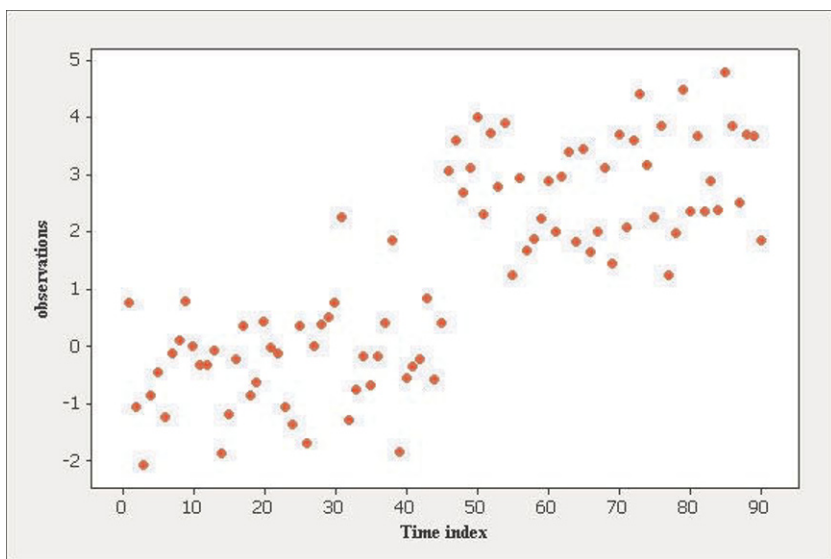


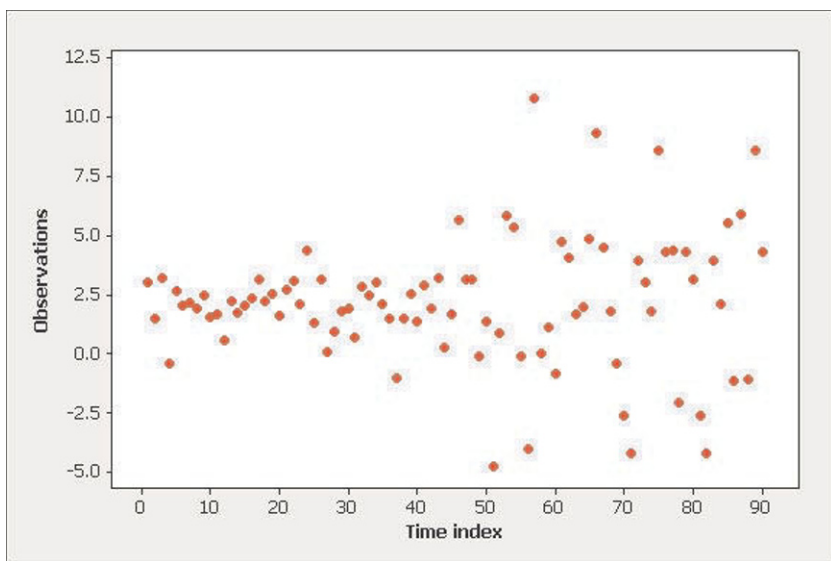
## Chapter 2

# Univariate Normal Model

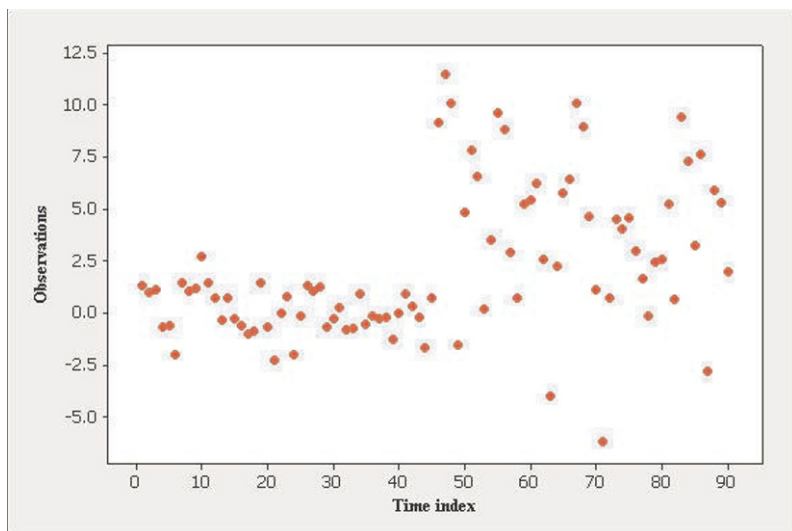
Let  $x_1, x_2, \dots, x_n$  be a sequence of independent normal random variables with parameters  $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots, (\mu_n, \sigma_n^2)$ , respectively. In this chapter, different types of change point problems with regard to the mean, variance, and mean and variance are discussed. For simplicity and illustration purposes, we familiarize readers with the three types of changes in the normal sequence by presenting the following three figures, where Figure 2.1 represents a sequence of normal observations with a mean change, Figure 2.2 shows a variance change in the normal observations, and Figure 2.3 indicates a mean and variance change in the sequence of normal observations.



