



## Fast estimation of regression parameters in a broken-stick model for longitudinal data

Ritabrata Das, Moulinath Banerjee, Bin Nan & Huiyong Zheng

**To cite this article:** Ritabrata Das, Moulinath Banerjee, Bin Nan & Huiyong Zheng (2015): Fast estimation of regression parameters in a broken-stick model for longitudinal data, Journal of the American Statistical Association, DOI: [10.1080/01621459.2015.1073154](https://doi.org/10.1080/01621459.2015.1073154)

**To link to this article:** <http://dx.doi.org/10.1080/01621459.2015.1073154>



View supplementary material [↗](#)



Accepted author version posted online: 07 Aug 2015.



Submit your article to this journal [↗](#)



Article views: 82



View related articles [↗](#)



View Crossmark data [↗](#)

# Fast estimation of regression parameters in a broken-stick model for longitudinal data

Ritabrata Das, Moulinath Banerjee, Bin Nan, Huiyong Zheng \*

## Abstract

Estimation of change-point locations in the broken-stick model has significant applications in modeling important biological phenomena. In this article we present a computationally economical likelihood-based approach for estimating change-point(s) efficiently in both cross-sectional and longitudinal settings. Our method, based on local smoothing in a shrinking neighborhood of each change-point, is shown via simulations to be computationally more viable than existing methods that rely on search procedures, with dramatic gains in the multiple change-point case. The proposed estimates are shown to have  $\sqrt{n}$ -consistency and asymptotic normality – in particular, they are asymptotically efficient in the cross-sectional setting – allowing us to provide meaningful statistical inference. As our primary and motivating (longitudinal) application, we study the Michigan Bone Health and Metabolism Study cohort data to describe patterns of change in log estradiol levels, before and after the final menstrual period, for which a two change-point broken stick model appears to be a good fit. We also illustrate our method on a plant growth data set in the cross-sectional setting.

**Keywords:** Asymptotic efficiency; Change-point; Hormone profile; Piecewise linear model.

---

\*Ritabrata Das is Ph.D. candidate, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: [ritob@umich.edu](mailto:ritob@umich.edu)). Moulinath Banerjee is Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: [moulib@umich.edu](mailto:moulib@umich.edu)). Bin Nan is Professor, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: [bnan@umich.edu](mailto:bnan@umich.edu)). Huiyong Zheng is Associate Research Scientist, Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109 (E-mail: [zhenghy@umich.edu](mailto:zhenghy@umich.edu)). The work of Banerjee was supported in part by NSF grant DMS-1308890. The work of Nan was supported in part by NIH grant R01-AG036802 and NSF grants DMS-1007590 and DMS-1407142. The authors gratefully acknowledge the editor, associate editor and the referees for their helpful comments and suggestions.

# 1 Introduction

In regression models, it is often assumed that the regression function throughout the domain of interest has the same parametric form. But it is also important to consider situations where the regression function has different functional forms in separate portions of the domain of interest. A special case of this is the continuous piecewise linear model, popularly referred to as the “broken-stick model”. This model is frequently useful in environmental and biological setups where the locations of the change-points are of interest. The broken-stick with  $r$  (known a priori) change-points can be canonically written as

$$E(Y|X, Z) = \beta_0 + \beta_1 X + \sum_{k=1}^r \beta_{k+1} f(X, \tau_k) + Z^T \lambda, \quad (1)$$

where

$$f(x, \tau) = (x - \tau)^+ = \begin{cases} x - \tau, & x > \tau; \\ 0, & x \leq \tau; \end{cases}$$

and  $\tau_j$ ’s are ordered change-points to be estimated together with other regression parameters  $\beta$ ’s and  $\lambda$ . Such a modelling strategy is of particular interest for the Michigan Bone Health and Metabolism Study (MBHMS).

MBHMS is a population-based longitudinal natural history study of ovarian aging conducted in a cohort of 664 White women from Tecumseh, Michigan during their young and mid-adulthood (24 – 44) years. Sowers et al. (2008) studied the serum estradiol (E2) hormone levels in 629 women enlisted in the MBHMS over a fifteen year period starting from 1992. The goal of Sowers et al. (2008) was to describe the E2 profile changes before and after Final Menstrual Period (FMP). A semiparametric mixed model approach was implemented in Sowers et al. (2008), and smoothing splines were used for estimation. Referring to Fig. 1 in Sowers et al. (2008), it is clear that the mean function can be fit nicely by a piecewise linear model with multiple change-points, whose identification is of considerable significance. However, existing methods of change-point estimation in a broken-stick model are fairly slow for large sample sizes. In the particular scenario of Sowers et al. (2008), the effective sample-size is of order of  $10^4$  and hence a fast method of estimating the change-points precisely is of considerable importance.

In the early literature, it was generally assumed that either the exact location of the change-point  $\tau$  is known (Poirer 1973), or, at worst, it is known which two observation points  $\tau$  should lie between (Robinson 1964). Also, most of the early work focused on detection of whether a change-point existed at all. In this article, however, the existence of change-point(s) is assumed and the exact number of change-points, denoted by  $r$ , is assumed known. The focus here is to propose a quick estimation procedure that gets around the non-smoothness of the model without compromising asymptotic efficiency in the process.

The principal difficulty in the estimation problem arises when the locations of the change-points are unknown. For an independent and identically distributed error case, if the location of these change-points were known, we would have a standard linear regression problem. Even if a relatively small set of plausible values were known, one could perform least squares for the slope and intercept parameters for each of these plausible values to find the overall least squares estimate. However, in most scenarios, this is unlikely, and the set of plausible values over which one needs to search is typically all  $r$  tuples of ordered  $X_i$ 's, leading to a very high number of linear regressions,  $\binom{n}{r}$  in principle,  $n$  being the sample size; see Hudson (1966).

Bellman and Roth (1969) proposed an alternative method based on dynamic linear programming but this method is even slower than the previous method. Feder (1975a) considered a general case of segmented regression problems and showed that the exact least squares estimate obtained by Hudson (1966) is asymptotically normal. In particular, for the broken-stick model, the estimate is  $\sqrt{n}$ -consistent.

Tishler and Zang (1981) were the first to suggest estimation of change-points using a maximum likelihood approach based on smoothing. They argued that the non-differentiability of the 'maximum' and 'minimum' operators in piecewise regression was the main problem in using maximum likelihood. However, if these operators were substituted by smoothed versions, maximum likelihood could readily be used for fast computation. Tishler and Zang (1981) suggested using a quadratic approximation, where the length of the interval on which  $f$  is smoothed was taken as an arbitrary small value. However, the behavior of these estimates, as the interval-length shrinks to 0, was not investigated. It

is clear that unless the length of the interval is allowed to decrease with sample size, their algorithm cannot yield a consistent estimate.

Recent articles for broken-stick models include Bhattachaya (1987), Huskova (1998) and Muggeo (2003). While the first two deal with theory, Muggeo (2003) tries to develop an estimation strategy, but does not provide any asymptotic results and thus fails to address the theoretical efficiency of the approach. For a detailed description of Bayesian methods of change-point estimation refer to Chen et al. (2011).

In sum, the lack of a suitable method of estimation, that is optimal in terms of both precision and computational economy, has forced statisticians to fall back on the search-based algorithm of Hudson or related algorithms thereof. Our paper fills this gap in the literature.

We should note here that alternative approaches to studying ‘kink-type’ phenomena have also been investigated. Chiu et al. (2006) suggested that in certain scenarios, instead of using the broken-stick as the true model, it might be better to use, what they referred to as, the “bent-cable model”, which is a quadratic smoothing of the broken-stick model in a  $\gamma$  neighborhood around  $\tau$ . Their change-point parameter was defined as the mid-point of the interval  $(\tau - \gamma, \tau + \gamma)$  on which the smoothing was done; here both  $\tau$  and  $\gamma$  are unknown parameters. It was shown that  $\hat{\tau}_n$ , the least squares estimate of  $\tau$ , is  $\sqrt{n}$ -consistent. Also, for  $\gamma_0 > 0$ ,  $\hat{\gamma}_n$  is  $\sqrt{n}$ -consistent as well. In a previous article, Chiu et al. (2002) had shown that for  $\gamma_0 = 0$ , i.e., when the smooth model reduces to the broken-stick model, the asymptotics are complex and  $\hat{\gamma}_n$  is at most  $n^{1/3}$ -consistent.

In this chapter, equation (1) is used as the *true* mean model. Ideally, one would want to minimize the residual sum of squares in this model by a Newton-Raphson type algorithm, but this is not viable owing to the non-differentiability of  $f$  at  $\tau$ . To this end, we use a twice differentiable perturbation of  $f$ , denoted by  $q_n$ , as our working model, where  $q_n$  coincides with  $f$  outside a shrinking neighborhood of  $\tau$ , say  $(\tau - \gamma_n, \tau + \gamma_n)$ , with  $\gamma_n$ , a user-specified tuning parameter, going to 0. Because  $q_n$  is differentiable, the minimization can be done by Newton’s algorithm very quickly. For the iid error case, we show that our estimate of  $\tau$  is indeed  $\sqrt{n}$ -consistent for  $\tau$ , and furthermore has an asymptotic normal distribution with the same asymptotic variance as the exact least squares estimate of Hudson (1966). For

the longitudinal model, the same method yields  $\sqrt{n}$ -consistent and asymptotically normal estimates for the change-points even for misspecified variance structures.

In sections 2 and 3, we introduce the model for both the cross-sectional and longitudinal set-ups respectively and outline the main steps of the estimation. The main theoretical results are presented along with the main ideas of the proofs. Section 4 contains simulation results indicating the efficiency of the proposed method while the method is applied to two real data —plant growth data (cross-sectional study) in section 5.1 and estradiol profile analysis (longitudinal study) in section 5.2. Sketch of the proofs are provided in the appendices while the details are presented in the Supplement.

