PLISKA

STUDIA MATHEMATICA BULGARICA

ПЛИСКА

БЪЛГАРСКИ МАТЕМАТИЧЕСКИ СТУДИИ

Pliska Stud. Math. Bulgar. **17** (2005), 241–248

PLISKA
STUDIA MATHEMATICA
BULGARICA

MODELLING COVARIATES IN MULTIPATH CHANGE

N. Sanjari Farsipour [1]

In the multipath change - point problems, it is often of interest to assess the impact of covariates on the change point itself as well as on the parameter before and after the change point. In this paper, we consider a simple model for the change-point distribution, and then through hazard of change, we include covariates in the change point distribution. Maximum likelihood estimation is discussed.

1. Introduction

The analysis of survival data has a long history and date back to the systematic study of time tables by Germans. Our important points of view in this field, are a vast spectral of some kinds of applications. For example in economy and industry (Lancaster 1990). Also in biology there are some application under topic of analysis of history and events by bolsfeld and Rohor (1995).

We use maximum likelihood methods for parameter estimation or testing hypothesis in a change at a sequence of independent random variable. The time intervals between explosions in British coal mines between 1875 and 1950, in which more than ten people were killed, have been analysed in an early paper by Maguire, Pearson & Wynn (1952). They concluded that the data has an exponential distribution with constant mean over time. Cox & Lewis (1966, ch.3)

[1]The paper was written when the author spend her sabbatical leave at the department of Mathematics and Statistics, Mcgill university of Canada.

2000 *Mathematics Subject Classification:* 62N02

Key words: Covariate, maximum likelihood, modelling, multipath change - point problems.

reanalysed the data with more powerful techniques and found strong evidence that the mean did not remain constant, but that it followed a quadratic trend in time.

This suggests a model for the mean time interval which remains constant up to an unknown point in the sequence and then changes to a different mean which remains constant for the rest of the sequence. The model can be formulated as follows. Let $X_1, ..., X_n$ be the sequence of n independent time intervals between accidents, ordered in time, and let $\mu_i = E(X_i)$ for $i = 1, \ldots, n$. Consider the model for a change in mean after the kth observation

$$H : \mu_i = \begin{cases} \mu & i = 1, \ldots, k \\ \mu^* & i = k+1, \ldots, n \end{cases} \tag{1}$$

where μ and μ^* are the unknown means before and after the unknown change - point k. Since Page (1954) first formally studied stopping rules in the context of quality control, there has been a surge of papers on change-point problems. These have addressed fixed sample size and sequential procedures, frequentis and Bayesian approaches, univariate and multivariate settings, parametric and non parametric models and have allowed the observations to be either independent or dependent. These papers have focussed almost exclusively on what we term the single - path change - point problem where by interest is restricted to a single sequence of observations. The extension to the multipath setting, in which more than one sequence is allowed, each with a possible change point, leads to a considerably expanded class of applications. (In partieular, see Joseph 1989, Joseph & Wolfson 1992, 1993, Joseph, Wolfson, du Berger & Lyle 1996, Joseph, Vandal & Wolfson 1996, Joseph, Wolfson, du Berger & Lyle 1997, and Belisle, Joseph, MacGibbon, Wolfson & du Berger 1998.) For a recent review of the literature on change-point problems see Chen & Gupta (2000) who emphasize univariate and multivariate normal distributions, and see the references cited in Asgharian (2001).

When it is desired to study the effect of covariates in change - point problems one is inevitably led to a multipath setting, each path arising from a different set of covariate values. Therefore, we devote our attention to this issue, which was one of those that arose from a study in 1987 of the effects of calcium supplementation on blood pressure.

Using a randomized clinical trial, Lyle et al. (1987) examined the effects of calcium supplementation on the blood pressure of 75 males, both white and black, aged 19 to 52 years. The subjects were followed for a period of 16 weeks. The first four weeks were taken as a baseline period. During this period, weekly blood

pressure was recorded. After this period, within each racial group, the patients were randomly assigned to a calcium in take group (10 black men and 27 white men) and a placebo group (11 black men and 28 white men). The subjects were then given three calcium tablets per day, and blood pressure measurements were taken every other week for the next 12 weeks, resulting in 6 measurements after the tablets were taken. Lyle et al. (1987) applied standard repeated measures methods to assess the effect of calcium intake on blood pressure in this study.

In this work first we introduce multipath change-point problems and describe a latent Markov structure in change-point problems that is key to our modelling approach, and then we show how change-point models may be specified through the hazard of the change, and how covariates may be introduced through "the hazard of change". The likelihood is derived and the quasi-identifiability of the parameters is stated.

