Open Journal of Applied Sciences, **2013, 3, 1-6**
doi:10.4236/ojapps.2013.31B1001 Published Online April 2013 (http://www.scirp.org/journal/ojapps)

Changepoint Analysis by Modified Empirical Likelihood Method in Two-phase Linear Regression Models

Hualing Zhao[1], Hanfeng Chen[2], Wei Ning[2]
[1]Department of Statistics, School of Science, Wuhan University of Technology, Wuhan, Hubei , P.R. of China
[2]Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio , USA
Email: hualingbo324@126.com, hchen@bgsu.edu, wning@bgsu.edu.

Received 2013

ABSTRACT

A changepoint in statistical applications refers to an observational time point at which the structure pattern changes during a somewhat long-term experimentation process. In many cases, the change point time and cause are documented and it is reasonably straightforward to statistically adjust (homogenize) the series for the effects of the changepoint. Sadly many changepoint times are undocumented and the changepoint times themselves are the main purpose of study. In this article, the changepoint analysis in two-phrase linear regression models is developed and discussed. Following Liu and Qian (2010)'s idea in the segmented linear regression models, the modified empirical likelihood ratio statistic is proposed to test if there exists a changepoint during the long-term experiment and observation. The modified empirical likelihood ratio statistic is computation-friendly and its p-value can be easily approximated based on the large sample properties. The procedure is applied to the Old Faithful geyser eruption data in October 1980.

Keywords: Changepoint; Extreme-Value Distribution; Modified Empirical Likelihood Ratio; Segmented Linear Regression

1. Introduction

In recent years increasing interest has been shown in changepoint analysis in two-phrase linear regression models. A changepoint in statistical applications refers to an observational time point at which the structure pattern changes during a somewhat long-term experimentation process. In many cases, the change point time and cause is documented and it is reasonably straightforward to statistically adjust (homogenize) the series for the effects of the changepoint. Sadly many changepoint times are undocumented and the changepoint times themselves are the main interest of study. For example, one of the most important problems in economics is to determine as early as possible the starting as well as ending time point of a suspected ongoing recession. In the environmental sciences, scientists are of great interest to understand when the global warming started or the Earth's mean surface temperature rise in the past decades should be explained by the normal variability of the Earth's surface temperature over time. (Indeed the official position of the World Natural Health Organization in regards to global warming is that there is no global warming and claims that global warming is nothing more than just another hoax. See their official website: http://www.wnho.net.)

The two-phrase linear regression model may be expressed as follows:

$$y_i = x_i'\alpha I(i \le k) + x_i'\beta I(i > k) + \epsilon_i,\ i = 1, \cdots, n, \quad (1)$$

where $x_i \in R^p$ are covariates, α and β are p-dimensional regression parameters, $1 \le k \le n$ is a putative changepoint at which the liner regression model changes from one phrase to another, and the ϵ_i are assumed to be independent and identically distributed unobservable measurement errors. The main interest in the two-phrase linear regression model is to determine whether such a change of phrase occurs or not and if it does, when the change happens during the experiment or observation. In the special case of simple linear regression, the model (1) is often called segmental linear regression model. As remarked by Liu and Qian (2010), widespread applications of two-phrase linear regression model (1) have appeared in diverse research areas. See, e.g., in environmental sciences, Piegorsch and Bailer (1997), in medical science, Simith and Cook (1980), in epidemiology Pastor and Gullar (1998), in econometrics Fiteni (2004) and Koul and Qian (2002), just for a few.

As described above, with responses y_i and covariates x_i, central to the problem is to determine whether there exists a changepoint during the long-term experiment or observation. In terms of statistical inference, that is to test

$$H_0 : \alpha = \beta \quad \text{versus} \quad H_1 : \alpha \ne \beta$$

Put $y=(y_1,\cdots,y_n)'$, $\theta=(\alpha',\beta')'$, and $\epsilon=(\epsilon_1,\cdots,\epsilon_n)'$. Let

$$\theta=\begin{pmatrix}\alpha\\ \beta\end{pmatrix},\quad X_k=\begin{pmatrix}X_{1k} & 0\\ 0 & X_{2k}\end{pmatrix},$$

where $X_{1k}=(x_1,\cdots,x_k)'$, $X_{2k}=(x_{k+1},\cdots,x_n)'$.

Then the model(1) has the matrix expression:

$$y=X_k\theta+\epsilon \tag{2}$$

Dong (2004) proposes an empirical likelihood-type Wald statistic to infer the changepoint. More recently, Liu and Qian (2010) proposes an interesting and computationally easy empirical likelihood detecting procedure in the segmented linear regression model. In this paper, their ideas are applied to the model (1) to present a modified empirical likelihood ratio statistic to test H_0.