## 2 Cross-sectional Study

We assume that the covariate  $X$  is contained in  $[M_1, M_2]$ . The regression parameter in the cross-sectional set-up (1), is denoted by  $\theta^T = (\beta^T, \tau^T, \lambda^T)$ . We assume  $\theta$  belongs to a compact set  $\Theta = \mathbb{B} \times [M_1 + \delta, M_2 - \delta]^r \times \Lambda$ ,  $\tau_k < \tau_{k+1}$ ,  $k = 1, 2, \dots, r-1$ , and  $|M_1|, |M_2| < \infty$  and  $\delta$  is a known small positive constant, indicating the change-points need to be well-separated from the boundaries of the  $X$ -space. Without loss of generality, take  $M_1 = 0$  and  $M_2 = M$ . We write  $\beta^T = (\beta_0, \beta_1, \dots, \beta_{r+1}) \in \mathbb{B}$ , where  $\beta_0$  is the intercept and  $\sum_{j=1}^k \beta_j$  is the slope of the  $k^{th}$  segment,  $k = 1, 2, \dots, r+1$ . We assume  $\mathbb{B}$  is a compact set in  $\mathbb{R}^{k+2}$  and  $\Lambda$  is a compact set in  $\mathbb{R}^l$ ; the restriction  $0 < \zeta \leq |\beta_k| \leq B < \infty$ ,  $2 \leq k \leq r+1$  is imposed for the sake of identifiability. We also write  $\tau^T = (\tau_1, \tau_2, \dots, \tau_r)$ , with  $\tau_k$  denoting the  $k$ 'th smallest change-point, and again identifiability requires the conditions  $\tau_k < \tau_{k+1}$ ,  $k = 1, 2, \dots, r-1$ ,  $\tau_1 \geq \delta$  and  $\tau_r \leq M - \delta$ . The errors  $\varepsilon_i$  are assumed to be independent and identically distributed with mean 0 and variance  $\sigma^2$ . The true parameter vector  $\theta^0 = (\beta_0^0, \beta_1^0, \dots, \beta_{r+1}^0, \tau_1^0, \tau_2^0, \dots, \tau_r^0, \lambda^0)^T$  is assumed to be an interior point of the compact set  $\Theta$ . Our data are independent and identically distributed observations  $\{Y_i, X_i, Z_i\}_{i=1}^n$  from (1) and henceforth,  $\mathbb{P}_n$  denotes the empirical measure of the data. Note that,  $Y_i$  and  $X_i$  are scalars while  $Z_i$  is an  $l$ -dimensional vector of covariates with no change-points.

## 2.1 Estimation

Define,

$$M(\theta, x, y, z) = (y - E_\theta(Y|X = x, Z = z))^2 = \left[ y - \left\{ \beta_0 + \beta_1 x + \sum_{k=1}^r \beta_{j+1} f(x, \tau_k) + z^T \lambda \right\} \right]^2.$$

The exact least squares procedure aims to obtain the minimizer of

$$\mathbb{P}_n(M(\theta, X, Y, Z)) = \frac{1}{n} \sum_{i=1}^n M(\theta, X_i, Y_i, Z_i).$$

As  $f(x, \tau)$  is not differentiable at  $\tau$ , one cannot obtain the minimizer of  $\mathbb{P}_n(M(\theta, X, Y, Z)) \equiv \mathbb{P}_n(M(\theta))$  (for notational convenience) by Newton's algorithm. So, we resort to minimize  $\mathbb{P}_n(M_n(\theta))$ , where

$$M_n(\theta) \equiv M_n(\theta, x, y, z) = [y - \{\beta_0 + \beta_1 x + \sum_{k=1}^r \beta_{j+1} q_n(x, \tau_k) + z^T \lambda\}]^2$$

is *our working model*, a smoothed approximation of  $M(\theta)$ . Basically each of the  $f$ 's in  $M(\theta)$  is replaced by its corresponding smoothed version  $q_n$  to obtain  $M_n(\theta)$ .

As far as the functional form of  $q_n$  is concerned, the motivation lies in the work of Tishler and Zang (1981) and Chiu et al. (2006). Tishler and Zang (1981) suggested using a quadratic approximation, where the length of the interval on which  $f$  is smoothed was taken as an arbitrary small value, while Chiu et al. (2006) considered the length of the corresponding interval as a parameter. We consider the same functional form for  $q_n$  as in these papers, but in our model, the length of the interval on which we smooth  $f$  is a user-specified tuning parameter shrinking to 0 with  $n$  at an appropriate rate as  $n \rightarrow \infty$ . More specifically,

$$q_n(x, \tau) = \begin{cases} 0, & \text{if } x < \tau - \gamma_n; \\ \frac{(x - \tau + \gamma_n)^2}{4\gamma_n}, & \text{if } \tau - \gamma_n \leq x \leq \tau + \gamma_n; \\ (x - \tau), & \text{if } x > \tau + \gamma_n; \end{cases} \quad (2)$$

where  $\gamma_n$  is a deterministic sequence, that approaches zero as  $n \rightarrow \infty$ .

Define  $\{\hat{\theta}_n\}$  as a sequence in  $\Theta$  that solves the surrogate empirical estimating equation  $\mathbb{U}_n(\theta) = \partial \mathbb{P}_n(M_n(\theta))/\partial \theta = 0$ , with probability increasing to 1. The existence of such a  $\{\hat{\theta}_n\}$  is shown in the Supplement. In our proposed method, we use Newton-Raphson algorithm to solve  $\mathbb{U}_n(\theta) = 0$ . It is not clear whether  $\hat{\theta}_n$  is uniquely defined but this is not an issue, since for our asymptotics, how we pick the zero is immaterial, meaning that the results hold true for any choice of zero. Our numerical results, however, suggest that generally there is a unique zero.

**Remark 1.** *The assumed compactness of the parameter space facilitates the mathematical arguments leading to consistency of our estimator. This has little bearing on the implementation of the Newton-Raphson algorithm in practice: in applications, the compact set for  $\beta$  i.e.  $\mathbb{B}$  is typically not known a priori, so by thinking of the set as being arbitrarily large, the iterates of the algorithm, whenever it converges, can be viewed as living in the set under consideration.*

## 2.2 Asymptotic Results

For model (1) we consider the following regularity conditions.

*Condition 1.1:* There are  $r$  distinct change-points  $\tau_1, \dots, \tau_r$  in model (1) for a fixed integer  $r \geq 1$ ;  $r$  is known.

*Condition 1.2:* Covariate  $X \in [0, M]$ ,  $M < \infty$ , follows a continuous distribution  $F_X$  such that  $\text{pr}(\tau_k < X \leq \tau_{k+1}) > 0$ ,  $k = 0, \dots, r$  with  $\tau_0 = 0$  and  $\tau_{r+1} = M$ . The joint distribution of  $Z$  is denoted by  $F_Z$ .

*Condition 1.3:* Changes of slope parameters satisfy  $0 < \zeta \leq |\beta_k| \leq B < \infty$ ,  $k = 2, 3, \dots, r+1$ , for some constants  $\zeta$  and  $B$ .

**Theorem 1.** *Under Conditions 1.1-1.3,  $\hat{\theta}_n$  is a consistent estimator for  $\theta^0$  for any deterministic sequence  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$ .*

The proof of Theorem 1 consists of two major steps. The first is to show that  $\theta^0$  is the unique solution of  $U(\theta) = \partial P(M(\theta))/\partial \theta = 0$ , and the second, to show  $\|\mathbb{U}_n - U\| = o_p(1)$ ; here  $Pf = \int f dP$ ,  $P$  being the probability measure that generates the data, and  $\|\cdot\|$  is the



supremum norm. For the case with single change-point and covariate  $Z$  absent, a sketch of the proof is presented in the Appendix 1, while the details are provided in the Supplement . The proof for the case with multiple change-points or with other covariates is an exercise involving extensive algebraic derivations following the same line.

**Theorem 2.** For  $\gamma_n = n^{-\alpha}$  with  $\alpha > 1/2$ , under Conditions 1.1-1.3, we have that  $n^{1/2}(\hat{\theta}_n - \theta^0)$  converges in distribution to a normal random variable with mean 0 and covariance matrix  $2\sigma^2\dot{U}_*^{-1}(\theta^0)$ . The  $kl$ -th element of matrix  $\dot{U}_*$  is  $(\dot{U}_*(\theta^0))_{kl} = 2P(H^T(\theta^0)H(\theta^0))$ , where

$$H(\theta) = (1, X, f(X, \tau_1), \dots, f(X, \tau_r), -\beta_2 \mathbf{1}(X > \tau_1), \dots, -\beta_{r+1} \mathbf{1}(X > \tau_r), Z)_{1 \times (2+2r+l)}.$$

The proof of Theorem 2 also consists of two major steps. To this end, let us define  $\theta_n$  as the the root of  $U_n(\theta) = \partial P(M_n(\theta))/\partial \theta$  for sufficiently large  $n$ , in  $\Theta$  closest to  $\theta^0$ .  $\theta_n$  is well-defined follows from arguments showing that  $\hat{\theta}_n$  is well-defined (in the Supplement). The first step is to show that  $\theta_n$  converges to  $\theta^0$  with a faster than  $\sqrt{n}$  rate, which in fact is  $\gamma_n$ , and the second to show the asymptotic normality of  $n^{1/2}(\hat{\theta}_n - \theta_n)$ . Both steps rely on Taylor series expansions. For notational simplicity, we provide a sketch of the proof for the case with single change-point and absence of  $Z$  in the Appendix 1. Please refer to the Supplement for the detailed proof. The case with multiple change-points and other covariates is again a straightforward extension.

The following Corollary shows that our proposed local smoothing method does not lose any efficiency. Its proof is provided in the Supplement.

**Corollary 1.** The asymptotic distribution of our estimate, as stated in Theorem 2, is exactly the same as that in Feder (1975b) for the exact least squares estimate in the broken stick model.

**Remark 2.** Please note that for the sake of notational convenience and to keep the results and proofs terse, only one variable with change-points have been included in model (1). However, the method will work equally well for a model consisting of multiple variables, with multiple change-points in each variable, and the results will be analogous.

### 3 Longitudinal study

The model for the longitudinal study with a broken-stick mean function is

$$E(Y_{ij}|X_{ij}, Z_{ij}) = \mu_{ij} = \beta_0 + \beta_1 X_{ij} + \sum_{k=1}^r \beta_{k+1} f(X_{ij}, \tau_k) + Z_{ij}^T \lambda, \quad (3)$$

where  $Y_{ij}$  is the response of the  $i^{th}$  subject at the  $j^{th}$  time-point ( $t_{ij}$ ) and  $X_{ij}$  denotes the corresponding covariate with  $r$  change-points, while  $Z_{ij}$  are  $l$  other covariates with simple linear effects on  $Y_{ij}$ ,  $j = 1, \dots, m_i$ ,  $i = 1, \dots, n$ .

For the regression parameters,  $\beta^T = (\beta_0, \beta_1, \dots, \beta_{r+1})$ , we have the same assumptions as in the cross-sectional model. We assume the effect sizes  $\lambda \in \Lambda$  (a compact set in  $\mathbb{R}^l$ ) and  $\tau$  is the vector of change-points, as before. Here,  $\theta^T = (\beta^T, \tau^T, \lambda^T)$  is our parameter of interest and  $\theta^0$  is the true value of  $\theta$ .

As far as the variance function is concerned, we postulate the following form:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= g(\eta, t_{ij}, t_{ik}), \\ \text{Cov}(Y_{ij}, Y_{lk}) &= 0, i \neq l; j, k = 1, \dots, m_i; i = 1, \dots, n, \end{aligned} \quad (4)$$

where  $\eta$  is the vector of covariance parameters. We assume that the observations across individuals are independent and the correlation between different observations of the same individual can depend on the time-points but not on the mean parameters,  $\theta$ .

$Y_i$  is used to denote the vector of  $m_i$  observations for the  $i$ -th individual,  $i = 1, \dots, n$ .  $Y = (Y_1, \dots, Y_n)$  is the vector of all responses. We use similar definitions for  $X_i$  and  $X$ . Let  $\Sigma^{(i)}$  denote the dispersion matrix of  $Y_i$  and  $\Sigma$  the dispersion matrix of  $Y$ . The true dispersion matrix is denoted by  $\Sigma_0$ , which can be written as  $\Sigma(\eta^0)$ ,  $\eta^0$  being the true value of  $\eta$ .