2. The Multipath Change-Point Problem and its Markovian Structure

We start with the single-path, single change-point setting. Suppose that for a given $\tau = k, X_1, \ldots, X_\tau, X_{\tau+1}, \ldots, X_m$ is a sequence of random variables such that $X_1, \ldots, X_k$ have joint distribution F_0 and $X_{k+1}, \ldots, X_m$ have joint distribution $F_1 \neq F_0$. If $\tau < m$, we say that a change has occurred at τ. If $\tau = m$, we say that no change has occurred. In either case, we refer to as a change-point. When τ is unknown, inference about it (or its distribution from a Bayesian perspective) as well as about F_0 and F_1 falls in the field of change-point inference. Note that associated with each X_j we may associate an unobserved random variable θ_j defined as

$$\theta_j = \begin{cases} 0 & if \quad X_j \sim F_0 \\ 1 & if \quad X_j \sim F_1 \end{cases} \tag{2}$$

It is obvious that knowledge of realization of $\theta_1, \ldots, \theta_m$ is equivalent to knowledge of the change point τ. When τ is assumed to be random, as it will be in the sequel, the sequence $\theta_1, \ldots, \theta_m$ is random. Consequently, a sequence of random variables with a random change point may be represented by a sequence of random vectors $(X_1, \theta_1), \ldots, (X_m, \ldots, \theta_m)$. The randomness of τ is an introduction to random (τ_i) effects model. In a multipath change-point problem, there are several paths each having a possible change point. The observations may be described by a

matrix

$$\begin{matrix} \tau_1 \\ \vdots \\ \tau_n \end{matrix} \begin{pmatrix} X_{11} & \cdots & X_{1m} \\ \vdots & & \\ X_{n1} & \cdots & X_{nm} \end{pmatrix} \tag{3}$$

This is a rather general framework which covers panel data, balanced longitudinal data and repeated measurements as special cases. For random change points $\tau_1, \ldots, \tau_n$, conditional on $\tau_i = k_i, X_{i1}, \ldots, X_{ik_i}$ have joint distribution, and have joint distribution $F_{1i} \neq F_{0i}$, for $i = 1, \ldots, m$. The will be assumed to be independent. We will further assume that the observations in the above matrix are row-wise independent. Now assume that the Markovian sequence $\pi_1, \pi_2, \ldots, 0 \leq \pi_k \leq 1$, is stopped at time m, which will be the case if observation ceases at m. Asgharian (2001), consider following model for the change - point distribution of τ

$$P(\tau = k) = \begin{cases} \pi_{k+1} \prod_{l=1}^{k} (1 - \pi_1) & if \quad k = 1, \ldots, m-1 \\ \prod_{l=1}^{m} (1 - \pi_1) & if \quad k = m \end{cases} \tag{4}$$

with $\pi_1 = 0$. Define $h(k) = p(\tau = k | \tau \geq k)$. It is important for our purposes to note that π_{k+1} is equal to the hazard, $h(k)$, $k = 1, \ldots, m-1$, and that $h(m)$, since

$$h(k) = \frac{P(\tau = k)}{P(\tau \geq k)} = \frac{\pi_{k+1} \prod_{l=1}^{k} (1 - \pi_1)}{\prod_{l=1}^{m} (1 - \pi_1)} = \pi_{k+1} \tag{5}$$

Modelling the change-point Distribution

For finding the form of the distribution of τ, by choosing some form for $p(\tau = k)$, the case $\pi_k = \pi$ for $k = 2, \ldots, m$, was considered by Asgharian (2001) which leads to a constant hazard function. For instance, in a clinical trial, it may be believed that the hazard for a change in response is increasing, or perhaps, constant. We consider the case where π_k is piecewise constant, i.e. a simple model $\pi_k = \begin{cases} \gamma & k = 1, 2, \ldots, m-1 \\ \eta & k = m \end{cases}$. Then the distribution of is

$$P(\tau = k) = \begin{cases} \gamma (1 - \gamma)^k & k = 1, 2, \ldots, m-1 \\ (1 - \eta)^m & k = m \end{cases} \tag{6}$$

with hazard function

$$h(k) = \begin{cases} \frac{\gamma(1-\gamma)^k}{2-\gamma-(1-\gamma)^{k-1}} & k = 1, 2, \ldots, m-1 \\ \frac{(1-\eta)^m}{1-(1-\eta)^{m-1}} & k = m \end{cases} \tag{7}$$

3. Introducing Covariates into the Model

In this section we introduce covariates in a change - point model. Up to now we model the hazard as a function of time. Consider the model (6), and impose a logistic form on $\pi(z_i)$. Therefore, we shall suppose that for subject i, with covariate vector $z_i^{'} = (1, z_{i1}, \ldots, z_{ir})$ and regression coefficient vector $\beta^{'} = (\beta_0, \ldots, \beta_r)$.