The article is organized as follows. The modified empirical likelihood ratio test procedure and its computational issues are present and discussed in the next section. The null distribution of the modified empirical likelihood ratio test statistic is studied for large samples and the results are put in the Appendix for interested readers. The modified empirical likelihood method is applied to a real-life data set for changepoint analysis in Section 3.

2. The Modified Empirical Likelihood Method

Following Liu and Qian (2010)'s ideas, the modified empirical likelihood method for changepoint analysis in the two-phrase linear regression mode(1) is described as follows: For each given *k*, estimate the regression parameters by least-square methods for each segment, fit the response y_i at x_i via the least-square estimate of the regression parameters for the segment of counter-part, and then construct the empirical likelihood ratio statistic based the fitting residuals. In the notations introduced in the last section, the least-square estimates for α and β are

$$\hat{\alpha}_k=(X'_{1k}X_{1k})^{-1}X'_{1k}y_{1k};\ \hat{\beta}_k=(X'_{2k}X_{2k})^{-1}X'_{2k}y_{2k}$$

where $y_{1k}=(y_1,\cdots,y_k)'$ and $y_{2k}=(y_{k+1},\cdots,y_n)'$.

Define

$$\tilde{e}_i(k)=y_i-\{x'_i\hat{\beta}_kI(i\le k)+x'_i\hat{\alpha}_kI(i>k)\}, \tag{3}$$

for $i=1,\cdots,n$ and

$$\mathcal{R}(k)=\sup\left\{\prod_{i=1}^{n}nw_i\,|w_i\tilde{e}_i(k)=0,w_i\ge 0,w_i=1\right\}. \tag{4}$$

The modified empirical likelihood ratio statistic is

$$M_n=\max_{p\le k\le n-p}\{-2\log\mathcal{R}(k)\}. \tag{5}$$

Recall that p is the dimension of the covariates, so equal to the number of regression parameters in each phrase. Reject the null hypothesis $H_0:\alpha=\beta$ and assert that a changepoint occurs, whenever M_n is significantly large.

It should be noted that the residuals $\tilde{e}_i(k)$ are not the ordinary least-squares fitting residuals but the residuals of fitting y_i at x_i with swapped least-square estimates of the regression parameters. Motivation leading to the modified empirical likelihood ratio statistics M_n is that $E\tilde{e}_i(k)=0$ if and only if $\alpha=\beta$, i.e. H_0 holds.

Through simulation studies, Liu and Qian (2010) investigate whether $\sqrt{M_n}$ has an asymptotic Gumbel extreme value distribution under the null hypothesis. We establish the null asymptotic theory of $\sqrt{M_n}$ that is given in the Appendix for interested readers. It is proved under regular conditions that if the null hypothesis $\alpha=\beta$ is true, $\sqrt{M_n}$ can be approximated by $\sqrt{T_n}$ in probability with an approximation error in size $n^{-\tau_2}$ for some constant $\tau_2>0$, where

$$T_n^{1/2}=\max_{p\le k\le n-p}n^{1/2}\{|\bar{e}(k)|/s(k)\}, \tag{6}$$

with

$$\bar{e}(k)=\frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i(k),\quad s^2(k)=\frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2(k).$$

It is then shown that for any *t*,

$$\lim_{n\to\infty}P\left\{A(\log n)T_n^{1/2}\le t+D_p(\log n)\right\}=\exp\{-2e^{-t}\},$$

where $D_p(x)=2\log x+(p/2)\log\log x-\log\Gamma(p/2)$ and $A(x)=(2\log x)^{1/2}$. Thus for any *t*,

$$\lim_{n\to\infty}P\left\{A(\log n)M_n^{1/2}\le t+D_p(\log n)\right\}=\exp\{-2e^{-t}\}. \tag{7}$$

The above formula indicates that the limiting extreme-value distribution has a convergence rate of $\log n$. For this reason, the authors suggest to use the distribution of $\sqrt{T_n}$ under null hypothesis to approximate the *p*-value of $\sqrt{M_n}$ in applications. As the asymptotic null distribution is free of any population distribution, one can easily approximate the *p*-value of $\sqrt{M_n}$ by Monte Carlo methods through simulating the null distribution of $\sqrt{T_n}$.

The main advantage of the modified empirical likelihood testing procedure based on M_n is its easiness of computation. The R package **emplik** can be used to compute $\mathcal{R}(k)$.