To establish the asymptotic results rigorously, the problem needs to be cast in a proper mathematical framework. We assume that, the number of repeated measures is denoted by the random variable  $D$  which takes values in the integer-space  $\{1, 2, \dots, L\}$  with probabilities  $p_1, p_2, \dots, p_L$  respectively. Note that this  $L$  is assumed fixed and known. Also we

have a triangular array of  $X$ -values,

$$\begin{array}{ccccccc} X_1^{(1)} & & & & & & \\ X_1^{(2)} & X_2^{(2)} & & & & & \\ X_1^{(3)} & X_2^{(3)} & X_3^{(3)} & & & & \\ \vdots & & & & \ddots & & \\ X_1^{(L)} & X_2^{(L)} & X_3^{(L)} & \dots & X_L^{(L)} & & \end{array}.$$

When  $D = d$ , the  $d$ -th row of this array is selected as the set of  $X$ -covariates. The same is true for the  $Z$ -covariates and measurement errors  $\{\varepsilon_{ij}\}$ . We assume  $\{\varepsilon_{ij}\}$ 's are uncorrelated with  $\{X_{ij}\}$ 's and  $\{Z_{ij}\}$ 's. Thus,

$$Y^{(D)} = \beta_0 + \beta_1 X^{(D)} + \sum_{k=1}^r \beta_{k+1} f(X^{(D)}, \tau_k) + Z^T \lambda + \varepsilon^{(D)} = \mu^{(D)} + \varepsilon^{(D)}.$$

Note that, here  $f : \mathbb{R}^{D+1} \rightarrow \mathbb{R}^D$  is defined as  $f(x, \tau) = ((x_1 - \tau)^+, \dots, (x_k - \tau)^+)^T$  for any  $x \in \mathbb{R}^D$  and  $\tau \in \mathbb{R}$ . This is the general definition of the  $f$  function and (1) is a special case of this for  $D = 1$ . The observation for each individual consists of  $(D, Y^{(D)}, X^{(D)}, Z^{(D)})$  and our data comprise of  $n$  iid copies of this array. As with most inference methods in longitudinal studies, we allow for ignorable dropouts (Rubin 1976).

### 3.1 Estimation

The estimation process is divided into three steps:

**Step 1 :** Assume working independence, i.e. take  $\Sigma^{(i)} = I, i = 1, \dots, n$ . As for the cross-sectional study, replace each of the  $f$ 's by their respective smoothed version  $q_n$ 's. Now, find the corresponding estimate  $\hat{\theta}_n^{(I)}$  as the solution to the estimating equation

$$\frac{\partial}{\partial \theta} \mathbb{P}_n[(Y^{(D)} - \mu_n^{(D)})^T (Y^{(D)} - \mu_n^{(D)})] = 0,$$

where  $\mu_n^{(D)}$  is the smoothed version of  $\mu^{(D)}$ .

**Step 2 :** Then use  $\hat{\theta}_n^{(I)}$  to estimate the nuisance parameter  $\eta$ . The specifics will depend on the nature of the covariance function  $g$  in (4).

**Step 3 :** Use  $\hat{\eta}_n$  obtained in step 2 to estimate  $\theta$ . So the final  $\hat{\theta}_n$  is the solution to the estimating equation

$$\frac{\partial}{\partial \theta} \mathbb{P}_n[(Y^{(D)} - \mu_n^{(D)})^T \hat{\Sigma}_n^{-1} (Y^{(D)} - \mu_n^{(D)})] = 0,$$

where  $\hat{\Sigma}_n^{-1}$  is the block-diagonal dispersion matrix based on  $\hat{\eta}_n$ .

### 3.2 Asymptotic Results

As in the cross-sectional model, for the longitudinal model we consider similar regularity conditions.

*Condition 2.1:* There are  $r$  distinct change-points  $\tau_1, \dots, \tau_r$  in model (3) for a fixed integer  $r \geq 1$ ;  $r$  is known.

*Condition 2.2:* Conditional on  $D = d$ , covariate  $X \in [0, M]^d$ ,  $M < \infty$ , follows a continuous distribution  $F_X$  such that  $\text{pr}(\tau_k < X_j \leq \tau_{k+1}) > 0$ , for some  $j = 1, \dots, d$ , for all  $k = 0, \dots, r$  with  $\tau_0 = 0$  and  $\tau_{r+1} = M$ . Also we assume the covariates  $Z$  follow a joint distribution  $F_Z$ .

*Condition 2.3:* Changes of slope parameters satisfy  $0 < \zeta \leq |\beta_k| \leq B < \infty$ ,  $k = 2, 3, \dots, r+1$ , for some constants  $\zeta$  and  $B$ .

*Condition 2.4:* There exists a positive definite matrix  $W$ , such that estimated covariance matrix  $\hat{\Sigma}_n$  satisfies  $\sqrt{n}(\hat{\Sigma}_n - W) = O_p(1)$ .

**Theorem 3.** Under Conditions 2.1-2.4,

(a) The estimator  $\hat{\theta}_n$  is consistent for  $\theta^0$  given any deterministic sequence  $\gamma_n \rightarrow 0$  as  $n \rightarrow \infty$ .

(b) For  $\gamma_n = n^{-\alpha}$  with  $\alpha > 1/2$ ,  $n^{1/2}(\hat{\theta}_n - \theta^0)$  converges in distribution to a normal random variable with mean 0 and covariance matrix

$$K^{(W^{-1})} = 2 \sum_{d=1}^L P^{(d)} [(H^T(\theta^0) W^{-1} H(\theta^0))^{-1} (W^{-1} H(\theta^0))^T \Sigma_0 (W^{-1} H(\theta^0)) (H^T(\theta^0) W^{-1} H(\theta^0))^{-1}] p_d$$

where

$$H(\theta) = (1, X, f(X, \tau_1), \dots, f(X, \tau_r), -\beta_2 \mathbf{1}(X > \tau_1), \dots, -\beta_{r+1} \mathbf{1}(X > \tau_r), Z)_{d \times (l+2r+2)};$$

here  $P^{(d)}f = \int f dP^{(d)}$ ,  $P^{(d)}$  being the probability measure that generates the data given  $D = d$ .

**Remark 3.** If the matrix  $W$  in condition 2.4 is indeed the true covariance matrix  $\Sigma_0$ , i.e.,  $\hat{\Sigma}_n$  is a  $\sqrt{n}$ -consistent estimate of  $\Sigma_0$ , then for  $\gamma_n = n^{-\alpha}$  with  $\alpha > 1/2$ ,  $n^{1/2}(\hat{\theta}_n - \theta^0)$  converges in distribution to a normal random variable with mean 0 and covariance matrix

$$K^{(\Sigma_0^{-1})} = 2 \sum_{d=1}^L P^{(d)} [(H^T(\theta^0) \Sigma_0^{-1} H(\theta^0))^{-1}] p_d.$$

The proof of Theorem 3 is similar to the proofs of Theorems 1 and 2 in Section 2. The main proof is divided into three major parts. First, we show that  $\sqrt{n}(\hat{\theta}_n^{(I)} - \theta^0)$  converges to  $N(0, K^{(I)})$  in distribution. Next, we prove  $\sqrt{n}(\hat{\theta}_n^{(W^{-1})} - \theta^0)$  converges to  $N(0, K^{(W^{-1})})$  in distribution. Here,  $\hat{\theta}_n^{(W^{-1})}$  is defined as a zero of  $\mathbb{U}_n^{(W^{-1})}(\theta) = \frac{\partial}{\partial \theta} \mathbb{P}_n(Y - \mu_n)^T W^{-1} (Y - \mu_n)$ , with probability increasing to 1. Finally, we show that  $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}) = o_p(1)$ , which proves Theorem 3.

For the sake of notational convenience, the proof is presented in the Appendix 2, for  $r = 1$  and for fixed visit-times, i.e.  $D \equiv m$  or equivalently  $m_i = m$  for all  $i = 1, \dots, n$ . Also for brevity, we exclude the time-invariant covariates  $Z$  in the proof.

Though the proof provided in Appendix 2 is for a fixed number of visit-times, it holds true for variable number of visit-times, as stated in Theorem 3. Notice that, conditional on  $D = d$  (this event has probability  $p_d$ ), it is shown that  $n^{1/2}(\hat{\theta}_n - \theta^0)$  converges in distribution to a normal random variable with mean 0 and covariance matrix

$$\begin{aligned} K^{(W^{-1})} &= 2P^{(d)} [(H^T(\theta^0) W^{-1} H(\theta^0))^{-1} (W^{-1} H(\theta^0))^T \Sigma_0 (W^{-1} H(\theta^0)) \\ &\quad (H^T(\theta^0) W^{-1} H(\theta^0))^{-1}]. \end{aligned}$$

Now, the result of Theorem 3 easily follows.

**Corollary 2.** Denoting the mean function at  $X = x, Z = z$  by  $\mu(x, z, \theta)$ , we have  $\sqrt{n}(\mu(x, z, \hat{\theta}_n) - \mu(x, z, \theta^0))$  converges in distribution to a normal random variable with mean zero and variance  $a^T K^{(W^{-1})} a$ , where  $a^T = (1, x, f(x, \tau_1^0), \dots, f(x, \tau_r^0), -\beta_2^0 \mathbf{1}(x > \tau_1^0), \dots, -\beta_{r+1}^0 \mathbf{1}(x > \tau_r^0), z)$ .

This result is useful in providing pointwise confidence bands for the broken-stick mean function as illustrated in Fig 4. The proof is provided in the Supplement.

**Remark 4.** Note that the estimated confidence band for the mean at  $\hat{\tau}_n$ , as provided by Corollary 2 is discontinuous. The asymptotic distribution of  $\sqrt{n}(\mu(\hat{\tau}_n, z, \hat{\theta}_n) - \mu(\tau^0, z, \theta^0))$  is not a direct application of this result, but needs separate calculations—a direct application of the delta method.

## 4 Simulations

### 4.1 Cross-sectional set-up

Simulations were conducted to compare the proposed method with the existing one. Models with one and two change-points were both considered. Sample sizes were varied,  $n = 50, 200, 1000, 5000$ . For each of the two models, for a fixed  $n$ , 3 different sets of values of  $\theta$  were considered within the domain of interest. For each value of  $\theta$ , the proposed and existing (Hudson 1966) methods were both repeated  $N = 1000$  times. The run-times, a measure of computational efficiency, for each of the methods were then averaged over these 1000 repetitions and over the 3 different values of  $\theta$ . This was done to average out discrepancies being caused by individual  $\theta$ 's. Error standard deviation  $\sigma$  was taken to be equal to 0.1 for all cases and  $M = 1$ . For all simulations,  $\alpha$  was taken to be 1. The simulations were carried out on an Intel(R) Core(TM) i7 system with 1.6 GHz and 8 GB RAM in a 64-bit OS.