**Fig. 2.1** A change in the mean of the sequence of normal observations



**Fig. 2.2** A change in the variance of the sequence of normal observations



**Fig. 2.3** A change in both the mean and variance of the sequence of normal observations

## 2.1 Mean Change

Suppose that each  $x_i$  is normally distributed with mean  $\mu_i$  and common variance  $\sigma^2$ ,  $i = 1, 2, \dots, n$ . The interest here is about the mean change. This problem was first examined by Page (1954, 1955, 1957). Later, Chernoff and

Zacks (1964), Bhattacharya and Johnson (1968), Gardner (1969), Sen and Srivastava (1975a,b), Gupta and Chen (1996), Chen and Gupta (1997), and Chen and Gupta (1998, 2003) also contributed to the study of this problem. Throughout this section, the hypothesis of stability (the null hypothesis) is defined as

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n = \mu.$$

The mean change problem in this one-dimensional case can be one-sided or two-sided. Only the two-sided test is addressed here. That is, the interest is to test  $H_0$  versus

$$H_1 : \mu_1 = \cdots = \mu_k \neq \mu_{k+1} \cdots = \mu_n,$$

where  $k$  is the unknown location of the single change point. The testing procedure depends on whether the nuisance parameter  $\sigma^2$  is known or unknown.

### 2.1.1 Variance Known

Without loss of generality, assume that  $\sigma^2 = 1$ . Under  $H_0$ , the likelihood function is

$$L_0(\mu) = \frac{1}{(\sqrt{2\pi})^n} e^{-\sum_{i=1}^n (x_i - \mu)^2 / 2}$$

and the maximum likelihood estimator (MLE) of  $\mu$  is

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Under  $H_1$ , the likelihood function is

$$L_1(\mu_1, \mu_n) = \frac{1}{(\sqrt{2\pi})^n} e^{-(\sum_{i=1}^k (x_i - \mu_1)^2 + \sum_{i=k+1}^n (x_i - \mu_n)^2) / 2},$$

and the MLEs of  $\mu_1$ , and  $\mu_n$  are, respectively,

$$\hat{\mu}_1 = \bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i, \quad \text{and} \quad \hat{\mu}_n = \bar{x}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n x_i.$$

Let

$$S_k = \sum_{i=1}^k (x_i - \bar{x}_k)^2 + \sum_{i=k+1}^n (x_i - \bar{x}_{n-k})^2,$$

$$V_k = k(\bar{x}_k - \bar{x})^2 + (n - k)(\bar{x}_{n-k} - \bar{x})^2,$$

and  $S = \sum_{i=1}^n (x_i - \bar{x})^2$ ; then  $V_k = S - S_k$ . Simple algebra leads to  $U^2 = V_{k*} = \max_{1 \leq k \leq n-1} V_k$  is the likelihood procedure (see Lehmann, 1986) test statistic for testing  $H_0$  against  $H_1$ .

Hawkins (1977) derived the exact and asymptotic null distributions of the test statistic  $U$ . The following is based on his work. First, simple algebraic computation gives an alternative expression for  $V_k$  as

$$V_k = \frac{n}{k(n-k)} \left[ \sum_{i=1}^k (x_i - \bar{x}) \right]^2.$$

Let

$$T_k = \sqrt{\frac{n}{k(n-k)}} \left[ \sum_{i=1}^k (x_i - \bar{x}) \right];$$

then  $V_k = T_k^2$  or  $|T_k| = \sqrt{V_k}$ . Therefore,

$$U = \sqrt{V_{k*}} = \max_{1 \leq k \leq n-1} \sqrt{V_k} = \max_{1 \leq k \leq n-1} |T_k| \quad (2.1)$$

is the equivalent likelihood-based test statistic for testing  $H_0$  against  $H_1$ . After this computational preparation, the main theorem can be stated as follows.

**Theorem 2.1** *The null probability density function of  $U$  is given by*

$$f_U(x) = 2\phi(x, 0, 1) \sum_{k=1}^{n-1} g_k(x, x) g_{n-k}(x, x),$$

where  $\phi(x, 0, 1)$  is the pdf of  $N(0, 1)$ ,  $g_1(x, s) = 1$  for  $x, s \geq 0$ , and

$$g_k(x, s) = P[|T_i| < s, i = 1, \dots, k-1 | |T_k| = x], \quad (2.2)$$

for  $x, s \geq 0$ .

To prove this theorem, the following lemma is essential.

**Lemma 2.2**  $\{T_1, T_2, \dots, T_{n-1}\}$  is a Markov process.

*Proof.* From the Markov process property, it suffices to show that the partial covariance  $\sigma_{jk.m}$  between  $T_j$  and  $T_k$  when  $T_m$  is fixed equals zero for  $j < m < k$ , or equivalently the partial correlation coefficient  $\rho_{jk.m}$  is zero.

For  $m = 1, 2, \dots, n-1$ , and  $m < k$ , the correlation coefficient  $\rho_{mk}$  between  $T_k$  and  $T_m$  is

$$\begin{aligned}\rho_{mk} &= \frac{n}{\sqrt{km(n-k)(n-m)}} E \left[ \sum_{i=1}^k (x_i - \bar{x}) \sum_{j=1}^m (x_j - \bar{x}) \right] \\ &= \frac{n}{\sqrt{km(n-k)(n-m)}} \frac{m(n-k)}{n} \\ &= \sqrt{\frac{m(n-k)}{k(n-m)}}.\end{aligned}$$

Then, for  $j < m < k$ ,

$$\begin{aligned}\rho_{jk.m} &= \frac{\rho_{jk} - \rho_{jm}\rho_{mk}}{\sqrt{(1 - \rho_{jm}^2)(1 - \rho_{mk}^2)}} \\ &= \frac{\sqrt{\frac{j(n-k)}{k(n-j)}} - \sqrt{\frac{j(n-m)}{m(n-j)} \frac{m(n-k)}{k(n-m)}}}{\sqrt{\left[1 - \frac{j(n-m)}{m(n-j)}\right] \left[1 - \frac{m(n-k)}{k(n-m)}\right]}} \\ &= 0.\end{aligned}$$

This completes the proof of the above lemma. □

Now, it is time to prove the theorem.

