Biometrika (2000), **87**, 2, *pp.* 301–314

Printed in Great Britain

Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation

By J. V. BRAUN

Kings Mountain Research, 1942 Kings Mountain Road, Woodside,
California 94062-4234, U.S.A.
jerome.braun@kmri.com

R. K. BRAUN

Stottler Henke Associates, Inc., 1107 NE 45th Street, Suite 401, Seattle,
Washington 98105, U.S.A.
rbraun@shai-seattle.com

AND H.-G. MÜLLER

Division of Statistics, University of California, One Shields Avenue, Davis,
California 95616, U.S.A.
mueller@wald.ucdavis.edu

Summary

We consider situations where a step function with a variable number of steps provides an adequate model for a regression relationship, while the variance of the observations depends on their mean. This model provides for discontinuous jumps at changepoints and for constant means and error variances in between changepoints. The basic statistical problem consists of identification of the number of changepoints, their locations and the levels the function assumes in between. We embed this problem into a quasilikelihood formulation and utilise the minimum deviance criterion to fit the model; for the choice of the number of changepoints, we discuss a modified Schwarz criterion. A dynamic programming algorithm makes the segmentation feasible for sequences of moderate length. The performance of the segmentation method is demonstrated in an application to the segmentation of the Bacteriophage λ sequence.

Some key words: Bacteriophage lambda; Deviance; Generalised linear model; Model selection; Schwarz criterion; Step function.

1. Introduction

Observations in a regression setting can often be adequately modelled by assuming that the regression function is a step function. Examples are the Nile river flow data (Cobb, 1978), neuronal calcium flow data (Fredkin & Rice, 1992), incidence data (Christensen & Rudemo, 1996) and genetic sequence data. An example concerning the segmentation of DNA sequences is discussed in § 5. The case of estimating a step function by the least squares principle, which is geared towards normal data, was thoroughly investigated by

Yao & Au (1989); compare also Yao (1988). For an overview of the related area of changepoint problems, see Bhattacharya (1994).

The basic idea of our approach is to use quasi-deviance to measure the quality of the fitted model and to adapt the Schwarz criterion for the choice of the number of changepoints. This approach is geared towards nonnormal data such as quasi-Poisson distributed incidence data or multinomial genetic sequence data. The least squares approach is included as a special case. The Schwarz criterion includes a penalty term which increases with the complexity of the model, as measured by the number of changepoints included in the model. Once the number of changepoints is determined, their locations and the constant levels of the regression function on the segments defined by the changepoints are obtained by minimising the quasi-deviance. Computational costs increase exponentially for brute force search over all possible segmentations to achieve global minimisation. Efficient searches to locate optimal segmentations are a necessity. Such methods will be discussed in § 4; compare also Fu & Curnow (1990).

The choice of the number of changepoints or segments in particular is a difficult problem, especially if no prior information is assumed. Various approaches appear in the literature, such as sequential likelihood ratio tests (Haccou & Meelis, 1988), the Schwarz criterion (Yao, 1988) and the assumption of a prior distribution for the number of changepoints in a Bayesian approach (Barry & Hartigan, 1992).

Bayesian approaches provide a particularly promising avenue for the number-of-segments problem, if one can justify a prior distribution. An example is the step-function case where the changepoints arise from a renewal process with identically geometrically distributed inter-arrival times. If in addition the segment means are mutually independent from a common normal prior, and additive measurement noise is also normally distributed, Yao (1984) determined Bayes estimators of the mean function as well as linear approximations to the Bayes estimators. These estimators were found to be sensitive to the normality assumption. Another example is the product partition approach (Hartigan, 1990; Barry & Hartigan, 1993) which includes implicitly some cases with non-constant variance function such as the Poisson case. The choice of whether or not to impose the extra structure required by Bayesian models is best made on a case-by-case basis, depending on the nature of the segmentation task. Our emphasis is to model nonnormal error structure of a wide variety of mean-variance structures, in a non-Bayesian framework.

The basic model and our proposed estimators are described in § 2. Consistency properties for the estimated number and location of the changepoints, and the estimated levels of the regression function on the segments are discussed in § 3. Computational issues are the topic of § 4. The methods are illustrated with the segmentation of the DNA sequence of Bacteriophage λ in § 5. Proofs and auxiliary results follow in an appendix.

2. Model and estimators

We assume that the data Y_i, for $i=1, \ldots, n$, are independent and are sampled at equispaced points

$$t_i = i/n \quad (i=1, \ldots, n).$$

Given a variance function $V(.)$ and a mean regression function $\mu(.)$, let

$$E(Y_i)=\mu(t_i), \quad \operatorname{var}(Y_i)=\sigma^2 V\{\mu(t_i)\}, \tag{1}$$

where σ^2 is an overdispersion parameter.

The regression function $\mu(.)$ is defined, first of all, by specifying the number R of changepoints along with their locations $0 < \tau_1 < \tau_2 < \ldots < \tau_R < 1$. We assume without loss of generality that the support of $\mu(.)$ is $[0, 1]$ and, for convenience, set $\tau_0 = 0$, $\tau_{R+1} = 1$. Thus,

$$\mu(t) = \begin{cases} \mu_1 & \text{if } t = 0, \\ \mu_r & \text{if } \tau_{r-1} < t \leqslant \tau_r,\ r = 1, \ldots, R+1, \end{cases} \tag{2}$$

where μ_i denotes the assumed constant mean between the $(i-1)$th and the ith changepoints τ_{i-1} and τ_i. The regression function $\mu(.)$ is a step function with $(R+1)$ segments $(\tau_i, \tau_{i+1}]$, on which it is constant. It follows from (1) that the variance of the observations satisfies a relation analogous to (2), i.e. the variance of the observations is constant in between changepoints, but may vary between segments. In addition, we make the stronger assumption that the distributions of all data Y_i recorded from the same segment are identical.