#### 4.1.1 Results

From Table 1, it is obvious that the proposed method is much faster than the exact least squares method, especially for two change-point problems. Also Tables 2 & 3 indicate that

the change-point estimates of the proposed method are almost as accurate as the exact least squares estimate. The biases are close to zero for all sample sizes, especially for the large samples. The standard deviation estimates are very close to the sample standard deviations indicating our standard deviation estimates work well, especially for large samples. The estimates are also very close to the theoretical standard deviations, indicating the asymptotic efficiency of our estimates. Although the bias and variances for the  $\beta$ 's have not been tabulated for the sake of brevity, we observed that our  $\beta$  estimates also have comparable Mean Squared Errors to their respective exact least squares estimates.

#### 4.1.2 Choice of $\alpha$ for finite samples

Although asymptotic results were established for all  $\alpha > 1/2$ , what a proper choice of  $\alpha$  should be for finite samples is a very pertinent question. We performed extensive simulations for different sample-sizes, to explore the robustness of different choices of  $\alpha$  values.

We tried a sample situation with one change-point,  $\beta_0^0 = 0.3$ ,  $\beta_1^0 = 1.5$ ,  $\beta_2^0 = 1$  and  $\sigma = 0.1$  with covariate  $X$ -space  $= [0, 1]$ . The  $\tau$ -values were varied between 0 and 1 and the Mean Square Errors were plotted against  $\log_{10} \alpha$  values for various sample-sizes. We found that the M.S.E. vs  $\log_{10} \alpha$  graphs are almost invariant with changing sample-sizes. To change the signal-to-noise ratio, the  $\beta$ -values were kept constant but  $\sigma$  was changed to 0.5 (Fig.2) and 1. The patterns are exactly similar for all parameter values and signal-to-noise ratios. However, for a small  $n$  and a very large value of  $\alpha$ , the algorithm occasionally breaks down because  $\dot{\mathbb{U}}_n(\theta)$  becomes (almost) singular for computational purposes. This is clearly indicated by the very large average MSE for  $\alpha = 50$  or 100 when sample-size is small ( $n = 50$ ). So, very large  $\alpha$ 's (greater than 10) are not recommended for small samples (less than 50). We would also like to point out the robustness of the M.S.E.'s to the choice of tuning parameter in the range of  $\alpha$ 's for which the algorithm is numerically stable. This is reflected by the flat stretch of the M.S.E. curves for each  $n$ , before numerical instability sets in. In other words, so long as the algorithm works, any choice of  $\alpha > 1/2$  is essentially as effective as any other. So, searching for an optimal  $\alpha$  is unlikely to yield any significant gains. Our recommendation is to use  $\alpha = 1$ , which works very well in terms of M.S.E. for all sample-sizes, as low as 30. The same  $\alpha$  value (1) is used for all data analyses in the

subsequent sections. The simulations indicate that computational efficiency is insensitive to the choice of  $\alpha$ . A more detailed version of Fig.2 is provided in the Supplement.

## 4.2 Longitudinal set-up

Simulations were conducted for the longitudinal case as well to compare the efficiency of our proposed method to the search-based algorithm. We considered an AR(1) correlation structure with  $\rho = 0.6$  to model the dependence among observations within subject. For each subject, we considered 10 observations in scenarios G and H (Table 4). For set-up J, we considered varying number of observations for each individual, which is uniformly distributed over integer-space  $\{1, 2, \dots, 20\}$ . Error standard deviation  $\sigma$  was taken to be equal to 0.1 for all cases and  $M = 1$ . For all simulations,  $\alpha$  was taken to be 1. The computational efficiency of our proposed method is huge compared to the search-based algorithm, as in the cross-sectional case (Table 1). So, in Table 4, we have just compared the bias and standard errors to illustrate the validity and estimation efficiency of our method. Results from Table 4, clearly indicate that our method yields almost the same standard error estimates as the search-based algorithm. Although for both methods with small sample-sizes, the bias is comparatively high and the standard deviation estimates are higher than the theoretical values, the differences become smaller for larger sample sizes. The M.S.E.'s for the slope and intercept parameters also behave similar to those of the change-points.

## 5 Applications

### 5.1 Plant growth data analysis

Vernalization, a requirement for plants to experience a period of cool conditions to accelerate flowering, is an important determinant of flowering date in winter wheat. In Brooking and Jameison (2002), controlled environment studies were carried out to quantify the response of vernalization rate to temperature for two near-isogenic lines of the wheat cultivar Batten: Spring Batten, vernalization insensitive; and Winter Batten, vernalization sensitive. Plants were sampled for dissection at intervals during the treatment and post-treatment pe-



riod, until the flag leaf could be distinguished. The authors investigated the co-ordination of primordium initiation and leaf appearance, quantified by the Haun stage. The authors observed that Spring Batten plants grown under fully inductive conditions, 25/20°C, 16 hrs photoperiod, produced eight leaves on average, and the rate of primordium initiation per emerged leaf increased markedly with the transition from leaf initiation to spikelet initiation. This represents an important phase transition in the growth of the plant. From Fig. 3, it is quite clear that the model which best fits the scenario is a broken-stick model with two change-points. The authors had estimated the change-points by naked eye and then fitted three line-segments for the three regions. We provide a fast as well as statistically rigorous analysis using the approach developed in this paper.

The change-point estimates of Brooking and Jameison (2002) by naked eye were 2.6 and 5 on the Haun stage scale, whereas ours are 2.931(2.715, 3.147) and 4.764(4.647, 4.881), with 95% confidence intervals provided in parentheses. From Fig. 3 we see that the estimates in Brooking and Jameison (2002) do not lie within our confidence intervals, emphasizing the importance of a principled analysis such as the one we have proposed. The main conclusion in Brooking and Jameison (2002) was that the rate of primordium initiation per emerged leaf, the slope parameter, jumped from 1.9 primordia per leaf to 7.11 primordia per leaf and then became constant. Our estimates of the slopes of the three segments are 2.67(2.46, 2.88), 8.19(7.84, 8.54) and  $-0.02(-0.16, 0.12)$  primordia per leaf. Our estimates, qualitatively, corroborate their conclusion that there are two sharp phase transitions in the growth pattern whereby the initial growth rate gets more than tripled and then becomes more or less constant.

## 5.2 Estradiol hormone profile analysis

We applied our proposed method to analyze the longitudinal estradiol data as discussed in Section 1. For our purpose, we considered only women whose Final Menstrual Period (FMP) had already been observed. This was done so as to avoid scenarios with censored FMP's (Lu et al. 2010). Among all these women, eight were left out either because their observed FMP was too early or too late or had less than three data-points. The remaining

sample of  $n = 156$  women with identified FMP was our sample of interest who in total gave 1396 observations, with each woman contributing 3 to 10 observations over time, covering 11 years before to 10 years after FMP. This gave an average of about 8.95 observations per woman. There were 75(48%) smokers at baseline and the baseline BMI mean(SD) was 27.4(6.56). Please note that the data we use here have longer follow-up and hence more subjects with identified FMP, compared to the data on which the analysis in Fig. 1 in Sowers et al. (2008) is based. A log transformation was applied to the Estradiol hormone level to make the normality assumption more plausible.

Denote by  $Y_{ij}$  the  $j$ th log-transformed E2 value measured at day  $t_{ij}$  centered around FMP  $T_i$ , for the  $i$ th woman and by  $SMOKE_i$  and  $BMI_i$  baseline smoking habit (0 meaning smoker at baseline, 1 otherwise) and the baseline body mass index, centered at the grand mean, respectively. We consider the following model:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 f(X_{ij}, \tau_1) + \beta_3 f(X_{ij}, \tau_2) + \lambda_1 SMOKE_i + \lambda_2 BMI_i + b_i + U_i(t_{ij}) + \varepsilon_{ij} \quad (5)$$

where  $X_{ij} = t_{ij} - T_i$ , the  $b_i$  are random intercepts following a  $N(0, \phi)$  distribution, the  $U_i(t)$  are mean zero Gaussian processes modeling serial correlation and  $\varepsilon_{ij}$  are independent measurement errors following a  $N(0, \sigma^2)$  distribution. We assume  $U_i(t)$  is a nonhomogenous OrnsteinUhlenbeck process, which satisfies  $\text{Var}(U_i(t)) = \xi(t)$  where  $\log \xi(t) = \xi_0 + \xi_1 t + \xi_2 t^2$  and  $\text{corr}(U_i(t), U_i(s)) = \rho^{|t-s|}$ . We also assume that for each  $i$ ,  $\varepsilon_i$ ,  $b_i$  and  $U_i(t)$  are independent of one another. Further, we assume,  $-11.9 \leq \tau_1 < \tau_2 \leq 9.9$  (in general, we assume for all our theoretical results that the covariate  $X$  is contained in some compact interval  $[M_1, M_2]$ ; here from the nature of the study and previous work we knew the scope of the study was between 12 years before and 10 years after the FMP). Further to make (5) identifiable, and  $\Theta$  compact, we assume that  $10^{-6} \leq |\beta_2| \leq 10^6$  and  $10^{-6} \leq |\beta_3| \leq 10^6$ . Also, the variance function part does not include any mean function parameters and so even in the presence of unknown change-points, the model remains identifiable.

As illustrated in section 3, we estimate the regression parameters in a three step procedure. In the first step, we assume working independence to estimate  $\hat{\theta}_n^{(I)}$ . Then  $\eta =$

$(\phi, \sigma^2, \xi_0, \xi_1, \xi_2, \rho)$  is estimated by maximizing the conditional log-likelihood,

$$l(\eta) = -1/2 \sum_{i=1}^n \left[ (Y_i - \mu_{n,i}^{(I)})^T \Sigma(\eta) (Y_i - \mu_{n,i}^{(I)}) + \log |\Sigma^{(i)}(\eta)| \right]. \quad (6)$$

Therefore,  $\hat{\Sigma}_n = \Sigma(\hat{\eta}_n)$  which is subsequently used in Step 3 to obtain  $\hat{\theta}_n$ . Condition 2.4 is verified to hold for this model; in fact  $W$  here turns out to be  $\Sigma_0 = \Sigma(\eta^0)$ . The proof for this is provided in the Supplement.

Our results indicate the presence of change-points at  $-2.174$  ( $-2.554, -1.794$ ) and  $1.733$  ( $1.513, 1.953$ ) years (Table 5). In Sowers et al. (2008), the change-points had been roughly thought to be around 2 years before and after FMP. Although this was a good estimate, we can see that actually the 95% confidence interval for the second change point does not contain 2 years after FMP, indicating the change of estradiol levels actually happen slightly sooner than anticipated in Sowers et al. (2008). Also BMI and smoking habits do not seem to alter this pattern significantly. But, our contribution, above all, is providing statistically meaningful inference about the change-points. Also the form of the confidence bands indicate that a two-change point model is indeed a good fit for the E2-hormone profile.

## 6 Discussion

We have proposed a method of estimating change-points in a broken stick model which is computationally much more efficient than existing methods, and demonstrated that it is asymptotically as efficient. The method of estimation is also numerically stable. An added advantage of this method is that, as shown in section 3, it can be readily extended to generalized linear models with repeated measures, examples of which are abundant. The estimates in those frameworks have shown similar desirable asymptotic properties.

