)/2\right).$$

If $-(\Delta_\lambda^0/4)\bar{G}(\phi^0)/\varrho(\phi, \phi^0) > 1$, then $J_1 \le -(\Delta_\lambda^0/2)\rho(\phi, \phi^0)$. Letting

$$C_2 = \min\{(\Delta_\lambda^0/2)^2|\bar{G}(\phi^0)|/(2\varrho(\phi, \phi^0)), \Delta_\lambda^0/2\},$$

inequality (B.1) gives that $J_1 \le -C_2\rho(\phi, \phi^0)$.

Setting $C_1 = (\Delta_\lambda^0/2)^2|\bar{G}(\phi^0)|/2$, we finally have that

$$J_1 \le -\max\{C_1\|\lambda - \lambda^0\|_\infty, C_2\rho(\phi, \phi^0)\},$$

which concludes the proof.   □

Proof of Lemma 2.2. With Part 1 of Assumption 2.3 in mind, equation (6) can be achieved by induction with respect to m_2. The induction method is similar to the one used in Móricz, Serfling and Stout (1982), so its proof is omitted here. Using Part 2 of Assumption 2.3, equation (7) can be proven similarly by the same induction method.   □

Proof of Theorem 2.1. Let

$$\Lambda_\delta = \{\lambda \in \Lambda: \|\lambda - \lambda^0\|_\infty > \delta\}, \qquad \Phi_\delta = \{\phi \in \Phi: \rho(\phi, \phi^0) > \delta\},$$

$$\Phi = \Theta_1 \times \Theta_2 \times \cdots \times \Theta_{k+1} \times \Psi,$$

$$\Lambda = \{(\lambda_1, \lambda_2, \ldots, \lambda_k)|\lambda_j = n_j/n, j = 1, 2, \ldots, k;$$

$$0 < n_1 < n_2 < \cdots < n_k < n\}.$$

Then, for any $\delta > 0$, it follows from Lemma 2.1 that

$$-\max_{\lambda \in \Lambda_\delta, \phi \in \Phi} J_1 \ge C_1\delta \quad \text{and} \quad -\max_{\phi \in \Phi_\delta, \lambda \in \Lambda} J_1 \ge C_2\delta.$$

Therefore, we obtain that

$$
\begin{aligned}
&P_r(\|\hat{\lambda} - \lambda^0\|_\infty > \delta) \\
&\quad \leq P_r\left(\max_{\lambda \in \Lambda_\delta, \phi \in \Phi} J > 0\right) \leq P_r\left(\max_{\lambda \in \Lambda_\delta, \phi \in \Phi} J_2 > -\max_{\lambda \in \Lambda_\delta, \phi \in \Phi} J_1\right) \leq P_r\left(\max_{\lambda \in \Lambda_\delta, \phi \in \Phi} |J_2| > C_1\delta\right) \\
&\quad \leq P_r\left(\max_{\lambda \in \Lambda_\delta, \phi \in \Phi} \sum_{j=1}^{k+1} \frac{1}{n}\left|\sum_{i=n_{j-1}+1}^{n_j} \{\log f_j(\psi, \theta_j; X_i) - E[\log f_j(\psi, \theta_j; X_i)]\}\right| > \frac{C_1}{2}\delta\right) \\
&\qquad + P_r\left(\sum_{j=1}^{k+1} \frac{1}{n}\left|\sum_{i=n_{j-1}^0+1}^{n_j^0} \{\log f_j(\psi^0, \theta_j^0; X_i) - E[\log f_j(\psi^0, \theta_j^0; X_i)]\right| > \frac{C_1}{2}\delta\right) \\
&\quad \leq \sum_{j=1}^{k+1} P_r\left(\max_{0 \leq n_{j-1} < n_j \leq n, \theta_j \in \Theta_j, \psi \in \Psi} \frac{1}{n}\left|\sum_{i=n_{j-1}+1}^{n_j} \{\log f_j(\psi, \theta_j; X_i) - E[\log f_j(\psi, \theta_j; X_i)]\}\right|\right. \\
&\qquad\qquad \left. > \frac{C_1\delta}{2(k+1)}\right) \\
&\qquad + \sum_{j=1}^{k+1} P_r\left(\frac{1}{n}\left|\sum_{i=n_{j-1}^0+1}^{n_j^0} \{\log f_j(\psi^0, \theta_j^0; X_i) - E[\log f_j(\psi^0, \theta_j^0; X_i)]\}\right| > \frac{C_1\delta}{2(k+1)}\right).
\end{aligned}
$$

It follows from Lemma 2.2 that

$$
P_r(\|\hat{\lambda} - \lambda^0\|_\infty > \delta) \leq 2\left[\frac{2(k+1)}{C_1\delta}\right]^2 \left(\sum_{j=1}^{k+1} A_j\right) n^{r-2} \to 0 \qquad \text{as } n \to +\infty,
$$

noting that $r < 2$.