*Proof of Theorem 2.1* Let  $A$ ,  $B$ ,  $C$ , be the following events.

$$\begin{aligned}A &= \{|T_k| \in (x, x + dx)\}, \\ B &= \{|T_j| < |T_k|, j = 1, \dots, k-1\}, \\ C &= \{|T_j| < |T_k|, j = k+1, \dots, n-1\}.\end{aligned}$$

$U = \max_{1 \leq k \leq n-1} |T_k|$ ; then

$$\begin{aligned}F_U(x + dx) - F_U(x) &= P[U \in (x, x + dx)] \\ &= P \left\{ \bigcup_{k=1}^{n-1} [|T_k| \in (x, x + dx)] \cap [|T_k| > |T_j|, j \neq k] \right\} \\ &= \sum_{k=1}^{n-1} P[ABC] \\ &= \sum_{k=1}^{n-1} P[A]P[B|A]P[C|AB].\end{aligned}$$

Next,  $T_k \sim N(0, 1)$ , therefore  $P[A] = 2\phi(x, 0, 1) + o(dx)$ . Moreover,

$$\begin{aligned} P[B|A] &= P[|T_j| < x, i = 1, \dots, k-1 | |T_k| = x] + o(dx) \\ &= g_k(x, x) + o(dx). \end{aligned}$$

Finally, from the fact that  $\{T_1, T_2, \dots, T_{n-1}\}$  is a Markovian,  $\{T_1, T_2, \dots, T_{k-1}\}$  and  $\{T_{k+1}, T_2, \dots, T_{n-1}\}$  are independent; that is,  $B$  and  $C$  are independent given  $T_k = x$ . Therefore,  $P[C|AB] = P[C|A]$ . According to the probability symmetry between  $B$  and  $C$ , similar to  $P[B|A]$ , we have

$$P[C|A] = g_{n-k}(x, x) + o(dx).$$

Thus, we obtain

$$P[U \in (x, x + dx)] = \sum_{k=1}^{n-1} 2\phi(x, 0, 1)g_k(x, x)g_{n-k}(x, x) + o(dx),$$

or

$$f_U(x) = 2\phi(x, 0, 1) \sum_{k=1}^{n-1} g_k(x, x)g_{n-k}(x, x).$$

This completes the proof of the theorem.  $\square$

To be able to use the null distribution of  $U$ , one needs to know how to evaluate  $g_k(x, s)$  or  $g_{n-k}(x, s)$ . The following theorem is given just for this purpose.

**Theorem 2.3** *The function  $g_k(x, s)$  is determined by the recursion:*

$$g_k(x, s) = \int_0^s g_{k-1}(y, s)[\phi(y, \rho x, \tau^2) + \phi(y, -\rho x, \tau^2)]dy,$$

where  $\rho = \rho_{k-1, k}$  is the correlation coefficient between  $T_{k-1}$  and  $T_k$ , and  $\tau^2 = \sqrt{1 - \rho^2}$ .

*Proof.* From (2.2) and the facts that  $\{T_1, T_2, \dots, T_{n-1}\}$  is a Markovian,  $T_k \sim N(0, 1)$ , the symmetry of  $T_k$  about 0, and  $T_{k-1}|T_k = x \sim N(\rho x, \tau^2)$ , we have:

$$\begin{aligned} g_k(x, s) &= P[|T_j| < s, j = 1, \dots, k-1 | |T_k| = x] \\ &= \int_{-s}^s P[|T_j| < s, j = 1, \dots, k-2 | T_{k-1} = y \quad \text{and} \quad |T_k| = x] \\ &\quad d[T_{k-1} < y | T_k = x] \\ &= \int_{-s}^s P[|T_j| < s, j = 1, \dots, k-2 | T_{k-1} = y] d[T_{k-1} < y | T_k = x] \end{aligned}$$

$$\begin{aligned}
&= \int_{-s}^s P[|T_j| < s, j = 1, \dots, k-2 | |T_{k-1}| = y] d[T_{k-1} < y | T_k = x] \\
&= \int_{-s}^s g_{k-1}(|y|, s) d[T_{k-1} < y | T_k = x] \\
&= \int_{-s}^s g_{k-1}(|y|, s) \phi(y, \rho x, \tau) dy \\
&= \int_0^s g_{k-1}(y, s) \phi(y, \rho x, \tau) dy + \int_{-s}^0 g_{k-1}(-y, s) \phi(y, \rho x, \tau) dy \\
&= \int_0^s g_{k-1}(y, s) [\phi(y, \rho x, \tau^2) + \phi(y, -\rho x, \tau^2)] dy. \quad \square
\end{aligned}$$

In addition to the null distribution of  $U$ , the distribution of the location  $k^*$  of the change point has also been derived, which is given in the following theorem.