It seems reasonable to assume that this idea should work equally efficiently in estimating change-points in a time-series framework, at least under short-range dependence. For instance, estimation of change-points is of considerable significance in climatic series data (Lund et al. 1995; Lund and Reeves 2002) and such data sets tend to be really large. Hence

our idea would likely prove even more economic in this setting. This is underscored by the fact that even for a sample size of 50, our method is more than a hundred times faster than the exact least squares method with multiple change points, and at large samples, thousands of times faster. Also, in a linear spline model with knot-locations unknown (number of knots known), the proposed method provides a faster alternative for locating these knots.

We cannot stress enough that this is a very generic idea which can be used for computational economy in several settings without giving up on asymptotic efficiency. For example, the same idea should be applicable for estimating change-points in a multivariable setup where the change-points are observed in more than one variable. While, for a search based algorithm, the computational time will increase many folds with the number of variables having change-points, it will scale much more favorably for our approach.

However, we would like to point out if the investigators feel that the linearity of broken-stick model is not best suited for their data, our method of estimation or for that matter any method of estimation based on the broken-stick model may not be reliable.

Also if the coefficients of two consecutive regimes are very close, then trying to fit separate segments for the two regimes is strongly discouraged. We performed extensive simulations and both our approach and the search-based approach yield poor estimates. Thus before fitting a broken-stick model, we would strongly suggest the investigators check that the assumptions for the model are valid.

In this article, we were interested in modeling the mean hormone profile of all subjects in the cohort discussed in Section (5.2). A possible way to model individual-specific hormone profiles is via multi-path change-points models. Major work done in regards to multi-path change-points include Joseph and Wolfson (1993) and Asgharian and Wolfson (2001). Most of this literature has treated change-point as the observation at which a transition has occurred, rather than a point in the  $X$ -space. Broken-stick models with random change-points and random intercept-slopes is a possible interesting avenue for future work in this field. The simplest possible model with one change-point is:

$$E(Y|X) = \beta_0 + \beta_1 X + \beta_2 (X - \tau)^+,$$

where  $\theta^T = (\beta_0, \beta_1, \beta_2, \tau)$  follows, say, a multivariate  $N(\theta^0, \Upsilon)$  distribution. Estimating methods will rely on minimizing criterion functions involving several integrals and is beyond the scope of this work.

## Appendix 1: Sketch proofs of Asymptotic Properties for Cross-sectional Model

We consider the simplest case with a single change-point. The case with multiple change-points is an algebraic exercise. The proofs heavily rely on the results in van der Vaart and Wellner (1996). We provide sketch proofs here with details given in the online Supplement.

### Sketch proof of Theorem 1

It is clearly seen that  $U(\theta^0) = 0$ , here  $U(\theta) = \partial P(M(\theta))/\partial \theta$  is the population score function. The proof of consistency is based on the following two facts:

FACT 1:  $\theta^0$  is the unique solution of  $U(\theta) = 0$ .

FACT 2:  $\|\mathbb{U}_n - U\| := \sup_{\theta \in \Theta} |\mathbb{U}_n(\theta) - U(\theta)| = o_p(1)$ .

In the online Supplementary Material, we show that there exists a zero of  $\mathbb{U}_n(\theta)$ , denoted as  $\hat{\theta}_n$ , in  $\Theta$  on a set with probability increasing to 1. Denote this set as  $\Omega_n$ . On the shrinking set  $\Omega_n^c$ ,  $\hat{\theta}_n$  is assigned some pre-fixed value (say,  $\theta_1$ ) which is an interior point of  $\Theta$ . To prove consistency of  $\hat{\theta}_n$ , observe that, for any  $\epsilon > 0$

$$pr(|\hat{\theta}_n - \theta^0| > \epsilon) = pr(\{|\hat{\theta}_n - \theta^0| > \epsilon\} \cap \Omega_n) + pr(\{|\hat{\theta}_n - \theta^0| > \epsilon\} \cap \Omega_n^c).$$

Notice that on the set  $\Omega_n$ ,  $\hat{\theta}_n$  is a zero of  $\mathbb{U}_n(\theta)$ , hence also a minimizer of  $\|\mathbb{U}_n(\theta)\|_2$  in  $\Theta$  while  $\theta^0$  is the unique zero of  $U(\theta)$  and hence the unique minimizer of  $\|U(\theta)\|_2$ , where  $\|\cdot\|_2$  is the Euclidean norm. From Fact 2 and using triangular inequality, we have  $\sup_{\theta \in \Theta} \|\mathbb{U}_n\|_2 - \|U\|_2 \leq \sup_{\theta \in \Theta} \|\mathbb{U}_n - U\|_2 \leq K\|\mathbb{U}_n - U\| = o_p(1)$  for a finite constant  $K$ . So, from the argmax (argmin) continuous mapping theorem (van der Vaart and Wellner

(1996), Corollary 3.2.3) for any  $\epsilon > 0$ , there exists a  $\nu > 0$  such that,

$$pr(\{|\hat{\theta}_n - \theta^0| > \epsilon\} \cap \Omega_n) \leq \frac{\nu}{2}.$$

Also, for sufficiently large  $n$ ,  $pr(\Omega_n^c) \leq \frac{\nu}{2}$ . This implies that  $pr(|\hat{\theta}_n - \theta^0| > \epsilon) \leq \nu$ , for sufficiently large  $n$ , hence proving Theorem 1.

Now, proving Fact 1 is a tedious algebraic exercise. The idea is to, by method of elimination, bring down all the equations in the various parameters to one equation involving only  $\tau$ . Then we show that this function is negative for all  $\tau > \tau^0$  and positive for all  $\tau < \tau^0$ . This allows us to argue Fact 1.

To establish Fact 2, observe that  $\|\mathbb{U}_n - U\| \leq \|\mathbb{U}_n - U_n\| + \|U_n - U\|$ . Direct calculation yields  $\|U_n - U\| = O(\gamma_n) = o(1)$ . On the other hand  $(\mathbb{U}_n - U_n)(\theta) = (\mathbb{P}_n - P)(\partial M_n(\theta)/\partial\theta)$ .

Observe that

$$\partial M_n(\theta)/\partial\theta = \begin{pmatrix} -2(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau)) \\ -2X(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau)) \\ -2q_n(X, \tau)(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau)) \\ -2\beta_2 \frac{\partial q_n}{\partial \tau}(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau)) \end{pmatrix}.$$

Now,  $\Theta$  being compact, it is clear that  $\beta_0, \beta_1 X, \beta_2 q_n(X, \tau)$  and  $\beta_2 \frac{\partial q_n}{\partial \tau}$  are all bounded monotones as functions of  $\theta$ . Theorem 2.7.5 in van der Vaart and Wellner (1996) shows that bounded monotone functions have a bracketing number of order  $(1/\epsilon)$ , wrt  $L_1(P)$  norm and hence a similar bound on the covering number with respect to the same norm. So they have bounded uniform entropy integral (BUEI) property and hence belongs to a Glivenko-Cantelli class. Theorem 3 in van der Vaart and Wellner (1999) shows that simple operations such as addition or multiplication preserves the Glivenko-Cantelli property and hence each component function of  $\partial M_n(\theta)/\partial\theta$  belongs to a Glivenko-Cantelli class. Thus  $\partial M_n(\theta)/\partial\theta$  belongs to a Glivenko-Cantelli class and hence  $\|\mathbb{U}_n - U_n\| = o_p(1)$ , which establishes Fact 2, and hence Theorem 1.

## Sketch proof of Theorem 2

Since  $n^{1/2}(\hat{\theta}_n - \theta^0) = n^{1/2}(\hat{\theta}_n - \theta_n) + n^{1/2}(\theta_n - \theta^0)$ , the asymptotic distribution of  $\hat{\theta}_n$  is a direct result of the following two facts when  $\gamma_n = n^{-\alpha}$  with  $\alpha > 1/2$ .

FACT 3:  $\|\theta_n - \theta^0\| = O(\gamma_n)$ .

FACT 4:  $n^{1/2}(\hat{\theta}_n - \theta_n)$  converges in distribution to  $N(\mathbf{0}, 2\sigma^2 \dot{U}_*^{-1}(\theta^0))$ .

We first show Fact 3. Observe that a simple Taylor series expansion of  $U_n(\theta_n)$  around  $\theta^0$  yields

$$U_n(\theta_n) - U_n(\theta^0) = \begin{pmatrix} \dot{U}_{1n}(\tilde{\theta}_n^{(1)}) \\ \dot{U}_{2n}(\tilde{\theta}_n^{(2)}) \\ \dot{U}_{3n}(\tilde{\theta}_n^{(3)}) \\ \dot{U}_{4n}(\tilde{\theta}_n^{(4)}) \end{pmatrix} (\theta_n - \theta^0) = A_n(\theta_n - \theta^0),$$

where each of  $\tilde{\theta}_n^{(i)}$ ,  $i = 1, 2, 3, 4$ , lies on the straight line joining  $\theta_n$  and  $\theta^0$  and  $\dot{U}_{in} = \frac{\partial U_n}{\partial \beta_{i-1}}$ ,  $i = 1, 2, 3$ , and  $\dot{U}_{4n} = \frac{\partial U_n}{\partial \tau}$ . Now, we know from the proof of Fact 1,  $\sup_{\theta \in \Theta} |P(M_n(\theta)) - P(M_n(\theta^0))| = o(1)$ , hence  $\theta_n - \theta^0 = o(1)$ . This in turn implies that for sufficiently large  $n$ ,  $\theta_n$  is an interior point of  $\Theta$  and hence, a zero of  $U_n(\theta_n)$ .

Now,  $U(\theta^0) = 0$ . Thus for sufficiently large  $n$ , the above equality becomes

$$U_n(\theta^0) - U(\theta^0) = -A_n(\theta_n - \theta^0).$$

It is clearly seen from

$$U_n(\theta) - U(\theta) = \begin{pmatrix} -2\beta_2 P(f(X, \tau) - q_n(X, \tau)) \\ -2\beta_2 P[X(f(X, \tau) - q_n(X, \tau))] \\ 2P[(f(X, \tau) - q_n(X, \tau))(Y - \beta_0 - \beta_1 X - \beta_2(q_n(X, \tau) + f(X, \tau)))] \\ 2\beta_2 P[(-1(X \geq \tau) - \frac{\partial}{\partial \tau} q_n(X, \tau))(Y - \beta_0 - \beta_1 X - \beta_2(q_n(X, \tau) + f(X, \tau)))] \end{pmatrix},$$

that  $\|U_n - U\| = O(\gamma_n)$ . Thus Fact 3 is established if  $A_n$  is invertible for large  $n$ . This is established by showing that  $\dot{U}_n(\theta)$  converges uniformly, implying  $A_n$  converges as well.