$$\gamma_k(z) = \frac{e^{\alpha_k + \beta^{'} z}}{1 + e^{\alpha_k + \beta^{'} z}},\ \eta_k(z) = \frac{e^{\delta_k + \lambda^{'} z}}{1 + e^{\alpha_k + \beta^{'} z}} \tag{8}$$

The components, Z_{ij}, of the vector z_i may be discrete or continuous. We also assume that the logistic model has been specified in such a way that the matrix M_z whose rows are the z_i has rank $r+1$. We write the full likelihood of the observed data, for the matrix of observations (3), $[X_{ik}]_{nm}$, where X_{ik} is the kth observation on the ith path, let $g_k^{z_i}(x_{i1}, \ldots, x_{im})$ be the density function of $X_i = (X_{i1}, \ldots, X_{im})$, given that the change takes place at k. For instance, if we assume that all the X_{ik} are independent and identically distributed before and after the change takes place, with density function $h_1^{z_i}$ and $h_2^{z_i}$, respectively, then conditional on a change of k,

$$g_k^{z_i}(x_{i1}, \ldots, x_{im}) = \prod_{l=1}^{k} h_k^{z_i}(x_{ie}) \prod_{l=k+1}^{m} h_k^{z_i}(x_{ie}) \qquad k = 1, \ldots, m-1$$

and

$$g_k^{z_i}(x_{i1}, \ldots, x_{im}) = \prod_{l=1}^{k} h_k^{z_i}(x_{ie}).$$

The unconditional density function for X_i is then

$$f_{z_i}(x_i) = \sum_{k=1}^{m} p_{z_i}(\tau_i = k)\, g_k^{z_i}(x_i) \tag{9}$$

where τ_i is the instant of change for the path. Thus the likelihood for a sample of n independent path is

$$\prod_{l=1}^{k} f_{z_i}(x_i) = \prod_{l=1}^{k} \left\{ \sum_{k=1}^{m} p_{z_i}(\tau_i = k)\, g_k^{z_i}(x_i) \right\}.$$

The expression (9) reduces to

$$f_{z_i}(x_i) = \sum_{k=1}^{m} p_{z_i}(\tau_i = k) \prod_{l=1}^{k} h_k^{z_i}(x_{ie}) \prod_{l=k+1}^{m} h_k^{z_i}(x_{ie})$$

under the assumption of independence of the observations within each path. Here the product $\prod_{l=k+1}^{m} h_2^{z_i}(x_{il})$ is defined to be equal to 1 for $k = m$. Now

$$\begin{aligned} f_{z_i}(x_i) &= \sum_{k=1}^{m-1} p_{z_i}(\tau_i = k) \prod_{l=k+1}^{k} h_2^{z_i}(x_{il}) \prod_{l=k+1}^{m} h_2^{z_i}(x_{il}) + p_{z_i}(\tau_i{=}m) \prod_{l=1}^{m} h_2^{z_i}(x_{i1}) \\ &= \sum_{k=1}^{m-1} \gamma^{1-I_m(k)}(z_i) \{1-\gamma(z_i)\}^k \prod_{l=1}^{k} h_1^{z_i}(x_{il}) \prod_{l=k+1}^{m} h_2^{z_i}(x_{ie}) \\ &+ (1-\eta(z_i))^m \prod_{l=1}^{m} h_1^{z_i}(x_{i1}) \\ &= \sum_{k=1}^{m-1} \left\{ \frac{e^{\alpha_k + \beta' z}}{1+e^{\alpha_k+\beta' z}} \right\}^{1-I_m(k)} \frac{\prod_{l=1}^{k} h_1^{z_i}(x_{il}) \prod_{l=k+1}^{m} h_2^{z_i}(x_{ie})}{\left\{1+e^{\alpha_k+\beta' z}\right\}^k} + \frac{\prod_{l=1}^{k} h_1^{z_i}(x_{il})}{\left\{1+e^{\alpha_k+\beta' z}\right\}^m} \end{aligned}$$

(10)

where $I_m(k) = \begin{cases} 1 & if \quad k = m \\ 0 & otherwise \end{cases}$, and

$\prod_{l=k+1}^{m} h_2^{z_i}(x_{il}) = 1$, for $k = m$.

4. Estimation of the Model Parameters and Quasi-identifiability of the Model

By finding the likelihood function, we can find the maximum likelihood estimators of the parameters, and then we determine the quasi-identifiability of the model.

Definition 5.1: A collection of families of probability measures $\{\{p_\theta^z; \theta \in \Theta\}, z \in Z\}$ is called quasi-identifiable with respect to θ if any $\theta \neq \theta^* \in \Theta$, there exists $z \in Z$, such that $P_\theta^z \neq P_{\theta^*}^z$.