As in Wedderburn (1974), the quasilikelihood for the mean μ of one observation $Y = y$ is defined by the integral

$$Q(\mu, y) = \int_y^{\mu} \frac{y-t}{\sigma^2 V(t)}\, dt.$$

A corresponding goodness-of-fit criterion, the quasi-deviance, is defined analogously to the deviance in the generalised linear model and is given by

$$D(\mu, y) = -\sigma^2 Q(\mu, y) = \int_{\mu}^{y} \frac{y-t}{V(t)}\, dt.$$

Note that D does not involve σ^2. A factor of 2 is often added, leading to $D^*(\mu, y) = 2D(\mu, y)$, because differences in the deviances between nested models will then follow an asymptotic χ^2 distribution. Maximising the quasilikelihood is equivalent to minimising the quasi-deviance.

Assume that data $Y_{l_1}, \ldots, Y_{l_2}$, for $l_1 \leqslant l_2$, are from the same segment between two changepoints. A simple calculation shows that the maximum quasilikelihood, or, equivalently, minimum quasi-deviance estimator for μ based on these data is the sample mean,

$$\hat{\mu}(l_1, l_2) = \arg\max_{\mu} \sum_{i=l_1+1}^{l_2} Q(\mu, Y_i) = \frac{1}{l_2 - l_1} \sum_{i=l_1+1}^{l_2} Y_i = \bar{Y}(l_1, l_2).$$

Our objective is to minimise the total quasi-deviance

$$G(l_1, \ldots, l_R) = \sum_{r=1}^{R+1} \sum_{i=l_{r-1}+1}^{l_r} D\{\hat{\mu}(l_{r-1}, l_r), Y_i\},$$

where $l_0 = 0$, $l_{R+1} = n$, and the interval segment defined by (l_i, l_{i+1}) is $(l_i/n, l_{i+1}/n]$. We assume that $l_1, \ldots, l_R$ are always ordered and define $(l_1, \ldots, l_R)$ to be the set of distinct indices, should some of the l_i's coincide.

Let $0 = l_0 < \hat{l}_1 < \ldots < \hat{l}_R < l_{R+1} = n$ be the values which minimise $G(l_1, \ldots, l_R)$ for a given number of changepoints R. The maximum quasilikelihood estimators for the changepoint locations, $\hat{\tau}_1, \ldots, \hat{\tau}_R$, are defined as

$$\hat{\tau}_r = \hat{l}_r / n \quad (r = 1, \ldots, R). \tag{3}$$

The maximum quasilikelihood mean parameter estimators are just the maximum quasilikelihood estimators for the means on each segment using the changepoint estimators; that is

$$\hat{\mu}_r = \hat{\mu}(\hat{l}_{r-1}, \hat{l}_r) \quad (r = 1, \ldots, R + 1). \tag{4}$$

The following average deviance will serve a similar function to that of the least squares estimator of the error variance σ^2 in a homoscedastic normal regression model:

$$\hat{v}_R^2 = \frac{1}{n} G(\hat{l}_1, \ldots, \hat{l}_R).$$

All these quantities are defined analogously in the more realistic situation where the number of changepoints must be estimated from the data by substituting an estimate $\hat{R}$ for R. The data-based choice of R corresponds to a model selection problem, which is addressed in § 3.

3. Asymptotic results and model selection

In order to obtain consistency properties for estimators $\hat{\tau}_r$ and $\hat{\mu}_r$, for $r = 1, \ldots, R$, we need an identifiability condition on adjacent means; if adjacent means were equal there would not be a changepoint. Conditions on higher-order moments of the observations and lower bounds on the variance function which may recede to zero asymptotically are also needed. This leads to the following assumptions.

Assumption 1. For each $r = 1, \ldots, R$, $\mu_r \neq \mu_{r+1}$.

Assumption 2. For all $i = 1, \ldots, n$, $E(Y_i^{2m}) < \infty$ for some $m \geqslant 3$.

Assumption 3. For data $U_1, \ldots, U_n$, with fixed mean μ and the same mean-variance structure as the Y_i's, if $\bar{U}(i, j) = (j - i + 1)^{-1} \sum_{l=i}^{j} U_l$, for $i \leqslant j$, it holds that

$$\max_{0 < i < j \leqslant n} \frac{D\{\mu, \bar{U}(i, j)\}}{\{\bar{U}(i, j) - \mu\}^2} = o_p(n^{1/3}).$$

We note that Assumption 3 is used for technical purposes of estimation only. One may easily establish that the quasi-normal, quasi-Poisson, quasi-binomial and quadratic variance function $V(\mu) = \mu^2(1 - \mu)^2$ models all satisfy Assumption 3. More details and further discussion of this condition are given in the Appendix.

We obtain the following asymptotic results for the estimators of the changepoints (3) and the estimators of the levels on the segments (4). The first result is on consistency of the estimators of the changepoint locations. It is clear that the locations of the changepoints can be determined at best to an order of n^{-1}, since the data themselves have that spacing. The n^{-1} rate is indeed obtained, as the following result shows.

Theorem 1. *We have that* $n(\hat{\tau}_r - \tau_r) = O_p(1)$, *for* $r = 1, \ldots, R$.

This and the following results require Assumptions 1–3, and the proofs are sketched in the Appendix. These results parallel those obtained by Yao & Au (1989) for the case of least squares estimation of a step function.

Regarding limiting distributions for the estimators of the changepoints and for the estimators of the mean parameters, it is known from previous work (Bhattacharya, 1994) that the limiting distribution for the estimators of the locations of the changepoints in a sequence of normally distributed random variables is that of the minimum of a two-sided symmetric random walk. As a result of there being different variances, the result here is more complex; the limit distribution is that of the minimum of a two-sided asymmetric random walk.