Let  $\dot{U}_*(\theta^0)$  denote the limit of  $A_n$ . It can be shown

$$\dot{U}_*(\theta^0) = 2 \begin{pmatrix} 1 & P(X) & P(f(X, \tau^0)) & -\beta_2^0 P(\mathbf{1}(X > \tau^0)) \\ P(X) & P(X^2) & P(Xf(X, \tau^0)) & -\beta_2^0 P(X\mathbf{1}(X > \tau^0)) \\ P(f(X, \tau^0)) & P(Xf(X, \tau^0)) & P(f^2(X, \tau^0)) & -\beta_2^0 P(f(X, \tau^0)) \\ -\beta_2^0 P(\mathbf{1}(X > \tau^0)) & -\beta_2^0 P(X\mathbf{1}(X > \tau^0)) & -\beta_2^0 P(f(X, \tau^0)) & (\beta_2^0)^2 P(\mathbf{1}(X > \tau^0)) \end{pmatrix}.$$

Now, for any vector  $a = (a_1, \dots, a_4)^\top \neq 0$ , we have

$$a^\top \dot{U}_*(\theta^0) a = 2P\{a_1 + a_2 X + a_3 f(X, \tau^0) - a_4 \beta_2^0 \mathbf{1}(X > \tau^0)\}^2 > 0,$$

which implies that  $\dot{U}_*(\theta^0)$  is positive definite and hence nonsingular. Thus  $A_n$  is nonsingular for large enough  $n$ , and we have

$$|\theta_n - \theta^0| = A_n^{-1} \|U_n - U\| = O_p(\gamma_n).$$

We next show Fact 4. Denote  $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$ . Again, observe that by Taylor Series expansion of  $U_n(\theta_n)$  around  $\hat{\theta}_n$  yields

$$U_n(\hat{\theta}_n) - U_n(\theta_n) = \begin{pmatrix} \dot{U}_{1n}(\theta_n^{*(1)}) \\ \dot{U}_{2n}(\theta_n^{*(2)}) \\ \dot{U}_{3n}(\theta_n^{*(3)}) \\ \dot{U}_{4n}(\theta_n^{*(4)}) \end{pmatrix} (\hat{\theta}_n - \theta_n) = B_n(\hat{\theta}_n - \theta_n),$$

where  $\theta_n^{*(i)}$ ,  $i = 1, 2, 3, 4$  lie on the straight line joining  $\theta_n$  and  $\hat{\theta}_n$ . Since, with probability increasing to 1,  $\mathbb{U}_n(\hat{\theta}_n) = U_n(\theta_n) = 0$ , the following equality holds with probability increasing to 1:

$$\mathbb{U}_n(\hat{\theta}_n) - U_n(\hat{\theta}_n) = -B_n(\hat{\theta}_n - \theta_n).$$



Now,

$$\begin{aligned}
n^{1/2}[\mathbb{U}_n(\hat{\theta}_n) - U_n(\hat{\theta}_n)] &= -2\mathbb{G}_n \begin{pmatrix} Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau) \\ X(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau)) \\ q_n(X, \tau)(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau)) \\ \beta_2 \frac{\partial}{\partial \tau} q_n(X, \tau)(Y - \beta_0 - \beta_1 X - \beta_2 q_n(X, \tau)) \end{pmatrix}_{\theta=\hat{\theta}_n} \\
&= -2\mathbb{G}_n \begin{pmatrix} g_n^{(1)}(\hat{\theta}_n) \\ g_n^{(2)}(\hat{\theta}_n) \\ g_n^{(3)}(\hat{\theta}_n) \\ g_n^{(4)}(\hat{\theta}_n) \end{pmatrix} \\
&= -2\mathbb{G}_n(g_n(\hat{\theta}_n)).
\end{aligned}$$

Next, we argue that  $g_n$ , as well as  $g$  (defined as the limit of  $g_n$ ), belongs to Donsker class. Then we show that,  $P(g_n^{(i)}(\hat{\theta}_n) - g^{(i)}(\hat{\theta}_n))^2 = o_p(1)$ ,  $i = 1, \dots, 4$ , where  $g_n^{(i)}$  are components of  $g_n$ . Then, by the asymptotic equicontinuity property, we have  $\mathbb{G}_n(g_n(\hat{\theta}_n) - g(\hat{\theta}_n)) = o_p(1)$  (van der Vaart and Wellner 1996). It can also be shown that  $P(g^{(i)}(\hat{\theta}_n) - g^{(i)}(\theta^0))^2 = o_p(1)$ ,  $i = 1, \dots, 4$ . Hence, again by the asymptotic equicontinuity property, we have  $\mathbb{G}_n(g(\hat{\theta}_n) - g(\theta^0)) = o_p(1)$  (van der Vaart and Wellner 1996). Now,

$$\begin{aligned}
\sqrt{n}[\mathbb{U}_n(\hat{\theta}_n) - U_n(\hat{\theta}_n)] &= -2\mathbb{G}_n(g_n(\hat{\theta}_n)) \\
&= -2\mathbb{G}_n(g_n(\hat{\theta}_n) - g(\hat{\theta}_n)) - 2\mathbb{G}_n(g(\hat{\theta}_n) - g(\theta^0)) - 2\mathbb{G}_n(g(\theta^0)) \\
&= -2\mathbb{G}_n(g(\theta^0)) + o_p(1).
\end{aligned}$$

Thus by the central limit theorem, the above expression converges in distribution to a normal random variable with mean zero and variance matrix  $2\sigma^2 \dot{U}_*(\theta^0)$ . By the same argument in the proof of Theorem 1 we know that  $B_n$  converges to  $\dot{U}_*(\theta^0)$  in probability. Thus  $n^{1/2}(\hat{\theta}_n - \theta_n)$  converges to  $N(0, 2\sigma^2 \dot{U}_*^{-1}(\theta^0))$  in distribution, establishing Fact 4 and thus Theorem 2.

## Appendix 2: Proof of Theorem 3

The three main steps of this proof has been already outlined in Section 3.2. The proof of Theorem 3 relies on the following lemma:

**Lemma 1.** *Under Conditions 2.1-2.3 in Section 3.2, and for  $D \equiv m$ , for any positive definite matrix  $V_{m \times m}$  and  $\gamma_n = n^{-\alpha}$  with  $\alpha > 1/2$ ,  $n^{1/2}(\hat{\theta}_n^{(V)} - \theta^0)$  converges to  $N(0, K^{(V)})$  in distribution, where*

$$K^{(V)} = 2P[(H^T(\theta^0)VH(\theta^0))^{-1}H^T(\theta^0)V^T\Sigma_0VH(\theta^0)(H^T(\theta^0)VH(\theta^0))^{-1}].$$

Here,  $\hat{\theta}_n^{(V)}$  is defined as a zero of  $\mathbb{U}_n^{(V)}(\theta) = \frac{\partial}{\partial \theta} \mathbb{P}_n(Y - \mu_n)^T V(Y - \mu_n)$  in  $\Theta$ , with probability increasing to 1.

The proof of lemma 1 is similar to the proof of Theorem 2 as shown in Appendix 1. The details of the proof have been excluded for brevity. They have been presented in the Supplement.

Now to prove Theorem 3, using  $V = I$ , we obtain from Lemma 1 that  $n^{1/2}(\hat{\theta}_n^{(I)} - \theta^0)$  converges to  $N(0, K^{(I)})$  in distribution, for  $\gamma_n = n^{-\alpha}$ , with  $\alpha > 1/2$ .

Now because of Condition 2.4  $\sqrt{n}(\hat{\Sigma}_n^{-1} - W^{-1}) = O_p(1)$  implying,

$$\begin{aligned} & \sqrt{n}(\mathbb{U}_n(\hat{\theta}_n^{(W^{-1})}) - \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})})) \\ &= \frac{\partial}{\partial \theta} \mathbb{P}_n \left[ (Y - \mu_n)^T \{ \sqrt{n}(\hat{\Sigma}_n^{-1} - W^{-1}) \} (Y - \mu_n) \right] \Bigg|_{\theta = \hat{\theta}_n^{(W^{-1})}} \\ &= O_p(1) \frac{\partial}{\partial \theta} \mathbb{P}_n \left[ (Y - \mu_n)^T (Y - \mu_n) \right] \Bigg|_{\theta = \hat{\theta}_n^{(W^{-1})}} \\ &= O_p(1) \mathbb{U}_n^{(I)}(\hat{\theta}_n^{(W^{-1})}). \end{aligned}$$

Again  $\|\mathbb{U}_n^{(I)} - U^{(I)}\| = o_p(1)$  (proof provided in the Supplement). This implies that,  $\mathbb{U}_n^{(I)}(\hat{\theta}_n^{(W^{-1})}) = U^{(I)}(\hat{\theta}_n^{(W^{-1})}) + o_p(1)$ . Also  $\hat{\theta}_n^{(W^{-1})}$  converges in probability to  $\theta^0$  and  $U^{(I)}$  is continuous, which together imply that  $U^{(I)}(\hat{\theta}_n^{(W^{-1})}) = U^{(I)}(\theta^0) + o_p(1) = o_p(1)$ , since

$U^{(I)}(\theta^0) = 0$ . Thus, we obtain  $\mathbb{U}_n^{(I)}(\hat{\theta}_n^{(W^{-1})}) = o_p(1)$ , which implies that

$$\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n^{(W^{-1})}) - \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})})) = o_p(1).$$

Also  $\mathbb{U}_n(\hat{\theta}_n) = \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})}) = 0$  implies that

$$\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n) - \mathbb{U}_n(\hat{\theta}_n^{(W^{-1})})) = -\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n^{(W^{-1})}) - \mathbb{U}_n^{(W^{-1})}(\hat{\theta}_n^{(W^{-1})})) = o_p(1).$$

Taylor series expansion of  $\mathbb{U}_n(\hat{\theta}_n)$  around  $\hat{\theta}_n^{(W^{-1})}$  provides

$$\sqrt{n}(\mathbb{U}_n(\hat{\theta}_n) - \mathbb{U}_n(\hat{\theta}_n^{(W^{-1})})) = \dot{\mathbb{U}}_n(\tilde{\theta}_n^*)\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}),$$

for some  $\tilde{\theta}_n^*$  lying on the straight line joining  $\hat{\theta}_n$  and  $\hat{\theta}_n^{(W^{-1})}$ . As shown for  $A_n$  in the proof of Theorem 2, we can show that  $\dot{\mathbb{U}}_n(\tilde{\theta}_n^*)$  converges in probability to  $\dot{U}_*(\theta^0)$ , which in turn implies that  $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}) = o_p(1)$ .

Also, notice that, clearly with  $V = W^{-1}$ , from Lemma 1,  $n^{1/2}(\hat{\theta}_n^{(W^{-1})} - \theta^0)$  converges to  $N(0, K^{(W^{-1})})$  in distribution. Along with the fact that  $\sqrt{n}(\hat{\theta}_n - \hat{\theta}_n^{(W^{-1})}) = o_p(1)$ , we have proved Theorem 3.