As many researchers remarked (Liu and Qian, 2010; Csörgő and Horváth, 1997), the statistic $-2\log R(k)$ is sensitive to outliers when *k* is too small or too close to the sample size. Csörgő and Horváth (1997) proposed the trimmed idea to overcome the problem. Let $d\le k_{n1}<k_{n2}\le n-d$. Define

$$M'_n=\max_{k_{n1}\le k\le k_{n2}}\{-2\log R(k)\}.$$

when it is assumed that as $n\to\infty$,$u_n=(n^2-k_{n1}k_{n2})$

$/[k_{n1}(n-k_{n2})] \to \infty,$ we have

$$\lim_{n\to\infty} P\{A(\log u_n)[M_n']^{1/2} \le t + D_p(\log u_n)\} = \exp(-e^{-t}),$$

when k_{n1} and $n-k_{n2}$ are chosen to be constant, $u_n \to \infty$. Liu and Qian (2010) suggests to use $k_{n1} = \log^2 n$ and $k_{n2} = n - k_{n1}$. Such a choice clearly satisfies. Another popular choice is $k_{n1} = 2\log n,\ k_{n2} = n - 2\log n$; see Perron and Vogelsang (1992). In particular, if for $0 < \lambda < 1,\ k_{n1} = [\lambda n]$ and $k_{n2} = n - [\lambda n]$ where $[x]$ is the greatest integer less than or equal to *x*, by Corollary A.3.1 of Csörgő and Horváth (1997),

$$\lim_{n\to\infty} P\{A(\log n)[M_n']^{1/2} \le x + D_d(\log n)\} = \exp(-2e^{-x}).$$

3. A Real-Life Example

We now apply the modified empirical likelihood method to the Old Faithful geyser in the Yellowstone National Park of USA. A geyser is a hot spring that occasionally becomes unstable and erupts hot water and steam into air. If we can find the relationship between the duration of the eruptions and the interval to next eruption, then the time of next eruption can be predicted. The data of 270 eruptions of the Old Faithful geyser in October 1980 can be found in Weisberg (2005). **Figure 1** is the scatterplot of intervals(y) to the next eruption versus the duration(X) of the eruptions. The scatter plot suggests that the relationship has two phases.

In this example, $n = 270$ and $p = 2$. We adopt $k_{n1} = [\log^2 n] = 31$ and $k_{n2} = n - k_{n1} = 239$. Thus $u_n = 68$, so that $\log u_n = 4.2$, $A(\log u_n) = 2.9$ and $D_2(\log u_n) = 9.84$.The function el.test in R package emplik is used to compute the test statistics M_n' and it appears that $\{M_n'\}^{1/2} = 30.34$. Thus

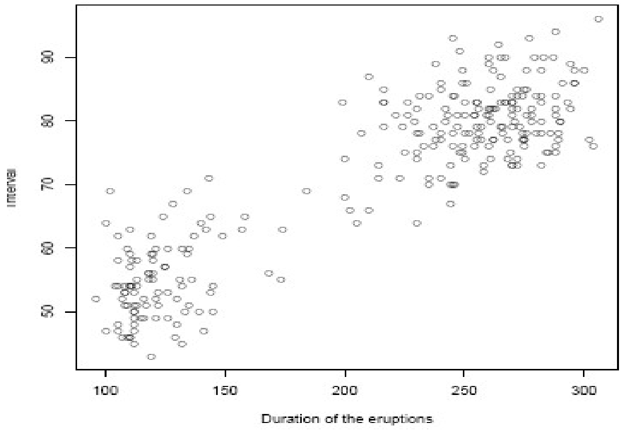

Figure 1. Scatter plot of 270 eruptions of the Old Faithful geyser in October 1980 in Yellowstone National Park USA.

$$\Xi_n = A(\log u_n)\{M_n'\}^{1/2} - D_2(\log u_n) = 78.15.$$

According the asymptotic null distribution discussed in last section, the *p*-value with the observed $\Xi_n = 78.15$ is approximately $1 - \exp\{-2e^{-78.15}\}$ that is very close to 0, leading to the assertion that there exists a change-point during the 270 eruptions of the Old Faithful geyser in October 1980.

4. Acknowledgements

The research is partially funded by the Fundamental Research Funds for the Central Universities (No. 2011-IV-116).

REFERENCES

[1] M. Csorgo and L. Horvnth, "Limit Theorem in Change-Point Analysi," *Wiley Series in Probability and Statistics,* John Wiley & Sons: New York, 1997.