THEOREM 2. *As $n \to \infty$, $\hat{\tau}_1, \ldots, \hat{\tau}_R$ are asymptotically independent, and $n\hat{\tau}_r - [n\tau_r]$ converges in distribution to the location L_r of the minimum for the random walk $\{\ldots, Z_{-1}^{(r)}, Z_0^{(r)}, Z_1^{(r)}, \ldots\}$, where $Z_0^{(r)} = 0$ and*

$$Z_j^{(r)} = \begin{cases} 2\sum_{i=1}^{j} \int_{\mu_r}^{\mu_{r+1}} \frac{W_i - t}{V(t)}\, dt & \text{if } j = 1, 2, \ldots, \\ 2\sum_{i=j}^{-1} \int_{\mu_{r+1}}^{\mu_r} \frac{W_i - t}{V(t)}\, dt & \text{if } j = -1, -2, \ldots; \end{cases}$$

the W_i are independently distributed with mean μ_r if $j < 0$, and mean μ_{r+1} if $j > 0$, and variance function $V(.)$.

THEOREM 3. *As $n \to \infty$, the random variables $\surd n(\hat{\mu}_r - \mu_r)$, for $r = 1, \ldots, R+1$, are asymptotically independent and normally distributed with means 0 and variances $V(\mu_r)/(\tau_r - \tau_{r-1})$. Also, $\surd n(\hat{\mu}_r - \mu_r)$, for $r = 1, \ldots, R+1$, and $n\hat{\tau}_r - [n\tau_r]$, for $r = 1, \ldots, R$, are asymptotically independent for different r.*

A further problem is the selection of the number of changepoints R. The Schwarz criterion was proposed and its consistency established for the case of normal observations in the work of Yao (1988) and Yao & Au (1989). An analogous result holds for the maximum quasilikelihood case. Let R^0 denote the true unknown number of changepoints. We assume that there is a finite upper bound R_U to the number of changepoints. The idea is to try to strike a balance between the quasi-deviance and the number of changepoints, by assessing a penalty for including too many changepoints.

THEOREM 4. *Define*

$$\hat{R} = \arg\min_{R} (n \log \hat{v}_R^2 + RC_n),$$

where $R < R_U$ and $\{C_n\}$ is a sequence which satisfies $C_n n^{-2/m} \to \infty$, and $C_n/n \to 0$. Then $\mathrm{pr}(\hat{R} = R^0) \to 1$ as $n \to \infty$.

We remark that, for the contiguous case, where $|\mu_{r+1} - \mu_r| = \Delta\gamma_n^{-1}$ for some sequence $\gamma_n \to \infty$, we obtain under suitable regularity conditions that $(n\gamma_n^{-1})(\hat{\tau}_r - \tau_r) = O_p(1)$. A second important extension is to the case of multivariate data $Z_i \in \Re^p$. In this case, we require

$$E(Z_i) = \mu(i/n), \quad \mathrm{cov}(Z_i) = \sigma^2 W\{\mu(i/n)\}.$$

One finds that the theoretical results for the univariate case readily generalise to the multivariate case, as long as the covariance matrix W meets the conditions for path-independence set forth in Chapter 9 of McCullagh & Nelder (1989).

As an example, we consider the multinomial distribution with $(q+1)$ categories. Here we assume that $\gamma < \mu_i < 1 - \gamma$ for some γ with $0 < \gamma < \frac{1}{2}$, where $\mu = (\mu_1, \ldots, \mu_q)^{\mathrm{T}}$ is the mean vector, subject to $\sum \mu_i + \mu_{q+1} = 1$. We have $Z \in [0, 1]^q$, and

$$W(\mu) = \mathrm{diag}\, \mu - \mu\mu^{\mathrm{T}}.$$

The mean vector μ specifies the probability of occurrence for each of the q categories. The inverse of the covariance function is

$$W^{-1}(\mu) = \operatorname{diag}\left(\frac{1}{\mu_1}, \ldots, \frac{1}{\mu_q}\right) + \frac{1}{\mu_{q+1}} ee^{\mathrm{T}},$$

where $e = (1, \ldots, 1)^{\mathrm{T}}$. This is the second derivative matrix of $b^*(\mu) = \sum_{j=1}^{q+1} \mu_j \log \mu_j$, a convex function. Fixing μ, we find that the deviance is

$$D(\mu, Y) = -\sum_{j=1}^{q+1} \left\{(Y_j - \mu_j) + Y_j \log \frac{\mu_j}{Y_j}\right\},$$

and the multivariate analogue of Assumption 3 is seen to be satisfied.

4. Segmentation algorithm and simulation results

There are two main issues to consider in the implementation of the quasi-deviance segmentation method described above. The first is the choice of criterion for the determination of the number of changepoints, and the second is the choice of the segmentation algorithm.

The first issue is the more difficult one. We require a concrete choice of the C_n penalty term to determine the number of changepoints in the quasilikelihood case. Yao (1988) gave a Bayesian justification for the Schwarz criterion which results in $C_n = 0{\cdot}5 \log n$ in the normal case. In practice, empirical studies are needed to determine suitable values for C_n.

The second issue is computational. When choosing the locations of changepoints (3), the number of possible choices of R changepoints in a sequence of length n is $n!/R!(n-R)!$. An exact solution for the global optimum may be found with $O(n^2)$ computational effort by use of a dynamic programming algorithm reported in Auger & Lawrence (1989). A brief discussion of the algorithm is given in the Appendix. For Monte Carlo studies and data analysis, we implemented this algorithm and several others in C++. The C++ code is available from the authors on request.

We report the results of a simulation study regarding the choice of the number of changepoints. For the penalty term C_n in the Schwarz criterion, we considered the simple functional form n^{α}. The parameter α allows us to tune the increase in the penalty for overfitting as n increases. Note that Assumption 2 is satisfied for all finite m, so that we have considerable freedom in choosing α. In simulations, we optimised over α using the absolute difference between true and estimated number of changepoints as a measure of fit.

Three factors were varied: sequence length with values $n = 1000$, 5000 and 10 000; true number of changepoints with values $R_0 = 0$, 1, 5 and 10; and range of jump sizes J with small, medium and large levels. Two replicates were drawn for each combination of levels. The sequences were analysed using the upper bound $R_U = 20$ on the number of changepoints.