For $\hat{\phi}$, we similarly obtain that

$$
\begin{aligned}
&P_r\left(\rho(\hat{\phi}, \phi^0) > \delta\right) \\
&\quad \leq P_r\left(\max_{\lambda \in \Lambda, \phi \in \Phi_\delta} J > 0\right) \\
&\quad \leq \sum_{j=1}^{k+1} P_r\left(\max_{0 \leq n_{j-1} < n_j \leq n, \theta_j \in \Theta_j, \psi \in \Psi} \frac{1}{n}\left|\sum_{i=n_{j-1}+1}^{n_j} \{\log f_j(\psi, \theta_j; X_i) - E[\log f_j(\psi, \theta_j; X_i)]\}\right|\right. \\
&\qquad\qquad \left. > \frac{C_2\delta}{2(k+1)}\right)
\end{aligned}
$$

$$+ \sum_{j=1}^{k+1} P_r\left(\frac{1}{n} \left| \sum_{i=n_{j-1}^0+1}^{n_j^0} \{\log f_j(\psi^0, \theta_j^0; X_i) - E[\log f_j(\psi^0, \theta_j^0; X_i)]\} \right| > \frac{C_2 \delta}{2(k+1)} \right).$$

Similarly, Lemma 2.2 shows that $P_r(\rho(\hat{\phi}, \phi^0) > \delta) \to 0$ as $n \to +\infty$. Noting the fact that $v(\psi, \theta_j; \psi^0, \theta_j^0) = 0$ if and only if $\psi = \psi^0$ and $\theta_j = \theta_j^0$, it follows that $\hat{\psi} \to_p \psi^0$ and $\hat{\theta}_j \to_p \theta_j^0$ for $j = 1, 2, \ldots, k+1$, which completes the proof. $\qquad\square$

Proof of Theorem 2.2. Let us first define

$$\Lambda_{\delta,n} = \{\lambda \in \Lambda : n\|\lambda - \lambda^0\|_\infty > \delta\}$$

for any $\delta > 0$. Because of the consistency of $\hat{\lambda}$, we need to consider only those terms whose observations are in $\tilde{n}_{j,j-1}$, $\tilde{n}_{j,j}$ and $\tilde{n}_{j,j+1}$ for all j in equation (5). Therefore, we have

$$P_r(n\|\hat{\lambda} - \lambda^0\|_\infty > \delta)$$

$$\leq \sum_{j=1}^{k+1} P_r\left(\max_{\lambda \in \Lambda_{\delta,n}, \phi \in \Phi} \left\{ \frac{1}{n} \sum_{t \in \tilde{n}_{jj}} [\log f_j(\psi, \theta_j; X_t) - E(\log f_j(\psi, \theta_j; X_t))] \right. \right.$$

$$- \frac{1}{n} \sum_{t \in \tilde{n}_{jj}} [\log f_j(\psi^0, \theta_j^0; X_t) - E(\log f_j(\psi^0, \theta_j^0; X_t))]$$

$$\left. \left. + \frac{1}{3(k+1)} J_1 \right\} > 0 \right)$$

$$+ \sum_{j=2}^{k+1} P_r\left(\max_{\lambda \in \Lambda_{\delta,n}, \phi \in \Phi} \left\{ \frac{1}{n} \sum_{t \in \tilde{n}_{j,j-1}} [\log f_j(\psi, \theta_j; X_t) - E(\log f_j(\psi, \theta_j; X_t))] \right. \right.$$

$$- \frac{1}{n} \sum_{t \in \tilde{n}_{j,j-1}} [\log f_{j-1}(\psi^0, \theta_{j-1}^0; X_t) - E(\log f_{j-1}(\psi^0, \theta_{j-1}^0; X_t))]$$

$$\left. \left. + \frac{1}{3k} J_1 \right\} > 0 \right)$$

$$+ \sum_{j=1}^{k} P_r\left(\max_{\lambda \in \Lambda_{\delta,n}, \phi \in \Phi} \left\{ \frac{1}{n} \sum_{t \in \tilde{n}_{j,j+1}} [\log f_j(\psi, \theta_j; X_t) - E(\log f_j(\psi, \theta_j; X_t))] \right. \right.$$

$$- \frac{1}{n} \sum_{t \in \tilde{n}_{j,j+1}} [\log f_{j+1}(\psi^0, \theta_{j+1}^0; X_t) - E(\log f_{j+1}(\psi^0, \theta_{j+1}^0; X_t))]$$

$$+ \frac{1}{3k} J_1 \Big\} > 0 \Big)$$

$$\equiv \sum_{j=1}^{k+1} I_{1j} + \sum_{j=2}^{k+1} I_{2j} + \sum_{j=1}^{k} I_{3j}.$$