## References

- Asgharian, M. and Wolfson, D. B. (2001), “Covariates in multipath change-point problems: Modelling and consistency of the MLE,” *Canadian Journal of Statistics*, 29, 515–528.
- Bellman, R. and Roth, R. (1969), “Curve fitting by segmented straight lines,” *Journal of the American Statistical Association*, 64, 1079–1084.
- Bhattachaya, P. K. (1987), “Maximum likelihood estimation of a change-point,” *J. Multivar. Analysis*, 23, 183–208.
- Brooking, I. R. and Jameison, P. D. (2002), “Temperature and photoperiod response of vernalization in near-isogenic lines of wheat,” *Field Crops Research*, 79, 21–38.
- Chen, C., Chan, J., Gerlach, R., and Hsieh, W. (2011), “A comparison of estimators for regression models with change points,” *Statistics and Computing*, 21, 395–414.
- Chiu, G., Lockhart, R., and Routledge, R. (2002), “Bent-cable Asymptotics when the Bend is Missing,” *Statistics and Probability Letters*, 59, 9–16.
- (2006), “Bent-cable regression theory and applications,” *Journal of the American Statistical Association*, 101, 542–553.
- Feder, P. I. (1975a), “The log likelihood ratio in segmented regression,” *The Annals of Statistics*, 3, 84 – 97.
- (1975b), “On asymptotic distribution theory in segmented regression problems-identified case,” *The Annals of Statistics*, 3, 49 – 83.
- Hudson, D. J. (1966), “Fitting segmented curves whose join points have to be estimated,” *Journal of the American Statistical Association*, 61, 1097 – 1129.
- Huskova, M. (1998), “Estimators in the location model with gradual changes,” *Commentationes Mathematicae Universitatis Carolinae*, 39, 147–157.
- Joseph, L. and Wolfson, D. (1993), “Maximum likelihood estimation in the multi-path change-point problem,” *Annals of the Institute of Statistical Mathematics*, 45, 511530.

- Lu, X., Nan, B., Song, P., and Sowers, M. F. (2010), “Longitudinal Data Analysis with Event Time as a Covariate,” *Stat Biosci*, 2, 65 – 80.
- Lund, R. B., Hurd, H. L., Bloomfield, P., and Smith, R. L. (1995), “Climatological time series with periodic correlation,” *J. Climate*, 8, 2787 – 2809.
- Lund, R. B. and Reeves, J. (2002), “Detection of undocumented changepoints: A revision of the two-phase regression model,” *J. Climate*, 15, 2547 – 2554.
- Muggeo, V. M. R. (2003), “Estimating regression models with unknown break-points,” *Statistics in Medicine*, 22, 3055–3071.
- Poirer, D. J. (1973), “Piecewise regression using cubic splines,” *Journal of the American Statistical Association*, 68, 515 – 524.
- Robinson, D. E. (1964), “Estimates for the points of intersection of two polynomial regressions,” *Journal of the American Statistical Association*, 59, 214 – 224.
- Rubin, D. (1976), “Inference and missing data,” *Biometrika*, 63, 581–592.
- Sowers, M. F. R., McConnell, D., Nan, B., Harlow, S., and Randolph Jr., J. F. (2008), “Estradiol rates of change in relation to the final menstrual period in a population-based cohort of women,” *J. Clin. Endocrinal Metab.*, 93(10), 3847 – 3852.
- Tishler, A. and Zang, I. (1981), “A new maximum likelihood algorithm for piecewise regression,” *Journal of the American Statistical Association*, 76, 980–987.
- van der Vaart, A. and Wellner, J. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer.
- (1999), “Preservation Theorems for Glivenko-Cantelli and Uniform Glivenko-Cantelli Classes,” *Technical Report No. 361, Department of Statistics, University of Washington*.

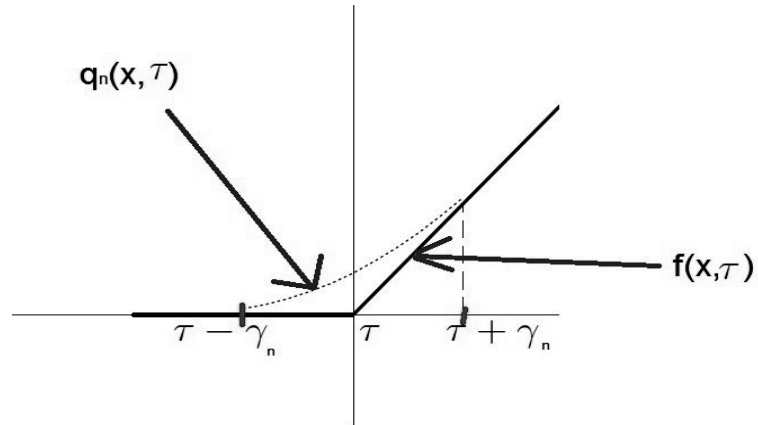


Figure 1:  $q_n$  is the smoothed version of  $f$ .

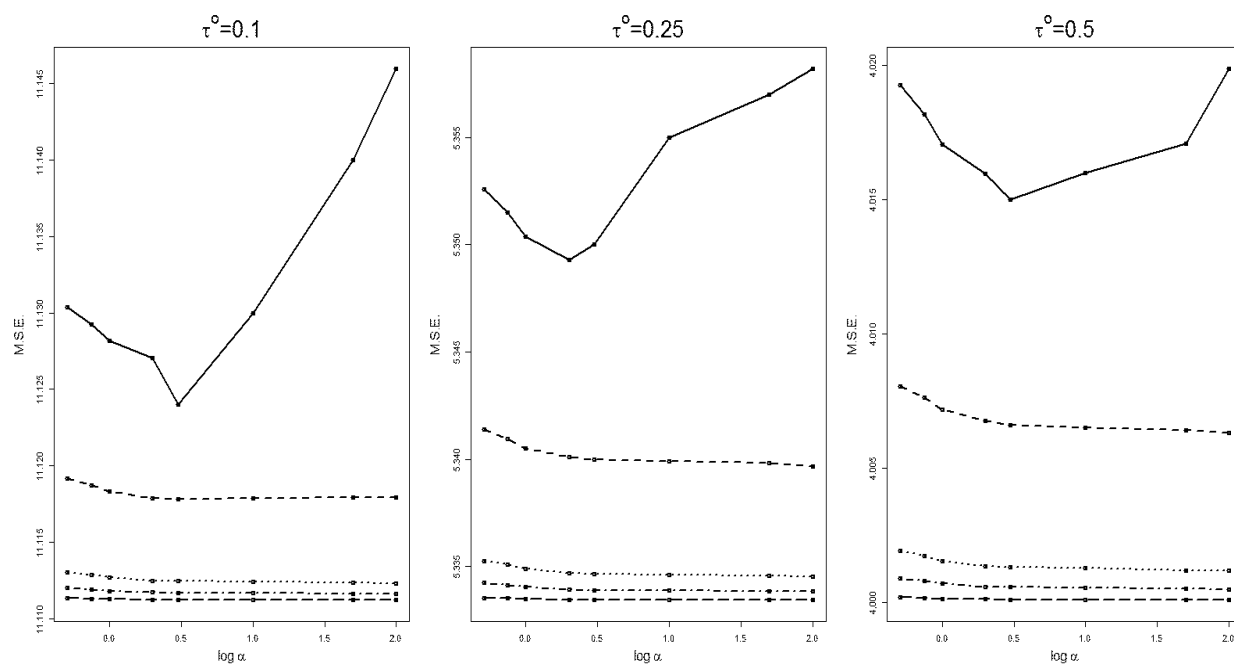


Figure 2: Mean Square Errors vs  $\log_{10} \alpha$  for varying sample-sizes with different  $\tau$ -values, where  $\beta_0^0 = 0.3$ ,  $\beta_1^0 = 1.5$ ,  $\beta_2^0 = 1$  and  $\sigma = 0.5$ . From the top below, the solid line corresponds to  $n = 50$ , dashed line corresponds to  $n = 100$ , the dotted line corresponds to  $n = 500$ , the dot-dash line corresponds to  $n = 1000$  and the longdash line corresponds to  $n = 5000$ .

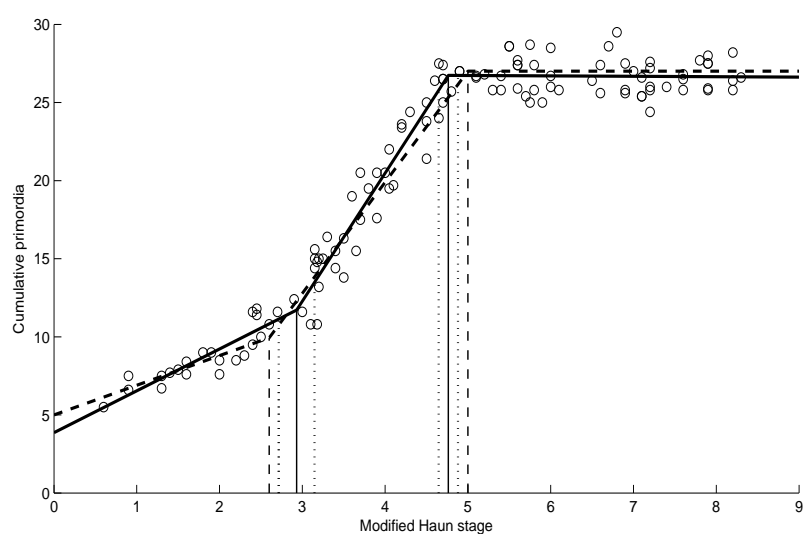


Figure 3: Co-ordination of primordium initiation and leaf emergence from Spring Batten treatments resulting in a final leaf number of 8 Brooking and Jameison (2002). The solid bold line represents the one estimated by our approach while the broken line represents the one estimated by Brooking and Jameison (2002). The dotted vertical lines give the confidence intervals for the estimated change-points given by the solid lines while the vertical broken-lines indicate the eye-estimated change-points.



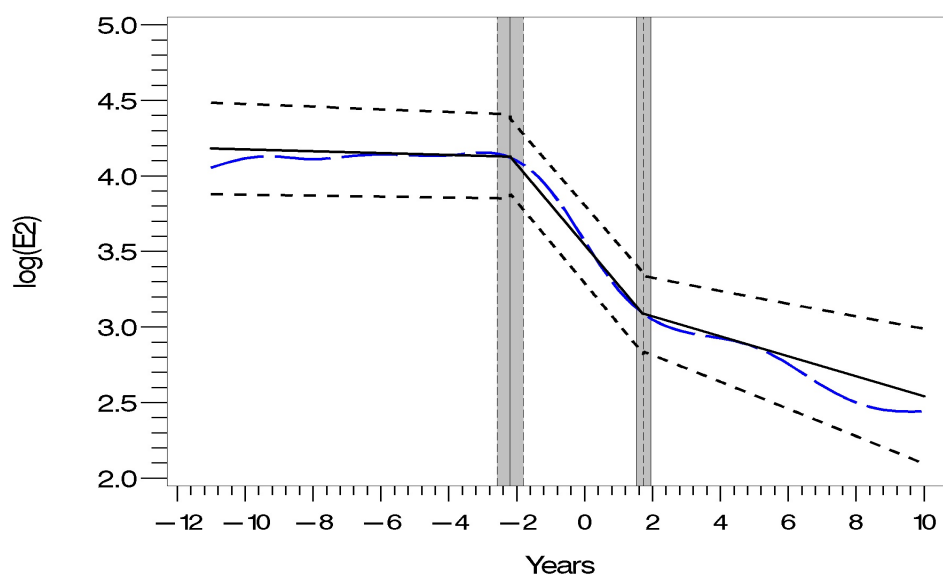


Figure 4: E2 profile analysis at baseline mean BMI for a non-smoker: the solid line represents the mean estimator using two change-point broken-stick model, the short-broken lines the corresponding pointwise 95% confidence bands; the long-broken lines represent the smooth estimator of the mean function from semiparametric mixed effects model using the same method as in Sowers et al. (2008); the shaded regions represent the 95% confidence intervals for the two change-points.