For each replicate, a random multinomial sequence with four components, representing a DNA sequence, was generated as follows. Given R_0, changepoint locations were chosen randomly in the sequence $1, \ldots, n$. The initial mean vector was obtained by normalising a set of uniform deviates; that is $\mu_i = U_i / \sum_i U_i$, for $i = 1, 2, 3, 4$, where $U_i \sim \mathrm{Un}(0, 1)$. Jumps were made on the logistic scale, and the resulting vectors normalised; that is, μ_i' was obtained by normalising expit(logit $\mu_i + U_i$), for $i = 1, 2, 3, 4$, where $U_i \sim \mathrm{Un}(-J, J)$, logit is the logistic transform, and expit is its inverse. The ranges of jump size were defined as small, medium and large for $J = 0{\cdot}2$, $0{\cdot}5$ and $0{\cdot}8$ respectively.

Errors in the estimated number of changepoints were assessed by the differences $\hat{R} - R_0$. These differences were not sensitive to n; however, they did generally increase with R_0. To summarise, for each jump size and sequence length, we calculated the median difference pooled over number of changepoints. As an additional overall measure of goodness of fit, we also calculated the sum of the absolute errors for each α, pooled over jump size, sequence length and number of changepoints. Table 1 displays the results for a range of α's. As the sequence length increases to 10 000, the criterion of median difference is less sensitive to the choice of α. In general, a smaller α provides a smaller penalty and leads to overfitting. If we use the criteria above, choosing $\alpha = 0{\cdot}23$ works reasonably well over the factors in this study, so as a point of departure we used this value in the data analysis of § 5.

Table 1. *Summary results from simulation study. Values are medians of* $\hat{R} - R_0$ *pooled over* R_0. *Values of* α *above and below those shown did not improve estimation*

						α				
Jump	n	0·08	0·11	0·14	0·17	0·20	0·23	0·26	0·29	0·32
0·2	1000	17	17	17	17	3·5	−0·5	−2	−2	−2
	5000	17	17	17	11	−1	−2·5	−2·5	−2·5	−2·5
	10 000	17	17	17	16·5	−1·5	−2	−2	−2	−2
0·5	1000	17	17	17	17	9	1	−2	−2	−2
	5000	17	17	17	15	2·5	−1·5	−1·5	−1·5	−1·5
	10 000	17	17	17	15	0	−0·5	−1	−1·5	−1·5
0·8	1000	17	17	17	17	9·5	−1	−1·5	−1·5	−1·5
	5000	17	17	17	16·5	2·5	−0·5	−1	−1	−1·5
	10 000	17	17	17	12·5	−0·5	−1	−1	−1·5	−1·5
Pooled sum of absolute values		153	153	153	137·5	30	10·5	14·5	15·5	16

Addressing the computational issues, we note that, while the dynamic programming algorithm works well for moderately sized sequences, $n \leqslant 10\,000$ say, it does not scale to much larger sequences. One would like to have an $O(n)$ algorithm to be able to work with truly large sequences. We examined a variety of such methods. Simple approaches include binary segmentation, i.e. splitting all previous segments into two, ordered segmentation, i.e. taking the best split out of the previous segments, and stochastic search techniques. These approaches are fast, requiring only $O(n)$ computations, but they are not guaranteed to find the global optimum in finite samples, and in simulations not reported here they were found to perform noticeably worse than the dynamic programming approach.

With more model structure, it is possible to produce models which allow faster computation. Examples of this are seen in several Bayesian approaches. For estimation of the underlying mean function, the approach of Yao (1984) requires $O(n^3)$ computational effort for exact calculations, while a filtering approximation can be performed with $O(n)$ effort. Barry & Hartigan (1993) showed that their product partition approach admits acceptable approximate results with $O(n)$ effort using Markov sampling; exact results again require $O(n^3)$ effort. It remains to be seen whether or not similar approaches can be developed for the quasilikelihood case.

5. Application to DNA sequence segmentation

We illustrate the quasilikelihood segmentation method for multinomial outcomes with the complete DNA sequence of Bacteriophage λ (Skalka, Burgi & Hershey, 1968; Churchill, 1992). The sequence is 48 502 bases in length, and each base is one of either adenosine (A), cytosine (C), guanine (G) or thymine (T). Thus, the observations are multinomial vectors.

Segmentation of DNA sequences serves several biological needs; see Churchill (1989), Curnow & Kirkwood (1989) and Karlin & Brendel (1993) for a discussion of pertinent issues. Specific segmentation methods which have been discussed include hidden Markov chains (Churchill, 1992), maximum likelihood segmentation (Fu & Curnow, 1990), high-order Markov chains (Scherer, McPeek & Speed, 1994), spectral methods (Stoffer, Tyler & McDougall, 1993), and scan statistics (Wallenstein, Naus & Glaz, 1994). While DNA sequences are clearly correlated, the independence assumption has been shown to provide a reasonable first-order approximation in many situations; see Braun & Müller (1998) for a review of such situations. An excellent on-line bibliography of correlation in DNA sequences is maintained by W. Li at http://linkage.rockefeller.edu/wli/dna_corr.

We used the Auger–Lawrence (1989) algorithm, using $C_n = n^{0{\cdot}23}$; no restriction was set on minimum segment size. Figure 1(a) shows the resulting value of the Schwarz criterion as a function of the number of changepoints. The minimum is found at $\hat{R} = 8$ changepoints, which is in general accordance with previous results for these data, using other methods. Since it is difficult to choose the number of changepoints, we show in Fig. 1(b) for comparison purposes the results of quasilikelihood segmentation for other assumed numbers of changepoints, up to $R_U = 40$. The locations remain stable: as more changepoints are added, this leads to subsegmentation of previously identified segments. Table 2 presents a brief summary of the estimated parameter vectors assuming $\hat{R} = 8$ changepoints.