First, consider the probability formulas I_{1j} in the above equation for any $j = 1, 2, \ldots, k+1$. The consistency of $\hat{\lambda}$ allows us to restrict our attention to the case $n_{jj} > \frac{1}{2}(n_j^0 - n_{j-1}^0)$. For this case, we have that

$$J_1 \leq \frac{n_j^0 - n_{j-1}^0}{2n} v(\psi, \theta_j; \psi^0, \theta_j^0).$$

Therefore, we obtain that

$$I_{1j} \leq P_r \Big(\sum_{t \in \tilde{n}_{jj}^*} \{ [\log f_j(\psi^*, \theta_j^*; x_t) - \log f_j(\psi^0, \theta_j^0; x_t)]$$

$$- v(\psi^*, \theta_j^*; \psi^0, \theta_j^0) \} > \frac{n_j^0 - n_{j-1}^0}{6(k+1)} |v(\psi^*, \theta_j^*; \psi^0, \theta_j^0)| \Big)$$

$$\leq P_r \Big(\max_{n_{j-1}^0 \leq s < t \leq n_j^0, \psi \in \Psi, \theta_j \in \Theta_j} \sum_{i=s+1}^{t} \{ [\log f_j(\psi, \theta_j; X_t) - \log f_j(\psi^0, \theta_j^0; X_t)]$$

$$- v(\psi, \theta_j; \psi^0, \theta_j^0) \} > \frac{E}{6(k+1)} (n_j^0 - n_{j-1}^0) \Big),$$

where $\tilde{n}_{jj}^*$, ψ^*, θ_j^* and λ^* are, respectively, the maximizing values of $\tilde{n}_{jj}$, ψ, θ_j and λ obtained through the maximization. Equation (7) of Lemma 2.2 can then be applied to show that $I_{1j} \to 0$ as $n, \delta \to \infty$.

Next, consider the probability formula I_{2j} for any $j = 2, \ldots, k+1$. In this case, $\lambda_{j-1} < \lambda_{j-1}^0$. We have that

$$I_{2j} \leq P_r \Big(\max_{\lambda \in \Lambda_{\delta,n}, \phi \in \Phi} \Big\{ \frac{1}{n} \sum_{t \in \tilde{n}_{j,j-1}} [\log f_j(\psi, \theta_j; X_t) - E(\log f_j(\psi, \theta_j; X_t))] + \frac{1}{6k} J_1 \Big\} > 0 \Big)$$

$$+ P_r \Big(\max_{\lambda \in \Lambda_{\delta,n}, \phi \in \Phi} \Big\{ -\frac{1}{n} \sum_{t \in \tilde{n}_{j,j-1}} [\log f_{j-1}(\psi, \theta_{j-1}; X_t) - E(\log f_{j-1}(\psi, \theta_{j-1}; X_t))]$$

$$+ \frac{1}{6k} J_1 \Big\} > 0 \Big)$$

$$\equiv I_{2j}^{(1)} + I_{2j}^{(2)}.$$

$I_{2j}^{(1)}$ and $I_{2j}^{(2)}$ can be handled in the same way, so we just show how to handle $I_{2j}^{(1)}$. Only two cases have to be considered.

If $n_{j-1}^0 - n_{j-1} \leq \delta$, then

$$I_{2j}^{(1)} \leq P_r \left(\max_{n_{j-1} \leq s < t \leq n_{j-1}^0, \theta_j \in \Theta_j, \psi \in \Psi} \left| \sum_{i=s+1}^{t} [\log f_j(\psi, \theta_j; X_i) - E(\log f_j(\psi, \theta_j; X_i))] \right| \right.$$
$$\left. > \frac{C_1 \delta}{6k} \right).$$

Equation (6) of Lemma 2.2 gives that $I_{2j}^{(1)} \to 0$ as $n, \delta \to +\infty$.