[2] I. Fiteni, "τ-estimators of Regression Models with Structural Change of Unknown Location," *Journal of Econometrics* , Vol. 119, No. 1, 2004, pp. 19-44. doi:10.1016/S0304-4076(03)00153-2

[3] Z. Liu and L. Qian, "Changepoint Estimation in a Segmented Linear Regression via Empirical Likelihood," *Communications in Statistics--Simulation and Computation,* Vol. 89**,** 2010**,** pp. 85-100.

[4] L. H. Koul and L. F. Qian, "Asymptotics of Maximum Likelihood Estimator in a Two-phaselinear Regression Model,"*Journal of Statistical Planning and Inference,* Vol. 108, No. 1-2, 2002, pp. 99-119. doi:10.1016/S0378-3758(02)00273-2

[5] A. B. Owen, "Empirical Likelihood for Linear Models," *Annals of Statistics*, Vol.19, No.19, 1991, pp. 1725-1747. doi:10.1214/aos/1176348368

[6] A. B. Owen, "Empirical Likelihood," New York: Chapman & Hall, 2001. doi:10.1201/9781420036152

[7] R. Pastor and E. Guallar, "Use of Two-segmented Logistic Regression to Estimate Changepoints in Epidemiologic Studies," *American Journal of Epidemiology*, Vol. 148, No. 7, 1998, pp. 631-642. doi:10.1093/aje/148.7.631

[8] P. Perron and T. J. Vogelsang, "Testing for a Unit Root in a Time Series with a Changing Mean: Corrections and Extensions," *J. Business Econom. Statist.*,Vol. 10,1992, pp. 467-470.

[9] W. W. Piegorsch and A. J. Bailer, "Statistics for Environmental Biology and Toxicology," London: Chapman and Hall, 1997.

[10] A. M. F. Smith and D. G. Cook, "Straight Lines with a Change Point: A Bayesian Analysis of Some Renal Transplant Data," *Applied Statistics*, Vol. 29, No. 2, 1980，pp. 180-189. doi:10.2307/2986304

[11] S. Weisberg, "Applied Linear Regeression," 3th Edition, John Wiley& Sons, Inc., Hoboken, New Jersey, 2005.

Appendix: Asymptotic Null Distribution

The asymptotic null distribution of the modified empirical likelihood ratio test statistic M_n is established under the two-phrase linear regression model (1) that includes the segmented simple linear regression model considered by Liu and Qian (2010) as a special case.

With $\tilde{e}_i(k)$'s, $\mathcal{R}(k)$is defined by (4), and by Lagrange multiplier method,

$$-2\log\mathcal{R}(k) = 2\left\{\sum_{i=1}^n \log[1+\hat{\lambda}(k)\tilde{e}_i(k)]\right\},$$

where $\hat{\lambda}(k)$ is the root of

$$\sum_{i=1}^n \frac{\tilde{e}_i(k)}{1+\lambda\tilde{e}_i(k)} = 0. \qquad (8)$$

According to (5), M_n is defined as follows:

$$M_n = \max_{p\le k\le n-p}\{-2\log\mathcal{R}(k)\}. \qquad (9)$$

Regular conditions needed are listed as follows. Assume

C.1 rank (X_{1k})=rank$(X_{2k}) = d$ for $p \le k \le n-p$.

C.2 There are some $\nu > 0$, $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$, and positive-definite matrices Σ_1, Σ_2 such that as $k\to\infty$ and $n-k\to\infty$,

$$\left|\frac{1}{k}X'_{1k}X_{1k} - \Sigma_1\right| = o(r(k)),$$

$$\left|\frac{1}{n-k}X'_{2k}X_{2k} - \Sigma_2\right| = o(r(n-k)), \qquad (10)$$

$$\left|\sum_{i=1}^k x_i)'(X'_{2k}X_{2k})^{-1}(\sum_{i=1}^k x_i) - \sigma_1^2\right| = o(r(n-k)), \quad (11)$$

$$\left|(\sum_{j=k+1}^n x_j)'(X'_{1k}X_{1k})^{-1}(\sum_{j=k+1}^n x_j) - \sigma_2^2\right| = o(r(k)) (12)$$

where $r(t) = 1/(\log t)^\nu$, and $|\cdot|$ is the ordinary norm: $|(a_{ij})| = (\sum_i\sum_j a_{ij}^2)^{1/2}$.

C.3 There is some $\delta > 0$ such that
$\max_{1\le i\le n}|x_i| = o(n^{1/(2+\delta)})$, and $E|\epsilon_i|^{2+\delta} < \infty$.