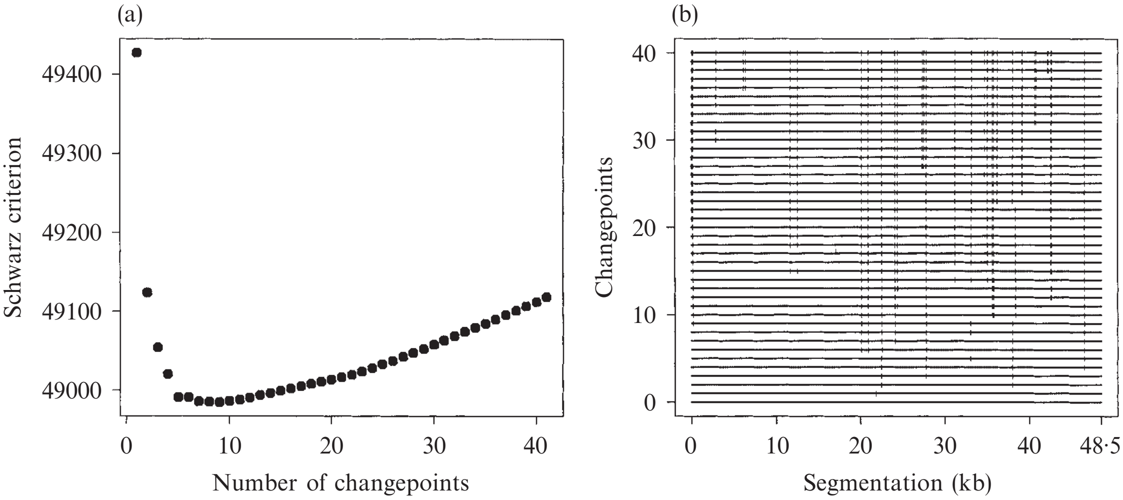

Fig. 1. (a) Value of the Schwarz criterion versus number of changepoints, using $C_n = n^{0{\cdot}23}$, for the multinomial segmentation of Bacteriophage λ; the minimum is found at $\hat{R} = 8$ changepoints. (b) Optimal minimum quasi-deviance segmentations of the multinomial Bacteriophage λ sequence for up to 40 changepoints.

To display the segmentation requires a choice of projections; since the mean vectors are constrained to sum to one, three suffice. Figure 2 displays the segmentation assuming $\hat{R} = 8$ changepoints for proportions of A + C, A + G and A + T. In order to compare the

Table 2. *Bacteriophage λ analysis. Values are estimated segments and estimated mean vectors for each segment*

Segment	(A, C, G, T)
0–20 091	(0·23, 0·25, 0·32, 0·20)
20 092–20 919	(0·29, 0·29, 0·30, 0·11)
20 920–22 544	(0·26, 0·24, 0·27, 0·23)
22 545–24 117	(0·29, 0·14, 0·16, 0·40)
24 118–27 829	(0·29, 0·20, 0·18, 0·33)
27 830–33 082	(0·23, 0·26, 0·22, 0·29)
33 083–38 029	(0·27, 0·22, 0·21, 0·31)
38 030–46 528	(0·30, 0·23, 0·26, 0·22)
46 529–48 502	(0·27, 0·18, 0·22, 0·33)

A, adenosine; C, cytosine; G, guanine; T, thymine.

results of the segmentation with known features of Bacteriophage λ, we made use of annotation information contained in GenBank, extracted from the GenBank database as LAMCG, Bacteriophage lambda, complete genome. Figure 3 shows the segmentation assuming $\hat{R} = 8$ changepoints; plotted above it are known coding sequences.

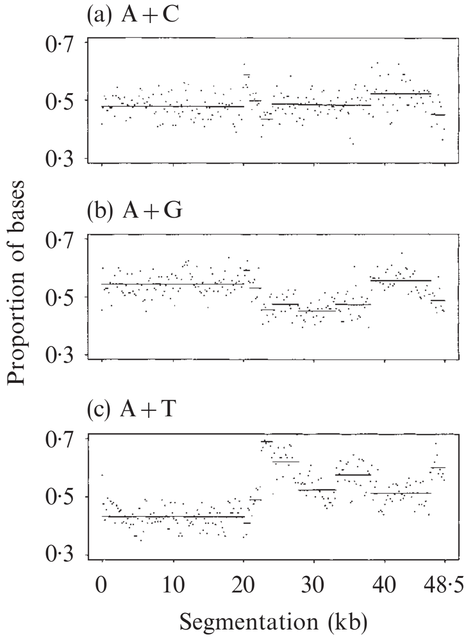

Fig. 2. Segmentation assuming $\hat{R} = 8$ changepoints. Shown are proportions of (a) adenosine + cytosine, (b) adenosine + guanine and (c) adenosine + thymine, in nonoverlapping 200-base bins.

The large segment, from 0 to 20 kb, represents a stretch of the phage genome which primarily contains code for the construction of the phage, e.g. head and tail components. The middle section, from 20 kb to about 38 kb, represents a stretch of the genome which contains some complex coding sequences that can produce different products by starting in different reading frames. The segmentation reflects broad shifts in transcription direc-

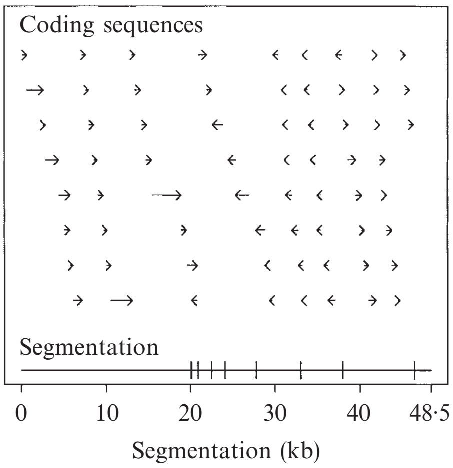

Fig. 3. Partially annotated Bacteriophage λ sequence with segmentation obtained assuming $\hat{R}=8$ changepoints. Shown above this are coding sequences with transcription direction.

tion; the major shifts are already detected with $\hat{R}=3$ and successive segmentations paint a more refined picture.