If $n_{j-1}^0 - n_{j-1} > \delta$ for the other case, then $J_1 \leq -C_1(n_{j-1}^0 - n_{j-1})/n$. Therefore, we obtain that

$$I_{2j}^{(1)} \leq P_r \left(\max_{n_{j-1} \leq s < t \leq n_{j-1}^0, \theta_j \in \Theta_j, \psi \in \Psi} \frac{n_{j-1}^0 - n_{j-1}}{n} \right.$$
$$\times \left(\frac{1}{n_{j-1}^0 - n_{j-1}} \sum_{i=s+1}^{t} [\log f_j(\psi, \theta_j; X_i) \right.$$
$$- E(\log f_j(\psi, \theta_j; X_i))]$$
$$\left. \left. - \frac{C_1}{6k} \right) > 0 \right)$$
$$\leq P_r \left(\max_{n_{j-1} \leq s < t \leq n_{j-1}^0, \theta_j \in \Theta_j, \psi \in \Psi} \left| \sum_{i=s+1}^{t} [\log f_j(\psi, \theta_j; X_i) - E(\log f_j(\psi, \theta_j; X_i))] \right| \right.$$
$$\left. > \frac{C_1}{6k}(n_{j-1}^0 - n_{j-1}) \right),$$

which converges to zero as $n, \delta \to 0$, by equation (6) of Lemma 2.2.

I_{3j} can be handled in the same way as I_{2j}. Therefore, Theorem 2.2 is proved.   □

Proof of Theorem 2.3. We first have the expansion

$$\hat{\ell}_\phi(\hat{\psi}, \hat{\theta}) - \hat{\ell}_\phi(\psi^0, \theta^0) = [\hat{\ell}_{\phi\phi}(\psi^0, \theta^0) + o_p(n)](\hat{\phi} - \phi^0).$$

The fact that $\hat{\ell}_\phi(\hat{\psi}, \hat{\theta}) = 0$ then gives that

$$\sqrt{n}(\hat{\phi} - \phi^0) = \left[-\frac{1}{n}\hat{\ell}_{\phi\phi}(\psi^0, \theta^0) + o_p(1) \right]^{-1} \frac{\hat{\ell}_\phi(\psi^0, \theta^0)}{\sqrt{n}}.$$

Now, consider the limit of $\hat{\ell}_\phi(\psi^0, \theta^0)/\sqrt{n}$. We have that

$$\frac{1}{\sqrt{n}}\hat{\ell}_\phi(\psi^0, \theta^0) = \frac{1}{\sqrt{n}}[\hat{\ell}_\phi(\psi^0, \theta^0) - \ell_\phi^0(\psi^0, \theta^0)] + \frac{1}{\sqrt{n}}\ell_\phi^0(\psi^0, \theta^0).$$

Because of the consistency of $\hat{\lambda}$, we can assume that $n_{j-1}^0 < \hat{n}_j < n_{j+1}^0$ for $j = 1, 2, \ldots, k$. It is then straightforward to obtain that

$$\frac{1}{\sqrt{n}}[\hat{\ell}_\phi(\psi^0, \theta^0) - \ell_\phi^0(\psi^0, \theta^0)]$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^{k+1}[\hat{\ell}_\phi^{(j)}(\psi^0, \theta_j^0) - \ell_\phi^{(j)}(\psi^0, \theta_j^0)]$$

$$= \frac{1}{\sqrt{n}}\sum_{j=1}^{k+1}\left\{ I(\hat{n}_j \geq n_j^0, \hat{n}_{j-1} \geq n_{j-1}^0) \right.$$

$$\times \left[\sum_{i=n_j^0+1}^{\hat{n}_j} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) - \sum_{i=n_{j-1}^0+1}^{\hat{n}_{j-1}} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) \right]$$

$$+ I(\hat{n}_j \geq n_j^0, \hat{n}_{j-1} < n_{j-1}^0)$$

$$\times \left[\sum_{i=n_j^0+1}^{\hat{n}_j} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) + \sum_{i=\hat{n}_{j-1}+1}^{n_{j-1}^0} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) \right]$$

$$+ I(\hat{n}_j < n_j^0, \hat{n}_{j-1} \geq n_{j-1}^0)$$

$$\times \left[-\sum_{i=\hat{n}_j+1}^{n_j^0} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) - \sum_{i=n_{j-1}^0+1}^{\hat{n}_{j-1}} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) \right]$$

$$+ I(\hat{n}_j < n_j^0, \hat{n}_{j-1} < n_{j-1}^0)$$

$$\left. \times \left[-\sum_{i=\hat{n}_j+1}^{n_j^0} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) + \sum_{i=\hat{n}_{j-1}+1}^{n_{j-1}^0} \frac{\partial}{\partial\phi}\log f_j(\psi^0, \theta_j^0; X_i) \right]\right\}.$$

It follows from Theorem 2.2 that

$$\frac{1}{\sqrt{n}}[\hat{\ell}_\phi(\psi^0, \theta^0) - \ell_\phi^0(\psi^0, \theta^0)] = \frac{1}{\sqrt{n}}O_p(1),$$

which converges to zero in probability as $n \to \infty$.