**Theorem 2.4** *If  $k^*$  is the position of the change point estimated by (2.1), then for  $k = 1, 2, \dots, n$ ,*

$$P[k = k^*] = \int_0^\infty g_k(x, x) g_{n-k}(x, x) \phi(x, 0, 1) dx.$$

*Proof.* In view of the facts that  $\{T_1, T_2, \dots, T_{n-1}\}$  is a Markovian,  $T_k \sim N(0, 1)$ , and the symmetry of  $T_k$  about 0, we obtain:

$$\begin{aligned}
P[k = k^*] &= P \left\{ \sqrt{V_{k*}} = \sqrt{V_k} = \max_{1 \leq k \leq n-1} |T_k| \right\} \\
&= P[|T_j| < |T_k|, j \neq k] \\
&= \int_0^\infty P[|T_j| < x, j \neq k | T_k = x] dP[T_k < x] \\
&= \int_0^\infty P[|T_j| < x, j = 1, \dots, k-1 | T_k = x] dP[T_k < x] \\
&\quad P[|T_j| < x, j = k+1, \dots, n | T_k = x] \\
&= \int_0^\infty P[|T_j| < x, j = 1, \dots, k-1 | |T_k| = x] dP[T_k < x] \\
&\quad P[|T_j| < x, j = k+1, \dots, n | |T_k| = x] \\
&= \int_0^\infty g_k(x, x) g_{n-k}(x, x) dP[T_k < x] \\
&= \int_0^\infty g_k(x, x) g_{n-k}(x, x) \phi(x, 0, 1) dx. \quad \square
\end{aligned}$$

Although the null distribution of the test statistic  $U$  has been obtained, the recursion formula requires moderate computations. Yao and Davis (1986) derived the asymptotic null distribution of  $U$ , which provides an alternative way to do formal statistical analysis when  $n$  is sufficiently large. The following is based on their work.

Let  $W_k = x_1 + x_2 + \cdots + x_k$ ,  $1 \leq k \leq n$ ; then simple algebra leads to

$$U = \max_{1 \leq k \leq n-1} \left| \frac{W_k}{\sqrt{n}} - \frac{k}{n} \frac{W_n}{\sqrt{n}} \right| \left/ \left[ \frac{k}{n} \left( 1 - \frac{k}{n} \right) \right]^{1/2} \right|.$$

Suppose  $\{B(t); 0 \leq t < \infty\}$  is a standard Brownian motion; then under  $H_0$ , from properties of the normal random variable,

$$\{(W_k - k\mu)/\sqrt{n}; 1 \leq k \leq n\} \stackrel{D}{=} \left\{ B\left(\frac{k}{n}\right); 1 \leq k \leq n \right\},$$

where “ $\stackrel{D}{=}$ ” means “distributed as”. Furthermore,

$$\begin{aligned} U &= \max_{1 \leq k \leq n-1} \left| \frac{W_k}{\sqrt{n}} - \frac{k}{n} \frac{W_n}{\sqrt{n}} \right| \left/ \left[ \frac{k}{n} \left( 1 - \frac{k}{n} \right) \right]^{1/2} \right| \\ &= \max_{nt=1, \dots, n-1} \left| \frac{W_k}{\sqrt{n}} - t \frac{W_n}{\sqrt{n}} \right| \left/ [t(1-t)]^{1/2} \right| \\ &= \max_{nt=1, \dots, n-1} \left| \frac{W_k}{\sqrt{n}} - \frac{k\mu}{\sqrt{n}} - t \left( \frac{W_n}{\sqrt{n}} - \frac{n\mu}{\sqrt{n}} \right) \right| \left/ [t(1-t)]^{1/2} \right| \\ &\stackrel{D}{=} \max_{nt=1, \dots, n-1} |B(t) - tB(1)|/[t(1-t)]^{1/2} \\ &= \max_{nt=1, \dots, n-1} |B_0(t)|/[t(1-t)]^{1/2}, \end{aligned}$$

where  $t = k/n$ , and  $B_0(t) = B(t) - tB(1)$  is the Brownian bridge. The following theorem shows that the asymptotic null distribution of  $U$  is a Gumbel distribution.

**Theorem 2.5** *Under  $H_0$ , for  $-\infty < x < \infty$ ,*

$$\lim_{n \rightarrow \infty} P[a_n^{-1}(U - b_n) \leq x] = \exp\{-2\pi^{1/2}e^{-x}\},$$

where  $a_n = (2 \log \log n)^{-1/2}$ ,  $b_n = a_n^{-1} + \frac{1}{2}a_n \log \log \log n$ .

The proof of this theorem is mainly based on the properties of Brownian motion and convergence rules from the theory of probability. The following lemmas are needed before the proof of the theorem is given.