6. Concluding remarks

We have seen that the proposed quasilikelihood procedure allows for efficient segmentation of multinomial data as they occur in biological sequences. The proposed quasilikelihood segmentation algorithm will also be useful for the segmentation of quasi-Poisson and similar sequences which include a non-constant variance function. In fact, even cases where the variance function is unknown could be included. In such situations, the variance function can be estimated and a nonparametric quasilikelihood can be constructed (Chiou & Müller, 1999). The construction of efficient $O(n)$ algorithms which scale up to large sequences remains an important research topic.

Acknowledgement

This research was supported by grants from the National Science Foundation and the National Security Agency of the U.S.A. The authors would like to thank the editor and referees for suggestions that substantially improved the exposition.

Appendix

Technical aspects

Dynamic Programming Algorithm. The algorithm relies upon the calculation of the deviance over each possible segment. For small sequences, these quantities can be stored in memory. This requires a half-matrix of $n(n-1)/2$ unique storage addresses. For large sequences, we also developed a version of the algorithm which stores appropriate columns of the matrix of the calculated deviances.

The algorithm is as follows. Assume we wish to estimate R changepoints. Let $c_{i,j}^{r+1}$ be the minimum quasi-deviance obtained by the best partition of the sequence $Y_{i+1}, \ldots, Y_j$ into $r+1$ segments with

r changepoints. Then

$$c^{R+1}_{1,n} = G(\hat{l}_1, \ldots, \hat{l}_R),$$

giving the global minimum quasi-deviance estimator.

One starts with

$$c^1_{i,j} = \sum_{l=i+1}^{j} D\{\hat{\mu}(i, j), Y_l\},$$

for $r = 0$ and for all $0 \leqslant i \leqslant j \leqslant n$.

Then, for each $r = 1, \ldots, R$, let

$$c^{r+1}_{1,j} = \min_{1 \leqslant l \leqslant j} (c^r_{1,l} + c^1_{l+1,j}).$$

Auger & Lawrence (1989) provide a proof that this algorithm leads to the global minimiser. Their proof extends readily to the case where a lower bound on the segment size is given.

Discussion of Assumption 3. Note that, in practice, we may impose a lower bound on the variance by defining a modified variance function as $V_n(.) = V(.) \vee \alpha_n^{-1}$, where α_n is a strictly increasing sequence with $\alpha_n = o(n^{1/3})$, not depending on μ. The truncated variance is used for estimation purposes only and the data are assumed to come from the mean and variance structure given by μ and $V(\mu)$. We note that, for any fixed $\mu \in \Omega$ and for all large enough n, $V_n(\mu) = V(\mu)$; also for all n, $V_n(\mu) \to V(\mu)$ as $\mu \to \infty$. Using the truncated variance, we find that

$$\max_{0 \leqslant i < j \leqslant n} \frac{D\{\mu, \bar{Y}(i, j)\}}{\{\bar{Y}(i, j) - \mu\}^2} = o(n^{1/3}),$$

so that Assumption 3 is satisfied. By explicitly calculating the deviances, we find that Assumption 3 is satisfied for many distributions. L'Hôpital's rule implies that

$$\lim_{y \to \mu} \frac{D(\mu, Y)}{(Y - \mu)^2} = \frac{1}{2V(\mu)}.$$

This fact and the continuity of $D(\mu, .)$ imply that Assumption 3 is satisfied for the cases of a normal distribution, Poisson distribution with $0 < \mu_0 \leqslant \mu$, binomial distribution and distributions with quadratic variance function with $0 < \mu_0 \leqslant \mu \leqslant 1 - \mu_0$, and also for exponential and inverse Gaussian distributions with $0 < \mu_0 \leqslant \mu$.

Auxiliary lemmas. In the following, let α_n be a sequence such that $\alpha_n \to \infty$, $\alpha_n n^{-1/3} \to 0$ as $n \to \infty$.

LEMMA A1. *Suppose that $Y_1, \ldots, Y_n$ are independently and identically distributed with mean μ and variance $V(\mu)$. Then, as $n \to \infty$, under Assumptions* 2 *and* 3,

$$\max_{0 \leqslant i < j \leqslant n} (j - i) D\{\mu, \bar{Y}(i, j)\} = O_p(n^{2/m} \alpha_n).$$

Proof. Let $U_i = Y_i - \mu$, for $i = 1, \ldots, n$. Then

$$\max_{0 \leqslant i < j \leqslant n} (j - i) D\{\mu, \bar{Y}(i, j)\} \leqslant \max_{0 \leqslant i < j \leqslant n} (j - 1) \frac{(U_{i+1} + \ldots + U_j)^2}{j - 1} \max_{0 \leqslant i < j \leqslant n} \frac{D\{\mu, \bar{Y}(i, j)\}}{\{\bar{Y}(i, j) - \mu\}^2}.$$

Applying Lemma 1 of Yao & Au (1989) to the first term shows that this term is $O_p(n^{2/m})$. An application of Assumption 3 to the second term completes the proof. □

Next we characterise the behaviour of the deviance function G. The proofs are straightforward.

LEMMA A2. *Suppose that $Y_1, \ldots, Y_n$ are independently and identically distributed with mean μ and variance $V(\mu)$. Let $l_1 < \ldots < l_R$ be arbitrarily chosen distinct indices chosen from the set $\{1, \ldots, n - 1\}$; set $l_0 = 0$ and $l_{R+1} = n$. Then*

$$G(l_1, \ldots, l_R) - \sum_{i=1}^{n} D(\mu, Y_i) \geqslant -(R + 1) \max_{0 \leqslant i < j \leqslant n} (j - i) D\{\mu, \bar{Y}(i, j)\}.$$

Lemma A3. *We have that*

$$G([n\tau_1], \ldots, [n\tau_R]) - \sum_{r=1}^{R+1} \sum_{i=[n\tau_{r-1}]+1}^{[n\tau_r]} D(\mu_r, Y_i) = O_p(1),$$

where $[x]$ *denotes the greatest integer less than* x.