Since

$$\frac{1}{\sqrt{n}} \ell_\phi^0(\psi^0, \theta^0) \xrightarrow{\mathcal{D}} N_{d+d_1+d_2+\cdots+d_{k+1}}(0, \bar{i}(\psi^0, \theta^0)),$$

it follows that

$$\frac{1}{\sqrt{n}} \hat{\ell}_\phi(\psi^0, \theta^0) \xrightarrow{\mathcal{D}} N_{d+d_1+d_2+\cdots+d_{k+1}}(0, \bar{i}(\psi^0, \theta^0)).$$

In a similar way, we easily obtain that

$$-\frac{1}{n} \hat{\ell}_{\phi\phi}(\psi^0, \theta^0) \xrightarrow{\mathcal{D}} \bar{i}(\psi^0, \theta_j^0).$$

Therefore, we have that

$$\sqrt{n}(\hat{\phi} - \phi^0) \xrightarrow{\mathcal{D}} N_{d+d_1+d_2+\cdots+d_{k+1}}(0, \bar{i}(\psi^0, \theta_j^0)^{-1}),$$

proving the result. □

Acknowledgements

References

Bahadur, R.R. (1971). *Some Limit Theorems in Statistics*. Philadelphia: SIAM. MR0315820

Battacharya, P.K. (1987). Maximum likelihood estimation of a change-point in the distribution of independent random variables: General multiparameter case. *J. Multivariate Anal.* **23** 183–208. MR0918253

Braun, J.V. and Muller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statist. Sci.* **13** 142–162.

Broemeling, L.D. and Tsurumi, H. (1987). *Econometrics and Structural Change*. New York: Marcel Dekker. MR0922263

Chen, J. and Gupta, A.K. (2000). *Parametric Statistical Change Point Analysis*. Boston: Birkhäuser. MR1761850

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *J. Amer. Statist. Assoc.* **68** 361–368.

Csörgo, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Chichester: Wiley.

Döring, M. (2007). Mehrdimensionale change-point-schätzung mit U-statistiken. Ph.D. thesis, University of Dresden, pp. 1–116 (in German).

Ferger, D. (2001). Exponential and polynomial tail bounds for change-point estimators. *J. Statist. Plann.* **92** 73–109. MR1809698

Fu, Y. and Curnow, R.N. (1990a). Locating a changed segment in a sequence of Bernoulli variables. *Biometrika* **77** 295–305. MR1064801

Fu, Y. and Curnow, R.N. (1990b). Maximum likelihood estimation of multiple change points. *Biometrika* **77** 563–573. MR1087847

Halpern, A.L. (2000). Multiple-change-point testing for an alternating segments model of a binary sequence. *Biometrics* **56** 903–908.

Hawkins, D.M. (2001). Fitting multiple change-point models to data. *Comput. Statist. Data Anal.* **37** 323–341. MR1856677

Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57** 1–17. MR0273727

Hinkley, D.V. (1972). Time-ordered classification. *Biometrika* **59** 509–523. MR0368317

Hinkley, D.V. and Hinkley, E.A. (1970). Inference about the change-point in a sequence of binomial variables. *Biometrika* **57** 477–488. MR0275556

Jandhyala, V.K. and Fotopoulos, S.B. (1999). Capturing the distributional behavior of the maximum likelihood estimator of a changepoint. *Biometrika* **86** 129–140. MR1688077

Jandhyala, V.K. and Fotopoulos, S.B. (2001). Rate of convergence of the maximum likelihood estimate of a change-point. *Sankhyā Ser. A* **63** 277–285. MR1897454

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86. MR0039968

Lombard, F. (1986). The change-point problem for angular: A nonparametric problem. *Technometrics* **28** 391–397.

Móricz, F., Serfling, R. and Stout, W. (1982). Moment and probability bounds with quasi-superadditive structure for the maximum partial sum. *Ann. Probab.* **10** 1032–1040. MR0672303

Reed, W.J. (1998). Determining changes in historical forest fire frequency from a time-since-sire map. *J. Agric. Biol. Environ. Stat.* **3** 430–450. MR1815291

Schervish, M.J. (1995). *Theory of Statistics*. New York: Springer-Verlag. MR1354146

Srivastava, M.S. and Worsley, K.J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *J. Amer. Statist. Assoc.* **81** 199–204. MR0830581

Wald, A. (1949). Note on the consistency of the maximum likelihood estimator. *Ann. Math. Statist.* **20** 595–601. MR0032169