Assumption C.2 is slightly weaker than C.9 in Csörgő and Horváth (1997, page 204) that assumes $\Sigma_1 = \Sigma_2$. In the two-phrase linear regression model, one is concerned with a slicing rule in the covariate variables. As a result, $(1/k)X'_{1k}X_{1k}$ and$(1/(n-k))X'_{2k}X_{2k}$may have different limits if existing. In the commonly adapted regression model that (y_i, x_i)'s are an independent and identically distributed sample with $E|(y_i,x_i)|^{2+\delta} < \infty$ for some $\delta > 0$, it is easily seen that C.2 and C.3 hold in probability one.

Theorem 1. Assume that H_0 hold and C.1-C.3 are satisfied with some $\nu > 2 + 27/\min(1,\delta)$. Then under the null model,

$$\begin{aligned} &\lim_{n\to\infty} P\left\{A(\log n)\sqrt{M_n} \le t + D_p(\log n)\right\} \\ = \ & \exp\{-2e^{-t}\}, \end{aligned} \qquad (13)$$

for any *t*, where $A(x) = (2\log x)^{1/2}$ and

$$D_p(x) = 2\log x + (p/2)\log\log x - \log\Gamma(p/2).$$

The main idea of proof of Theorem 1 is to use Owen (1991)'s arguments to obtain a quadratic approximation to $\mathcal{R}(k)$ so that the limit (1) follows from that of the classic parametric likelihood ratio test. The crucial step in Owen (1991)'s arguments is to approximate $\hat{\lambda}(k)$ up to an order of $O_P(n^{-1/2})$ uniformly in k in order to capture the leading terms in the Taylor's expansion of $\mathcal{R}(k)$. The first lemma gives an order estimate for $\max\{\tilde{e}_i(k)\}$. Denote $\bar{e}(k) = (1/n)\sum_{i=1}^n \tilde{e}_i(k)$ and $s^2(k) = (1/n)\sum_{i=1}^n \tilde{e}_i^2(k)$.

Lemma 1 Assume that H_0 and C.1-C.3 hold. Then

$$\max_{1\le i\le n}\{\max_{p\le k\le n-p}|\tilde{e}_i(k)|\} = O_P(n^{1/(2+\delta)}).$$

Proof. Under $H_0: \alpha = \beta$, both $\hat{\alpha}_k$ and $\hat{\beta}_k$ have the same mean. Let $\gamma_1 = (I_k, 0)\epsilon$ and $\gamma_2 = (0, I_{n-k})\epsilon$. Under H_0, we can express

$$\begin{aligned} \tilde{e}_i &= \epsilon_i - x'_i(X'_{2k}X_{2k})^{-1}X'_{2k}\gamma_2 I(i\le k) \\ &- x'_i(X'_{1k}X_{1k})^{-1}X'_{1k}\gamma_1 I(i > k) \end{aligned} \qquad (14)$$

By C.3, $E\{|\epsilon|^{(2+\delta)/2}\}^2 < \infty$. Thus from Lemma 11.2 in Owen (2001), $\max|\epsilon_i|^{(2+\delta)/2} = o(n^{1/2})$, implying

$$\max\{|\epsilon_i|\} = O_P(n^{1/(2+\delta)}) \qquad (15)$$

Next, by C.2 and the law of the iterated logarithm,

$$\max_{1\le i\le n}\left\{\max_{p\le k\le n-p}(n-k)^{1/2}\frac{|(X'_{2k}X_{2k})^{-1}X'_{2k}\gamma_2|}{[\log\log(n-k)]^{1/2}}\right\} = O_P(1). \qquad (16)$$

Therefore, by C.3 and (16)

$$\begin{aligned} &\max_{1\le i\le n}\{\max_{p\le k\le n-p}|x'_i(X'_{2k}X_{2k})^{-1}X'_{2k}\gamma_2 I(i\le k)|\} \\ = \ & O_P(n^{1/(2+\delta)})\max_{1\le i\le n}\{\max_{p\le k\le n-p}(n-k)^{-1/2} \\ & [\log\log(n-k_n)]^{1/2}\} \\ = \ & O_P(n^{1/(2+\delta)}). \end{aligned} \qquad (17)$$

Similarly,

$$\begin{aligned} &\max_{1\le i\le n}\{\max_{p\le k\le n-p}|x'_i(X'_{1k}X_{1k})^{-1}X'_{1k}\gamma_1 I(i > k)|\} \\ = \ & O_P(n^{1/(2+\delta)}). \end{aligned} \qquad (18)$$

The lemma follows by (14),(15),(17),(18).