Lemma A4. *Under Assumptions* 1–3, $\hat{\tau}_r - \tau_r = o_p(1)$, *for* $r = 1, \ldots, R$.

Proof. The proof adopts a strategy developed by Yao & Au (1989) for the least squares estimation of a step function and we sketch only the main steps. Choose $\delta > 0$ so that $\tau_s + 2\delta < \tau_{s+1} - 2\delta$ for each $s = 0, \ldots, R$. Define

$$A_r(n, \delta) = \left\{ \left(\frac{l_1}{n}, \ldots, \frac{l_R}{n} \right) : 0 < l_1 < \ldots < l_R < n, \left| \frac{l_s}{n} - \tau_r \right| > \delta,\ s = 1, \ldots, R \right\},$$

for $r = 1, \ldots, R$. For any fixed $(l_1/n, \ldots, l_R/n) \in A_r(n, \delta)$, we have that

$$G(l_1, \ldots, l_R) \geqslant G\{l_1, \ldots, l_R, [n\tau_1], \ldots, [n\tau_{r-1}], [n(\tau_r - \delta)], [n(\tau_r + \delta)], [n\tau_{r+1}], \ldots, [n\tau_R]\}.$$

Using Lemma A1 and Lemma A2, we obtain, under Assumption 3, that

$$\min_{(l_1/n, \ldots, l_R/n) \in A_r(n,\delta)} G(l_1, \ldots, l_R) \geqslant G([n\tau_1], \ldots, [n\tau_R]) + O_p(n^{2/m}\alpha_n) + [n\delta]\eta_r + o_p(n),$$

for a constant $\eta_r > 0$.

Thus we have that, with probability approaching 1,

$$\min_{(l_1/n, \ldots, l_R/n) \in A_r(n,\delta)} G(l_1, \ldots, l_R) > G([n\tau_1], \ldots, [n\tau_R]).$$

This implies that $(\hat{\tau}_1, \ldots, \hat{\tau}_R) \in A_r(n, \delta)$ with probability approaching 0 as $n \to \infty$. By the choice of δ, this implies that with probability approaching 1 exactly one of the $\hat{\tau}_1, \ldots, \hat{\tau}_R$ is between $\tau_r - \delta$ and $\tau_r + \delta$, for $r = 1, \ldots, R$; this will be $\hat{\tau}_r$. Since δ can be taken arbitrarily small, it follows that $\hat{\tau}_r - \tau_r = o_p(1)$. □

Proof of Theorem 1. We sketch the proof for the cases $r = 2, \ldots, R-1$; the cases $r = 1$ and $r = R$ are treated similarly. Fix $\varepsilon > 0$. It will suffice to show that there exists $M > 0$ such that $\text{pr}(n|\hat{\tau}_r - \tau_r| > M) < \varepsilon$ for all sufficiently large n.

Choose $\delta > 0$ such that $\tau_s + 2\delta < \tau_{s+1} - 2\delta$ for each $s = 0, \ldots, R$, and define

$$B(n, \delta) = \left\{ \left(\frac{l_1}{n}, \ldots, \frac{l_R}{n} \right) : 0 < l_1 < \ldots < l_R < n, \left| \frac{l_s}{n} - \tau_s \right| < \delta,\ s = 1, \ldots, R \right\},$$

$$B_r(n, \delta, M) = \left\{ \left(\frac{l_1}{n}, \ldots, \frac{l_R}{n} \right) \in B(n, \delta) : l_r - n\tau_r < -M \right\}.$$

By Lemma A4, $\text{pr}\{(\hat{\tau}_1, \ldots, \hat{\tau}_R) \in B(n, \delta)\} > 1 - \varepsilon/4$ for large enough n. We may show that there exists a suitable $M > 0$ such that, for large enough n, $\text{pr}\{n(\hat{\tau}_r - \tau_r) < -M\} < \varepsilon/2$, and similarly $\text{pr}\{n(\hat{\tau}_r - \tau_r) > M\} < \varepsilon/2$, so that $\text{pr}(n|\hat{\tau}_r - \tau_r| > M) < \varepsilon$. □

Proof of Theorem 2. We note that

$$E\{Z_j^{(r)}\} = \begin{cases} D(\mu_r, \mu_{r+1}) & \text{if } j > 0, \\ D(\mu_{r+1}, \mu_r) & \text{if } j < 0. \end{cases}$$

Thus, as $|j| \to \infty$, the random walk $\{Z_j^{(r)}\}$ diverges to ∞.

One may show that, for sufficiently large $M > 0$,

$$G([n\tau_1] + i_1, \ldots, [n\tau_R] + i_R) - G([n\tau_1], \ldots, [n\tau_R])$$

converges jointly in i_r with $|i_r| \leqslant M$, for $r = 1, \ldots, R$, in distribution to $\sum_{r=1}^{R} Z_{i_r}^{(r)}$, for sufficiently large n. Letting $M \to \infty$ in conjunction with some additional arguments comparing $\arg\min G(l_1, \ldots, l_R)$ with $\min_{|j| \leqslant M} Z_j^{(r)}$ yields the result. □

Proof of Theorem 3. By Theorem 1, we have that $\hat{\tau}_r - \tau_r = O_p(n^{-1})$. Note that

$$\hat{\mu}(n\hat{\tau}_{r-1}, n\hat{\tau}_r) - \hat{\mu}([n\tau_{r-1}], [n\tau_r]) = O_p(n^{-1}).$$