**Lemma 2.6**

$$\max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} = o_p(a_n).$$

*Proof.* For large  $n$  and  $0 < t \leq 1/\log n$ , clearly,  $\sqrt{t/(1-t)} \leq 2\sqrt{t}$ , and  $(1/\sqrt{1-t} - 1)/\sqrt{t} \leq \sqrt{t}$ . We obtain:

$$\begin{aligned} \left| \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \frac{|B(t)|}{\sqrt{t}} \right| &\leq \left| \frac{B_0(t)}{\sqrt{t(1-t)}} - \frac{|B(t)|}{\sqrt{t}} \right| \\ &= \left| \frac{B(t) - tB(1)}{\sqrt{t(1-t)}} - \frac{|B(t)|}{\sqrt{t}} \right| \\ &= \left| \frac{|B(t)|}{\sqrt{t}} \left( \frac{1}{\sqrt{1-t}} - 1 \right) - \sqrt{\frac{t}{1-t}} B(1) \right| \\ &\leq \left| \frac{|B(t)|}{\sqrt{t}} \left( \frac{1}{\sqrt{1-t}} - 1 \right) \right| + \left| \sqrt{\frac{t}{1-t}} B(1) \right| \\ &\leq \sqrt{t}|B(t)| + 2\sqrt{t}|B(1)| \\ &\leq (\log n)^{-1/2}(|B(t)| + 2|B(1)|). \end{aligned}$$

Then

$$\begin{aligned} &\max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} \\ &\leq \max_{1 \leq nt \leq [n/\log n]} \left| \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \frac{|B(t)|}{\sqrt{t}} \right| \\ &\leq (\log n)^{-1/2} \max_{1 \leq nt \leq [n/\log n]} (|B(t)| + 2|B(1)|) \\ &= O_p((\log n)^{-1/2}) \\ &= o_p(a_n). \end{aligned} \quad \square$$

**Lemma 2.7**

$$\max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} = O_p((\log \log \log n)^{1/2}).$$

*Proof.* Inasmuch as  $B(t)$  is distributed as  $N(0, t)$ , from the law of iterated logarithm, for large  $n$ ,

$$\frac{nB(t)}{2[nt \log \log (nt)]^{1/2}} = O_p(1).$$

Then

$$\frac{B(t)}{2[t \log \log(nt)]^{1/2}} = O_p(1),$$

because for large  $n$ ,  $n > \sqrt{n}$ . Let  $t$  be small, say  $t = 1/\sqrt{n}$ ; then, for large  $n$ ,

$$\frac{B(t)}{2\sqrt{t \log \log(t^{-1})}} = O_p(1),$$

or with probability 1,

$$|B(t)| < 2\sqrt{t \log \log(t^{-1})}.$$

Let  $t \rightarrow 0^+$ ; then for  $t \in [s, \frac{1}{2}]$ ,  $\log \log(t^{-1}) < \log \log(s^{-1})$ , and

$$\max_{t \in [s, 1/2]} \frac{|B(t)|}{\sqrt{t}} = O_p((\log \log(s^{-1}))^{1/2}).$$

Finally,

$$\begin{aligned} & \max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \\ &= \max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B(t) - tB(1)|}{\sqrt{t(1-t)}} \\ &\leq \max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B(t)|}{\sqrt{t(1-t)}} + \max_{[n/\log n] \leq nt \leq [n/2]} \sqrt{\frac{t}{1-t}} |B(1)| \\ &\leq \max_{[n/\log n] \leq nt \leq [n/2]} \frac{2|B(t)|}{\sqrt{t}} + \max_{[n/\log n] \leq nt \leq [n/2]} \sqrt{\frac{t}{1-t}} |B(1)| \\ &\leq 2 \max_{[1/\log n] \leq t \leq [1/2]} \frac{|B(t)|}{\sqrt{t}} + O_p(1) \\ &= O_p\left(\log \log\left(\frac{1}{\log n}\right)^{-1}\right)^{1/2} + O_p(1) \\ &= O_p((\log \log \log n)^{1/2}). \end{aligned} \quad \square$$

**Lemma 2.8** For  $-\infty < x < \infty$ ,

$$\lim_{n \rightarrow \infty} P \left[ a_n^{-1} \left( \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} - b_n \right) \leq x \right] = \exp\{-\pi^{1/2} e^{-x}\}.$$

*Proof.* Because  $\{(|B(t)|/\sqrt{t}); t = 1/n, \dots, [n/\log n]/n\} \stackrel{D}{=} \{|B(t)|/\sqrt{t}; t = 1, \dots, [n/\log n]\}$ , and from Theorem 2 of Darling and Erdős (1956),

$$\lim_{n \rightarrow \infty} P \left[ a_{[n/\log n]}^{-1} \left( \max_{t=1, \dots, [n/\log n]} \frac{|B(t)|}{\sqrt{t}} - b_{[n/\log n]} \right) \leq x \right] = \exp\{-\pi^{1/2} e^{-x}\} \quad (2.3)$$

or

$$\lim_{n \rightarrow \infty} P \left[ \max_{t=1, \dots, [n/\log n]} \frac{|B(t)|}{\sqrt{t}} \leq a_{[n/\log n]} x + b_{[n/\log n]} \right] = \exp\{-\pi^{1/2} e^{-x}\}.$$

Now, from L'Hospital's rule, one can show that

$$a_{[n/\log n]} = a_n + o(a_n) \quad \text{and} \quad b_{[n/\log n]} = b_n + o(b_n).$$

Hence, for  $-\infty < x < \infty$ ,

$$(2.3) = \lim_{n \rightarrow \infty} P \left[ a_n^{-1} \left( \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} - b_n \right) \leq x \right] = \exp\{-\pi^{1/2} e^{-x}\}.$$

□

**Lemma 2.9** *The following hold as  $n \rightarrow \infty$ ,*

(i)

$$\max_{1 \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} = o_p(a_n).$$

(ii)

$$\max_{1 \leq n(1-t) \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq n(1-t) \leq [n/\log n]} \frac{|B(t) - B(1)|}{\sqrt{t}} = o_p(a_n).$$