Lemma 2 . Assume that H_0 and C.1-C.3 hold. Then

(a) $\max_{p\le k\le n-p}|\bar{e}(k)| = O_P(n^{-1/2}\log\log^{1/2} n)$.

(b) $\max_{p\le k\le n-p}s^2(k) = O_P(1)$ and in probability,

$$\liminf_{n\to\infty}\max_{p\le k\le n-p}s^2(k) \ge \sigma^2 > 0.$$

Furthermore, if $k_n\to\infty$ as $n\to\infty$, we have

$$\max_{k_n\le k\le n-k_n}|s^2(k) - \sigma^2| = O_P(1).$$

Proof.

Let $\bar{\epsilon} = (1/n)\sum\epsilon_i$, $\gamma_1 = (I_k, 0)\epsilon$ and $\gamma_2 = (0, I_{n-k})\epsilon$, Under H_0,

$$\begin{aligned}\bar{e} \quad = \quad & \bar{\epsilon} - \tfrac{1}{n}(\textstyle\sum_{i=1}^{k} x_i)'(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2 \\ & +\tfrac{1}{n}(\textstyle\sum_{j=k+1}^{n} x_j)'(X_{1k}'X_{1k})^{-1}X_{1k}'\gamma_1\}. \end{aligned} \tag{19}$$

By C.2 and the law of iterated logarithm,

$$\max_{p\le k\le n-p}\left\{\frac{|(\sum_{i=1}^{k} x_i)'(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2|}{[(n-k)\log\log(n-k)]^{1/2}}\right\} = O_P(1).$$

Thus,

$$\begin{aligned} & \frac{1}{n}\max_{p\le k\le n-p}\left|(\sum_{i=1}^{k} x_i)'(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2\right| \\ = \; & O_P(n^{-1/2}(\log\log n)^{1/2}). \end{aligned} \tag{20}$$

Similarly,

$$\begin{aligned} & \frac{1}{n}\max_{p\le k\le n-p}\left|(\sum_{i=k+1}^{n} x_i)'(X_{1k}'X_{1k})^{-1}X_{1k}'\gamma_1\right| \\ = \; & O_P(n^{-1/2}(\log\log n)^{1/2}) \end{aligned} \tag{21}$$

Combining (19),(20),(21), we have

$$\max_{p\le k\le n-p}|\bar{e}(k)| = O_P(n^{-1/2}(\log\log n)^{1/2}).$$

The part (a) is proved.
Next consider $s^2(k)$. We may write

$$\begin{aligned} s^2(k) \quad = \quad & \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\epsilon_i^2 + \tfrac{1}{n}\gamma_1'X_{1k}(X_{1k}'X_{1k})^{-1}X_{2k}'X_{2k} \\ & (X_{1k}'X_{1k})^{-1}X_{1k}'\gamma_1 + \tfrac{1}{n}\gamma_2'X_{2k}(X_{1k}'X_{1k})^{-1} \\ & X_{1k}'X_{1k}(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2 - \tfrac{2}{n}\gamma_2'X_{2k}(X_{1k}' \\ & X_{1k})^{-1}X_{1k}'\gamma_1 - \tfrac{2}{n}\gamma_1'X_{1k}(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2 \end{aligned} \tag{22}$$

By C.2 and the law of iterated logarithm, we have

$$\begin{aligned} & \max_{p\le k\le n-p} n^{-1}|\gamma_1'X_{1k}(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2| \\ = \; & \max_{p\le k\le n-p} O_P\left\{\frac{[k(\log\log k)\log\log(n-k)]^{1/2}}{n(n-k)^{1/2}}\right\} \\ = \; & O_P(n^{-1/2}(\log\log n)^{1/2}). \end{aligned} \tag{23}$$

Similarly,

$$\begin{aligned} & \max_{p\le k\le n-p} n^{-1}|\gamma_2'X_{2k}(X_{1k}'X_{1k})^{-1}X_{1k}'\gamma_1| \\ = \; & O_P(n^{-1/2}(\log\log n)^{1/2}). \end{aligned} \tag{24}$$