By the properties of the quasilikelihood estimator and Slutsky's theorem we find that

$$\sqrt{n}\{\hat{\mu}(n\hat{\tau}_{r-1}, n\hat{\tau}_r) - \mu_r\} \to \mathcal{N}\left(0, \frac{V(\mu_r)}{\tau_r - \tau_{r-1}}\right)$$

in distribution independently for each $r = 1, \ldots, R+1$. Furthermore, for a suitably chosen $M > 0$ and for $|i_r| \leqslant M$, for $r = 1, \ldots, R$, we find the representation

$$G([n\tau_1] + i_1, \ldots, [n\tau_R] + i_R) - G([n\tau_1], \ldots, [n\tau_R]) = \sum_{r=1}^{R} S_{i_r}^{(r)} + O_p(n^{-\frac{1}{2}}),$$

for suitably chosen random variables $S_j^{(r)}$, where $S_0^{(r)} = 0$. The result follows from the properties of $\min_{|j| \leqslant M} S_j^{(r)}$ for large constants M. □

Proof of Theorem 4. Suppose $R > R^0$. By Lemma A2,

$$G([n\tau_1], \ldots, [n\tau_{R^0}]) \geqslant G([n\tau_1], \ldots, [n\tau_{R^0}]) - \sum_{r=1}^{R^0+1} (R+1) \max_{[n\tau_{r-1}] < i < j \leqslant [n\tau_r]} (j-i) D\{\mu_r, \bar{Y}(i,j)\}.$$

By Lemma A1 and (A3), $\hat{v}_{R^0}^2 - \hat{v}_R^2 = O_p(n^{2/m-1})$. Analogously,

$$\hat{v}_{R^0}^2 - G([n\tau_1], \ldots, [n\tau_{R^0}])/n = o_p(1).$$

By the law of large numbers,

$$\frac{1}{n} \sum_{r=1}^{R^0+1} \sum_{i=[n\tau_{r-1}]+1}^{[n\tau_r]} D(\mu_r, Y_i) \to \sum_{r=1}^{R^0+1} (\tau_r - \tau_{r-1}) E\{D(\mu_r, Q_r)\} \qquad \text{(A1)}$$

in probability, where Q_r has the same distribution as the Y_i for $i = [n\tau_{r-1}] + 1, \ldots, [n\tau_r]$. Denote the quantity on the right-hand side of (A1) by v^2. By Lemma A3,

$$G([n\tau_1], \ldots, [n\tau_{R^0}])/n - v^2 = o_p(1),$$

and thus $\hat{v}_{R^0}^2 - v^2 = o_p(1)$. Therefore, under Assumption 3, $n \log \hat{v}_{R^0}^2 - n \log \hat{v}_R^2 = O_p(n^{2/m}\alpha_n)$.

Since $C_n n^{-2/m}\alpha_n \to \infty$, we have with probability approaching 1 that

$$n \log \hat{v}_{R^0}^2 + RC_n > n \log \hat{v}_{R^0}^2 + R^0 C_n.$$

Exactly as noted in Yao & Au (1989), with probability approaching 1, the criterion will not overestimate the number of changepoints. By similar arguments, with probability approaching 1 the criterion will also not underestimate the number of changepoints. □

References

Auger, I. E. & Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51**, 39–54.

Barry, D. & Hartigan, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20**, 260–79.

Barry, D. & Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *J. Am. Statist. Assoc.* **88**, 309–19.

Bhattacharya, P. K. (1994). Some aspects of change-point analysis. In *Change-point Problems*, Ed. E. Carlstein, H.-G. Müller and D. Siegmund, pp. 28–56. Hayward, CA: Institute of Mathematical Statistics.

Braun, J. V. & Müller, H.-G. (1998). Statistical methods for DNA sequence segmentation. *Statist. Sci.* **13**, 142–62.

Chiou, J. & Müller, H.-G. (1999). Nonparametric quasi-likelihood. *Ann. Statist.* **27**, 36–64.

Christensen, J. & Rudemo, M. (1996). Multiple change-point analysis of disease incidence rates. *Prev. Vet. Med.* **26**, 53–76.

Churchill, G. A. (1989). Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51**, 79–94.

Churchill, G. A. (1992). Hidden Markov chains and the analysis of genome structure. *Comp. Chem.* **16**, 107–15.

Cobb, G. W. (1978). The problem of the Nile: Conditional solution to a change-point problem. *Biometrika* **65**, 243–51.

Curnow, R. N. & Kirkwood, T. B. L. (1989). Statistical analysis of deoxyribonucleic acid sequence data — a review. *J. R. Statist. Soc.* A **152**, 199–220.

Fredkin, D. R. & Rice, J. A. (1992). Bayesian restoration and single-channel patch clamp recordings. *Biometrics* **48**, 427–48.

Fu, Y.-X. & Curnow, R. N. (1990). Locating a changed segment in a sequence of Bernoulli variables. *Biometrika* **77**, 295–304.

Haccou, P. & Meelis, E. (1988). Testing for the number of change points in a sequence of exponential random variables. *J. Statist. Comp. Simul.* **30**, 285–98.

Hartigan, J. A. (1990). Partition models. *Commun. Statist.* A **19**, 2745–56.

Karlin, S. & Brendel, V. (1993). Patchiness and correlations in DNA sequences. *Science* **259**, 677–80.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. New York: Chapman and Hall.

Scherer, S., McPeek, M. S. & Speed, T. P. (1994). Atypical regions in large genomic DNA sequences. *Proc. Nat. Acad. Sci. USA* **91**, 7134–8.

Skalka, A., Burgi, E. & Hershey, A. D. (1968). Segmental distribution of nucleotides in the DNA of bacteriophage lambda. *J. Molec. Biol.* **34**, 1–16.

Stoffer, D. S., Tyler, D. E. & McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika* **80**, 611–22.

Wallenstein, S., Naus, J. & Glaz, J. (1994). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence. *Biometrika* **81**, 595–601.

Wedderburn, R. W. M. (1974). Quasilikelihood functions, generalized linear models and the Gauss–Newton method. *Biometrika* **61**, 439–47.

Yao, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *Ann. Statist.* **12**, 1434–47.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Prob. Lett.* **6**, 181–9.

Yao, Y.-C. & Au. S. T. (1989). Least-squares estimation of a step function. *Sankhyā* A **51**, 370–81.

[*Received December* 1997. *Revised September* 1999]