*Proof.* (i) From Lemma 2.6,

$$\max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} = o_p(a_n),$$

or

$$a_n^{-1} \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - a_n^{-1} \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} = o_p(1). \quad (2.4)$$

Apply it to Lemma 2.8; then we have:

$$\lim_{n \rightarrow \infty} P \left[ a_n^{-1} \left( \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - b_n \right) \leq x \right] = \exp\{-\pi^{1/2} e^{-x}\},$$

or

$$\lim_{n \rightarrow \infty} P \left[ a_n \left( \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \right) \leq a_n b_n + a_n^2 x \right] = \exp\{-\pi^{1/2} e^{-x}\}.$$

Letting  $n \rightarrow \infty$ , and then  $x \rightarrow +\infty$ , we have

$$\lim_{n \rightarrow \infty} P \left[ a_n \left( \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \right) \rightarrow 1 \right] \rightarrow 1;$$

that is,

$$(2 \log \log n)^{-1/2} \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \xrightarrow{n \rightarrow \infty} 1, \quad (2.5)$$

in probability. From Lemma 2.7,

$$\max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} = O_p((\log \log \log n)^{1/2}).$$

Then

$$a_n \max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} = o_p(1).$$

Combining the above with (2.5), as  $n \rightarrow \infty$ , we obtain

$$P \left[ \max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \geq \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \right] \rightarrow 0;$$

that is,

$$\max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} < \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}}. \quad (2.6)$$

Then, according to (2.6),

$$\begin{aligned} & \max_{1 \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \\ &= \max \left[ \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}}, \max_{[n/\log n] \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \right] \\ &\stackrel{P}{=} \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}}. \end{aligned}$$

Therefore, (2.4) becomes

$$a_n^{-1} \max_{1 \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - a_n^{-1} \max_{1 \leq nt \leq [n/\log n]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} = o_p(1),$$

which leads to the result (i).

(ii) From the symmetry property of Brownian motion with respect to  $t = \frac{1}{2}$ ,  $B_0(t) \stackrel{D}{=} B_0(1-t)$  for  $0 < t < 1$ . Because  $B(1-t) \stackrel{D}{=} B(1) - B(t) \sim N(0, 1-t)$ , then replacing  $t$  by  $1-t$  in (i), we obtain

$$\max_{1 \leq n(1-t) \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq n(1-t) \leq [n/\log n]} \frac{|B(1-t)|}{\sqrt{t}} = o_p(a_n),$$

or

$$\max_{1 \leq n(1-t) \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} - \max_{1 \leq n(1-t) \leq [n/\log n]} \frac{|B(t) - B(1)|}{\sqrt{t}} = o_p(a_n).$$

□

After the above preparation, we are ready to prove the theorem.

*Proof of Theorem 2.5*

$$\begin{aligned} P[a_n^{-1}(U - b_n) \leq x | H_0] &= P[U \leq a_n x + b_n | H_0] \\ &= P \left[ \max_{1 \leq nt \leq n-1} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \leq a_n x + b_n \right] \\ &= P \left[ \max_{1 \leq nt \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \leq a_n x + b_n, \right. \\ &\quad \left. \max_{1 \leq n(1-t) \leq [n/2]} \frac{|B_0(t)|}{\sqrt{t(1-t)}} \leq a_n x + b_n \right] \\ &= P \left[ \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} \leq a_n x + b_n, \right. \\ &\quad \left. \max_{1 \leq n(1-t) \leq [n/\log n]} \frac{|B(t) - B(1)|}{\sqrt{1-t}} \leq a_n x + b_n \right] + o_p(1) \\ &= P \left[ \max_{1 \leq nt \leq [n/\log n]} \frac{|B(t)|}{\sqrt{t}} \leq a_n x + b_n \right] \\ &\quad \cdot P \left[ \max_{1 \leq n(1-t) \leq [n/\log n]} \frac{|B(t) - B(1)|}{\sqrt{1-t}} \leq a_n x + b_n \right] + o_p(1) \\ &\stackrel{n \rightarrow \infty}{\longrightarrow} \exp(-\pi^{-1/2} e^{-x}) \cdot \exp(-\pi^{-1/2} e^{-x}) \\ &= \exp(-2\pi^{-1/2} e^{-x}). \end{aligned}$$

### 2.1.2 Variance Unknown

Under  $H_0$ , the likelihood function now is

$$L_0(\mu, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\sum_{i=1}^n (x_i - \mu)^2 / 2\sigma^2}$$

and the MLEs of  $\mu$  and  $\sigma^2$  are

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

respectively. Under  $H_1$ , the likelihood function is

$$L_1(\mu_1, \mu_n, \sigma_1^2) = \frac{1}{(\sqrt{2\pi})^n} e^{-\sum_{i=1}^k (x_i - \mu_1)^2 / 2\sigma_1^2 - \sum_{i=k+1}^n (x_i - \mu_n)^2 / 2\sigma_1^2},$$

and the MLEs of  $\mu_1$ ,  $\mu_n$ , and  $\sigma_1^2$  are,

$$\hat{\mu}_1 = \bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i, \quad \hat{\mu}_n = \bar{x}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n x_i,$$

and

$$\hat{\sigma}_1^2 = \frac{1}{n} \left[ \sum_{i=1}^k (x_i - \bar{x}_k)^2 + \sum_{i=k+1}^n (x_i - \bar{x}_{n-k})^2 \right],$$

respectively.