By C.2 and the the law of iterated logarithm again,

$$\begin{aligned} & \max_{p\le k\le n-p}\frac{1}{n}|X_{2k}(X_{1k}'X_{1k})^{-1}X_{1k}'\gamma_1|^2 \\ = \; & \max_{d\le k\le n-d}\frac{\log\log^{1/2}k}{k^{1/2}}O_P^2(1) = O_P^2(1), \end{aligned} \tag{25}$$

$$\begin{aligned} & \max_{p\le k\le n-p}\frac{1}{n}|X_{1k}(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2|^2 \\ = \; & \max_{d\le k\le n-d}\frac{\log\log^{1/2}(n-k)}{(n-k)^{1/2}}O_P^2(1) = O_P^2(1). \end{aligned} \tag{26}$$

It is clear that for $k_n \to \infty$,

$$\max_{k_n\le k\le n-k_n}[(\log\log k)/k]^{1/2}) = o_P(1),$$

$$\max_{k_n\le k\le n-k_n}\frac{\log\log^{1/2}(n-k)}{(n-k)^{1/2}} = o_P(1),$$

so that (25)and (26) become

$$\max_{k_n\le k\le n-k_n}\frac{1}{n}|X_{2k}(X_{1k}'X_{1k})^{-1}X_{1k}'\gamma_1|^2 = O_P(1) \tag{27}$$

$$\max_{k_n\le k\le n-k_n}\frac{1}{n}|X_{1k}(X_{2k}'X_{2k})^{-1}X_{2k}'\gamma_2|^2 = O_P(1). \tag{28}$$

The part (b) follows from(22-24),(27),(28). The proof is complete.

Lemma 3. Assume that H_0 and C.1-C.3 hold. Then for some $\tau > 0$,

$$\max_{p\le k\le n-p}|\hat{\lambda}(k) - \bar{e}(k)/s^2(k)| = O_P(n^{-1/2-\tau}).$$

Proof. Since $\hat{\lambda}(k)$ solves (8), similarly to Owen (2001), we consider

$$\begin{aligned} 0 = & \frac{1}{n}\left|\sum_{i=1}^{n}\frac{\tilde{e}_i(k)}{1+\hat{\lambda}(k)\tilde{e}_i(k)}\right| \\ = & \left|\bar{e}(k) - \tfrac{1}{n}\hat{\lambda}(k)\textstyle\sum_{i=1}^{n}\tfrac{\tilde{e}_i(k)^2}{1+\hat{\lambda}(k)\tilde{e}_i(k)}\right| \\ \ge & \quad |\hat{\lambda}(k)|\tfrac{s^2(k)}{1+|\hat{\lambda}(k)|\max|\tilde{e}_i(k)|} - |\bar{e}(k)|. \end{aligned}$$

So by Lemma 1, for some $\delta > 0$,

$$\frac{|\hat{\lambda}(k)|}{1+|\hat{\lambda}(k)|o(n^{1/(2+\delta)})} \le |\bar{e}(k)|/s^2(k). \tag{29}$$

By Lemma 2,

$$\max_{p\le k\le n-p}\{|\bar{e}(k)|/s^2(k)\} = O_P(n^{-1/2}(\log\log n)^{1/2}) \tag{30}$$

Therefore by (29),(30),

$$\max_{p\le k\le n-p}|\hat{\lambda}(k)| = O_P(n^{-1/2}(\log\log n)^{1/2}). \tag{31}$$

Now let $\eta_i(k) = \hat{\lambda}(k)\tilde{e}_i(k)$. By (31) and Lemma 1, it follows that $\max_{1\le i\le n}\{\max_{p\le k\le n-p}|\eta_i(k)|\} = O_P(1)$.

Using Taylor expansion,

$$\begin{aligned} 0 \quad = \quad & \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{e}_i(k)}{1+\eta_i(k)} \\ = \quad & \bar{e}(k) - s^2(k)\hat{\lambda}(k) + \frac{\hat{\lambda}^2(k)}{n}\sum_{i=1}^{n}\frac{\tilde{e}_i^3}{(1+\xi_i(k))^3} \end{aligned} \tag{32}$$

where $|\xi_i(k)| \le |\eta_i(k)| = |\hat{\lambda}(k)\tilde{e}_i(k)|$, By Lemma 1 and (31)

$$\begin{aligned} & \max\nolimits_{1\le i\le n}\max\nolimits_{p\le k\le n-p}|\xi_i(k)| \\ = \quad & O_P(n^{1/(2+\delta)})O_P(n^{-1/2}(\log\log n)^{1/2}) \\ = \quad & O_P(1). \end{aligned} \tag{33}$$

Therefore, by (31), Lemmas 1 and 2, we conclude that uniformly in *k*,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{\lambda}^2(k)\tilde{e}_i^3(k)}{(1+\xi_i(k))^3}\right| \quad \le \quad & \hat{\lambda}^2(k)s^2(k)O_P(1)\max_{1\le i\le n}|\tilde{e}_i(k)| \\ = \quad & O_P(n^{-(1+\delta)/(2+\delta)}\log\log n) \\ = \quad & O_P(n^{-1/2}). \end{aligned} \tag{34}$$

The lemma follows from Lemma 2, (32) and (34), with any $0 < \tau < \delta/[2(2+\delta)]$.