Quasi-identifiability means, identifiability of the conditional density given the covariates. It follows from (10) that the likelihood induced by the matrix of observations (3) with independent rows is given by

$$L(\theta) = \prod_{i=11}^{n} f_{z_i}(x_i;\theta) = \prod_{i=11}^{n} \left\{ \sum_{k=1}^{m-1} \xi_k(\alpha,\beta;z_i)\, g_k^{z_i}(x_i;v) + \phi_m(\delta,\lambda;z_i)\, q_k^{z_i}(x_i;v) \right\}$$

(11)

where $\theta = (\alpha, \beta, v, \delta, \lambda)'$, $v = (v_1, v_2)'$, and

$$\begin{array}{lll} \xi_k(\alpha, \beta; z) & = \gamma(\alpha, \beta; z)\{1 - \gamma(\alpha, \beta; z)\}^k & k = 1, \ldots, m-1 \\ \phi_m(\delta, \lambda; z) & = (1 - \eta(\delta, \lambda; z))^m & \\ \gamma(\alpha, \beta; z) & = \frac{e^{\alpha_k + \beta' z}}{1 + e^{\alpha_k + \beta' z}}, & \eta(\delta, \lambda; z) = \frac{e^{\delta_k + \lambda' z}}{1 + e^{\alpha_k + \beta' z}} \end{array}$$

The vector $v = (v_1, v_2)'$ specifies the parameters of the distributions before and after the change, while β, λ, α_k, δ_k are the unknown regression parameters.

Theorem 5.1: Suppose that the rank of M_z is $r+1$ and that h_s^z for $s = 1, 2$ are respectively quasi - identifiable with respect to γ_s for $s = 1, 2$. Then

$$f_z(x; \theta) = \sum_{k=1}^{m-1} \xi_k(\alpha, \beta; z_i)\, g_k^{z_i}(x_i; v) + \phi_m(\delta, \lambda; z_i)\, \phi_k^{z_i}(x; v)$$

is quasi - identifiable with respect to θ.

Acknowledgement: The author wishes to thanks the council of research of Shiraz university for financial support of visiting department of Mathematics and Statistics of Mcgill university.

REFERENCES

[1] Asgharian, M., D. Wolfson Covariates in multipath change-point problems modelling and consistency of the MLE. *The Canadian Journal of Statistics* **29(4)** (2001), 515–528.

[2] Belisle, P., L. Joseph, B. Mac Gibbon, D. B. Wolfson, R. du Berger Change point analysis of neuronspike train data. *Biometrics* **54** (1998), 113–123.

[3] Chen, J., A. K. Gupta Parametric statistical change point analysis. Birkhauser, Boston, 2000.

[4] Cox, D. R., P. A. W. Lewis The statistical analysis of series of events. London: Methuen, 1966.

[5] Ghosh, J. K., P. K. Sen On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. *In proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* June. 20 - July 1, 1983 (L. M. Le Com & R. A. Olshen, eds). Wadsworth Statistics/probability series (1985), 789–806.

[6] Joseph, L. The multi-path change-point. Unpublished Ph.D. Thesis, Department of Mathematics and Statistics, McGill University, Montreal, (1989).

[7] Joseph, L., D. B. Wolfson Estimation in multi-path change-point problems. *Communications in Statistics: Theory and Methods* **21** (1992), 897–918.

[8] Joseph, L., D. B. Wolfson Maximum likelihood estimation in the multi-path change-point problem. *Annals of the Institute of Statistical Mathematics* **45** (1993), 511–530.

[9] Joseph, L., D. B. Wolfson, R. du Berger, R. M. Lyle Change - point analysis of a randomized trial on the effects of calcium supplementation on blood pressure. In Bayesian Biostatistics (D. A. Berry & D. K. Stangle, eds.), Marcel Dekker, New York, 1996, 617–649.

[10] Joseph, L., D. B. Wolfson, R. du Berger, R. M. Lyle Analysis of Panel data with (1997)

[11] Lyle, R. M., C. L. Melby, G. C. Hyner, J. W. Edmonson, J. Z. Miller, M. H. Weinberger Blood pressure and metabolic effects of calcium supplementation in normotensive white and black men. *Journal of the American Medical Association* **257** (1987), 1772–1776.

[12] Maguire, B. A., E. S. Pearson, A. H. A. Wynn The time intervals between industrial accidents. *Biometrika* **38** (1952), 168–80.

[13] Page, E. Continuous inspection Schemes. *Biometrika* **41** (1954), 100–114.

N. Sanjari Farsipour
Dept. of Statistics
Shiraz University
Shiraz 71454, IRAN
e-mail: nsf@susc.ac.ir