Table 1: Simulation results comparing the run-times of the existing (Hudson 1966) and proposed methods for one and two change-point(s) model, with ratio of the time taken by the existing method with respect to that of the proposed one.

Sample Size $n$	Mean Time (Seconds)					
	One change-point			Two change-points		
	Existing	Proposed	Ratio	Existing	Proposed	Ratio
50	0.18	0.006	30	2.36	0.02	118
100	0.30	0.008	38	13.86	0.03	462
500	0.97	0.02	49	64.87	0.06	1015
1000	1.89	0.03	63	947.03	0.08	11838
5000	4.98	0.06	83	22843	0.20	114215

Table 2: Bias and variances for the change-point estimate  $\hat{\tau}_n$  compared for one change-point problem in 3 setups: A:  $\theta^T = (0.2, 1, 1, 0.6)$ , B:  $\theta^T = (0.3, 1.5, 1, 0.8)$  & C:  $\theta^T = (0.3, 1.5, -1, 0.2)$  (S.D.: Average of estimated standard deviations over 1000 replications; Emp. S.D. : Sample standard deviation based on 1000 replications).

Sample Size $n$	Bias (S.D., Emp. S.D.) $\times 10^{-3}$		Theoretical S.D. $\times 10^{-3}$
	Existing	Proposed	
Set-up A			
50	-8.2 (58.1, 60.3)	-12.4 (58.3, 61.0)	57.8
100	-4.3 (41.0, 41.2)	-5.2 (41.0, 41.3)	40.9
500	-2.1 (18.3, 18.4)	-2.9 (18.3, 18.5)	18.3
1000	-0.6 (12.9, 12.9)	-0.9 (12.9, 13.0)	12.9
5000	-0.0 (5.8, 5.8)	-0.1 (5.8, 5.8)	5.8
Set-up B			
50	-9.2 (71.6, 73.2)	-14.1 (71.6, 73.6)	70.7
100	-4.7 (50.1, 50.3)	-5.9 (50.1, 50.4)	50.0
500	-2.8 (22.4, 22.4)	-3.3 (22.4, 22.5)	22.3
1000	-0.9 (15.8, 15.8)	-1.2 (15.8, 15.8)	15.8
5000	-0.1 (7.1, 7.1)	-0.1 (7.1, 7.1)	7.1
Set-up C			
50	8.1 (71.6, 73.3)	12.8 (71.6, 73.5)	70.7
100	4.7 (50.1, 50.4)	6.2 (50.1, 50.5)	50.0
500	1.8 (22.4, 22.4)	2.0 (22.5, 22.6)	22.3
1000	0.2 (15.8, 15.8)	0.4 (15.8, 15.8)	15.8
5000	0.1 (7.1, 7.1)	0.1 (7.1, 7.1)	7.1

Table 3: Bias and variances for the change-point estimate  $\hat{\tau}_n$  compared for two change-points problem in 3 setups: D:  $\theta^T = (0.3, 1, 1, 1, 0.2, 0.8)$ , E:  $\theta^T = (0.2, 1, 2, 1, 0.4, 0.6)$  & F:  $\theta^T = (0.3, 1, -1, 1, 0.2, 0.8)$  (S.D.: Average of estimated standard deviations over 1000 replications; Emp. S.D. : Sample standard deviation based on 1000 replications).

Sample Size ( $n$ )	$\hat{\tau}_{1n}$ Bias (S.D., Emp. S.D.) $\times 10^{-3}$		Theo. S.D. $\times 10^{-3}$	$\hat{\tau}_{2n}$ Bias $\times 10^{-3}$ (S.D., Emp. S.D.) $\times 10^{-3}$		Theo. S.D. $\times 10^{-3}$
	Existing	Proposed		Existing	Proposed	
	Set-up D					
50	10.2 (63.2, 63.9)	20.3 (63.3, 64.1)	62.0	-11.1 (55.0, 55.5)	-19.2 (55.1, 55.6)	54.0
100	5.1 (44.2, 44.6)	6.3 (44.3, 44.8)	43.8	-4.8 (38.7, 39.2)	-5.9 (38.8, 39.3)	38.2
500	2.8 (19.7, 19.8)	3.7 (19.8, 19.9)	19.6	-3.0 (17.3, 17.5)	-4.0 (17.3, 17.5)	17.1
1000	0.9 (13.9, 13.9)	1.1 (13.9, 14.0)	13.9	-1.0 (12.2, 12.3)	-1.1 (12.2, 12.3)	12.1
5000	0.1 (6.2, 6.2)	0.2 (6.2, 6.2)	6.2	-0.1 (5.4, 5.5)	-0.1 (5.4, 5.5)	5.4
Set-up E						
50	-32.9 (14.1, 17.9)	-41.1 (14.8, 19.0)	11.0	40.0 (22.4, 24.1)	51.2 (22.8, 24.9)	21.0
100	-7.2 (9.6, 10.2)	-9.0 (10.0, 10.5)	7.7	8.1 (15.5, 16.4)	9.7 (15.8, 16.7)	14.8
500	-5.1 (3.7, 3.7)	-6.2 (3.7, 3.8)	3.5	6.0 (6.8, 7.0)	6.8 (6.9, 7.1)	6.6
1000	-1.3 (2.5, 2.5)	-1.4 (2.5, 2.5)	2.4	1.4 (4.8, 4.8)	1.5 (4.8, 5.0)	4.7
5000	-0.1 (1.1, 1.1)	-0.1 (1.1, 1.1)	1.1	0.1 (2.1, 2.1)	0.2 (2.1, 2.2)	2.1
Set-up F						
50	10.2 (63.2, 64.0)	19.8 (63.3, 64.7)	62.0	-10.4 (55.0, 0.155)	-19.0 (55.3, 55.8)	54.0
100	4.9 (44.0, 44.6)	6.0 (44.3, 44.8)	43.8	-5.3 (38.7, 39.1)	-6.1 (38.8, 39.3)	38.2
500	3.0 (19.7, 19.8)	4.0 (19.8, 19.9)	19.6	-3.8 (17.2, 17.5)	-4.3 (17.3, 17.4)	17.1
1000	0.9 (13.9, 14.0)	1.2 (13.9, 14.0)	13.9	-0.8 (12.1, 12.2)	-1.0 (12.2, 12.3)	12.1
5000	0.1 (6.2, 6.2)	0.1 (6.2, 6.2)	6.2	-0.1 (5.4, 5.5)	-0.1 (5.4, 5.5)	5.4

Table 4: Bias and variances for the change-point estimate  $\hat{\tau}_n$  compared for two change-points problem in 3 longitudinal setups: G:  $\theta^T = (0.3, 1, 1, 1, 0.2, 0.8)$ , H:  $\theta^T = (0.2, 1, 2, 1, 0.4, 0.6)$  & J:  $\theta^T = (0.2, 1, 2, 1, 0.4, 0.6)$ . 10 observations per individual in set-ups G and H. In set-up J, number of observations per individual  $D \sim \text{Discrete Uniform } \{1, 2, \dots, 20\}$  (S.D.: Average of estimated standard deviations over 1000 replications; Emp. S.D. : Sample standard deviation based on 1000 replications).

Sample Size ( $n$ )	$\hat{\tau}_{1n}$			$\hat{\tau}_{2n}$		
	Bias (S.D., Emp. S.D.) $\times 10^{-3}$		Theo. S.D. $\times 10^{-3}$	Bias $\times 10^{-3}$ (S.D., Emp. S.D.) $\times 10^{-3}$		Theo. S.D. $\times 10^{-3}$
	Existing	Proposed		Existing	Proposed	
Set-up G						
10	5.4 (74.2, 74.6)	6.9 (74.3, 74.8)	62.8	-5.3 (56.7, 57.3)	-6.6 (56.8, 58.3)	43.2
50	3.2 (27.7, 27.9)	4.0 (27.8, 27.9)	21.4	-3.3 (21.3, 21.5)	-4.3 (21.3, 21.5)	19.8
100	1.1 (19.8, 20.0)	1.4 (19.9, 20.1)	16.0	-1.2 (15.3, 15.4)	-1.5 (15.4, 15.5)	14.1
500	0.2 (7.7, 7.7)	0.3 (7.9, 8.2)	6.9	-0.2 (6.4, 6.5)	-0.2 (6.4, 6.5)	5.9
Set-up H						
10	-7.5 (13.6, 13.2)	-9.7 (13.8, 13.5)	8.7	8.1 (25.5, 26.4)	9.8 (25.8, 27.7)	17.1
50	-5.3 (6.7, 6.7)	-6.8 (6.7, 6.8)	4.6	6.3 (9.9, 10.2)	7.1 (9.8, 10.1)	8.2
100	-1.4 (4.5, 4.4)	-1.6 (4.5, 4.5)	2.9	1.4 (6.8, 6.4)	1.7 (6.8, 6.1)	5.5
500	-0.1 (2.8, 2.8)	-0.1 (2.8, 2.8)	1.9	0.1 (4.1, 4.1)	0.2 (4.1, 4.2)	2.8
Set-up J						
10	5.8 (77.1, 77.6)	7.1 (77.2, 77.8)	64.4	-6.1 (57.7, 59.4)	-6.8 (57.2, 59.1)	44.1
50	3.5 (28.5, 28.7)	4.1 (28.5, 28.7)	22.9	-3.9 (22.0, 22.2)	-4.4 (21.8, 22.0)	20.6
100	1.2 (20.3, 20.6)	1.4 (20.4, 20.6)	16.6	-1.3 (15.7, 15.8)	-1.5 (15.8, 15.9)	14.9
500	0.2 (7.9, 8.0)	0.3 (8.0, 8.2)	7.2	-0.2 (6.8, 6.9)	-0.2 (6.8, 6.9)	6.4

Table 5: Regression parameter estimates along with their respective standard errors

Parameter	Estimate	Standard Error
$\beta_0$	4.116	0.139
$\beta_1$	-0.006	0.002
$\beta_2$	-0.259	0.009
$\beta_3$	0.199	0.008
$\tau_1$	-2.192	0.197
$\tau_2$	1.738	0.11
$\lambda_1$	0.047	0.072
$\lambda_2$	0.005	0.005