Proof of Theorem 1. First, we use Lemmas 1, 2 and 3 to obtain a quadratic approximation to $-2\log R(k)$, uniformly in k. Following Owen (2001, page 221)'s arguments, denote $z_i = \hat{\lambda}(k)\tilde{e}_i(k)$. Using Taylor's expansion,

$$\begin{aligned} -2\log R(k) &= 2\sum_{i=1}^{n}\log(1+z_i) \\ &= 2\sum_{i=1}^{n}\left\{z_i - \frac{1}{2}z_i^2 + \frac{1}{3}\frac{z_i^3}{(1+\xi_i)^3}\right\}, \quad (35) \end{aligned}$$

where $|\xi_i| \le |z_i| = |\hat{\lambda}(k)\tilde{e}_i(k)| = O_P(1),$ uniformly in k as argued in (33). By Lemmas 1 and 3, for some $\delta > 0$,

$$\begin{aligned} &\max_{p\le k\le n-p}\sum_{i=1}^{n}\left|\frac{z_i^3}{(1+\xi_i)^3}\right| \\ \le\ & n\left\{\max_{p\le k\le n-p}[|\hat{\lambda}^3(k)|s^2(k)]\right\}\max_{1\le i\le n}|\tilde{e}_i(k)| \\ =\ & O_P\left\{n^{-\frac{3}{2}+1+\frac{1}{2+\delta}}\log\log^{3/2} n\right\} \\ =\ & O_P(n^{-\delta/(4+2\delta)}\log\log^{3/2} n). \qquad (36) \end{aligned}$$

Next by Lemma 3, for some $\tau > 0$,

$$\begin{aligned} 2\sum_{i=1}^{n} z_i &= 2n\bar{e}^2(k)/s^2(k) + n\bar{e}(k)o(n^{-1/2-\tau}) \\ &= 2n\bar{e}^2(k)/s^2(k) + O_P(n^{-\tau}\log\log^{1/2} n), \qquad (37) \end{aligned}$$

and

$$\begin{aligned} \sum_{i=1}^{n} z_i^2 &= n\bar{e}^2(k)/s^2(k) + nO_P(n^{-1/2-\tau})\bar{e}(k) \\ &= n\bar{e}^2(k)/s^2(k) + O_P(n^{-\tau}\log\log^{1/2} n), \qquad (38) \end{aligned}$$

Combining (35),(36),(37),(38) yields that for any

$$0 < \tau_1 < \min\{\delta/(4+2\delta), \tau\},$$

$$\max_{p\le k\le n-p}\left|-2\log R(k) - n\frac{\bar{e}^2(k)}{s^2(k)}\right| = O_P(n^{-\tau_1}). \quad (39)$$

Now applying Taylor expansion

$$(a+x)^{1/2} = a^{1/2} + x/(2a^{1/2}) + o(x/a^{1/2}),$$

we have for any $0 < \tau_2 < \tau_1$,

$$\begin{aligned} M_n^{1/2} &= \{\max_{d\le k\le n-d}[-\log R(k)]\}^{1/2} \\ &= \max_{p\le k\le n-p} n^{1/2}\{|\bar{e}(k)|/s(k)\} + O_P(n^{-\tau_2}). \qquad (40) \end{aligned}$$

Denote $M_n^{1/2} = T_n^{1/2} + O_P(n^{-\tau_2})$, i.e.,

$$T_n^{1/2} = \max_{p\le k\le n-p} n^{1/2}\{|\bar{e}(k)|/s(k)\}$$

Using the same arguments to the proof of Theorem 3.1.2 of Csörgő and Horváth(1997), we have

$$\lim_{n\to\infty} P\{A(\log n)T_n^{1/2} \le t + D_d(\log n)\} = \exp(-2e^{-t})$$

for all *t* Since $A(\log n)O_P(n^{-\tau_2}) = o(1)$, it follows from (40) that

$$\begin{aligned} & A(\log n)M_n^{1/2} - D_d(\log n) \\ =\ & A(\log n)T_n^{1/2} - D_d(\log n) + O_P(1). \end{aligned}$$

The proof is complete.