Let

$$S = \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{and} \quad T_k^2 = \frac{k(n-k)}{n} (\bar{x}_k - \bar{x}_{n-k})^2.$$

The likelihood procedure-based test statistic is then given by

$$V = \max_{1 \leq k \leq n-1} \frac{|T_k|}{S}. \quad (2.7)$$

Worsley (1979) obtained the null distribution of  $V$ . His result is presented in the following.

Let  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , and  $\mathbf{y} = (y_1, y_2, \dots, y_n)$ , where  $y_i = (x_i - \bar{x})/\sqrt{S}$ ,  $i = 1, 2, \dots, n$ . Also, define

$$c_k = \sqrt{k/(n-k)} \quad \text{and} \quad b_{ki} = \begin{cases} n^{-1/2} c_k^{-1}, & i = 1, \dots, k \\ -n^{-1/2} c_k, & i = k+1, \dots, n \end{cases},$$

for  $k = 1, 2, \dots, n-1$ . Let  $\mathbf{b}_k$  be the vector such that  $\mathbf{b}_k = (b_{k1}, \dots, b_{kn})$ . Then,  $T_k = \mathbf{b}'_k \mathbf{x}$ ,  $k = 1, \dots, n-1$ . It is easy to see that  $\mathbf{b}'_k \mathbf{1} = \mathbf{0}$ , where

$\mathbf{1} = (1, 1, \dots, 1)'$  is the  $n \times 1$  unit vector, hence  $\mathbf{b}'_k \mathbf{y} = \mathbf{b}_k \mathbf{x} / \sqrt{S}$  and  $V = \max_{1 \leq k \leq n-1} |\mathbf{b}'_k \mathbf{y}|$ .

Next, let  $\mathbf{y} \in R^n$ . Under  $H_0$ ,  $\mathbf{y}'\mathbf{y} = 1$ , thus  $\mathbf{y}$  is uniformly distributed on the surface of the unit  $(n-1)$ -ball  $C = \{\mathbf{y} : \mathbf{y}'\mathbf{y} = 1 \text{ and } \mathbf{1}'\mathbf{y} = 0\}$ . Let  $D = \{\mathbf{y} : \mathbf{1}'\mathbf{y} = 0, |\mathbf{b}'_k \mathbf{y}| \leq v, k = 1, \dots, n-1\}$ , then the event  $\{V \leq v\} = \{\mathbf{y} \in C \cap D\}$ . Therefore,

$$\begin{aligned} P\{V \leq v\} &= \text{surface area of } C \text{ inside } D \\ &= P\left\{\mathbf{y} \in \bigcap_{k=1}^{n-1} (A_k^+ \cup A_k^-)^c\right\}, \end{aligned} \quad (2.8)$$

where for  $k = 1, \dots, n-1$ ,

$$\begin{aligned} A_k^+ &= \{\mathbf{y} : \mathbf{y} \in C, \mathbf{b}'_k \mathbf{y} > v\}, \\ A_k^- &= \{\mathbf{y} : \mathbf{y} \in C, -\mathbf{b}'_k \mathbf{y} > v\}, \\ A_k &= A_k^+ \cup A_k^-, A_k^+ \cap A_k^- = \phi. \end{aligned}$$

From DeMorgan's law, (2.8) is reduced to

$$\begin{aligned} P\{V \leq v\} &= 1 - P\left\{\mathbf{y} \in \bigcup_{k=1}^n (A_k^+ \cup A_k^-)\right\} \\ &= 1 - \sum_{k=1}^n P\{\mathbf{y} \in A_k\} + \sum \sum_{1 \leq k_1 \leq k_2 \leq n-1} P\left\{\mathbf{y} \in A_{k_1} \cap A_{k_2}\right\} \\ &\quad + \dots + (-1)^p \sum \sum \dots \sum_{1 \leq k_1 \leq k_2 \leq \dots \leq k_p \leq n-1} P\left\{\mathbf{y} \in \bigcap_{j=1}^p A_{k_j}\right\} \\ &\quad + \dots + (-1)^{n-1} P\left\{\mathbf{y} \in \bigcap_{j=1}^{n-1} A_j\right\}. \end{aligned} \quad (2.9)$$

Because  $P\{\mathbf{y} \in A_k\} = P\{\mathbf{y} \in C, \text{ and } \mathbf{b}'_k \mathbf{y} > v\} + P\{\mathbf{y} \in C, \text{ and } \mathbf{b}'_k \mathbf{y} < -v\}$ , and  $\mathbf{b}'_k \mathbf{y} = \mathbf{b}_k \mathbf{x} / \sqrt{S} = T_k / \sqrt{S}$ , then  $P\{\mathbf{y} \in A_k\}$  can be calculated via the distribution of the statistic  $T_k / \sqrt{S}$ .

Now, under  $H_0$ ,  $T_k \sim N(0, \sigma^2)$ ,  $S_k^2 = \sum_{i=1}^k (x_i - \bar{x}_k)^2 + \sum_{i=k+1}^n (x_i - \bar{x}_{n-k})^2$  is distributed as  $\sigma^2 \chi_{n-2}^2$ , and  $T_k$  is independent of  $S_k$ ; then  $T_k / [S_k / \sqrt{n-2}] \sim \chi_{n-2}^2$ . But simple algebra shows that  $S = S_k^2 + T_k^2$ , then  $T_k / S = (S_k^2 / T_k^2 + 1)^{-1/2}$ , hence  $P\{\mathbf{y} \in A_k\}$  can be calculated via a  $t_{n-2}$  distribution.

For other terms in (2.9), we need only to consider a general one such as  $P\{\mathbf{y} \in \bigcap_{j=1}^p \tilde{A}_{k_j}\}$  for  $1 < p < n-1$ , with  $k_1 < k_2 < \dots < k_p$ , and  $\tilde{A}_{k_j}$  is either  $A_{k_j}^+$  or  $A_{k_j}^-$ .

Let  $B$  be a  $p \times n$  matrix such that  $B = (\mathbf{b}_{k1}^{*'}, \dots, \mathbf{b}_{kp}^{*'})'$ , and  $\mathbf{I}_p$  be a  $p \times 1$  unit vector, where

$$\mathbf{b}_{ki}^{*'} = \begin{cases} b_{kj} & \text{if } \tilde{A}_{kj} = A_{kj}^+ \\ -b_{kj} & \text{if } \tilde{A}_{kj} = A_{kj}^- \end{cases}$$

Then,  $P\{\mathbf{y} \in \bigcap_{j=1}^p \tilde{A}_{kj}\} = P\{B\mathbf{y} > v\mathbf{I}_p\}$ . Now the following theorem gives the null probability density function of  $V = B\mathbf{y}$  at  $v$ .

**Theorem 2.10** *Under  $H_0$ , the pdf of  $V = B\mathbf{y}$  at  $v$  is given by*

$$f_p(v) = \begin{cases} \frac{\Gamma\left[\frac{n-1}{2}\right]}{\pi^{p/2} \Gamma\left[\frac{n-1-p}{2}\right]} |\Sigma|^{-1/2} [1 - v' \Sigma^{-1} v]^{(n-3-p)/2}, & \text{if } v' \Sigma^{-1} v < 1 \\ 0, & \text{otherwise} \end{cases}$$

where  $\Sigma = BB'$ .

To prove this theorem, we need the following results.

**Lemma 2.11**  *$B$  can be written as  $B = G\Gamma$ , where  $G$  is a  $p \times p$  positive definite matrix, and  $\Gamma$  is  $p \times n$  with  $\Gamma\Gamma' = I_p$ .*

*Proof.* It follows directly from Theorem 1.39 on page 11 of Gupta and Varga (1993).  $\square$

For the purpose of deriving the null distribution, WLOG, we write:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_n = 0.$$

Let  $J = I_n - (1/n)\mathbf{1}\mathbf{1}'$ , and augment the matrix  $\Gamma$  in Lemma 2.11 to:

$$Q = \begin{pmatrix} \Gamma \\ \Gamma_0 \end{pmatrix},$$

where  $\Gamma_0$  is  $(n-p) \times n$  such that  $Q$  is an  $n \times n$  orthogonal matrix. Also, let  $M = QJQ'$ , and let the first  $p \times p$  principal minor of  $M$  be  $M_p$ ; then we have the following Lemma 2.12.

**Lemma 2.12**  *$V = B\mathbf{y}$  has the pdf:*

$$f_p(v) = \frac{\Gamma\left[\frac{n-1}{2}\right]}{\pi^{p/2} \Gamma\left[\frac{n-1-p}{2}\right]} |M_p^{-1}|^{1/2} |G^{-1}| [1 - v' G'^{-1} M_p^{-1} G^{-1} v]^{(n-3-p)/2},$$

over the region  $v' G'^{-1} M_p^{-1} G^{-1} v < 1$ , and zero otherwise, where  $G$  is as in Lemma 2.11.



*Proof.* Clearly,  $\mathbf{y} = J\mathbf{x}/\sqrt{\mathbf{x}'J\mathbf{x}}$  and  $J\mathbf{1} = \mathbf{0}$ . From Lemma 2.11,

$$\begin{aligned} V &= B\mathbf{y} = G\Gamma\mathbf{y} \\ &= G(I_p:0)Q\mathbf{y} \quad (0 \text{ is a } p \times (n-p) \text{ zero matrix}) \\ &= G(I_p:0)QTQ'(Q\mathbf{x}')/\sqrt{(Q\mathbf{x})'QJQ'(Q\mathbf{x})}. \end{aligned}$$

Let  $Z = Q\mathbf{x}$ ; then under  $H_0$ ,  $Z \sim N(\mathbf{0}, \sigma^2 I_n)$ . Let  $\mathbf{t} = \Gamma\mathbf{y} = (I_p:0)MZ/\sqrt{Z'MZ}$ , then from Worsely (1979, 1983), the pdf of  $\mathbf{t}$  is:

$$f(t_1, \dots, t_p) = \frac{\Gamma\left[\frac{n-1}{2}\right]}{\pi^{p/2}\Gamma\left[\frac{n-1-p}{2}\right]} |M_p^{-1}|^{1/2} [1 - \mathbf{t}'M_p^{-1}\mathbf{t}]^{(n-3-p)/2}$$

for  $\mathbf{t}'M_p^{-1}\mathbf{t} < 1$ , and zero otherwise.

Therefore,  $V = B\mathbf{y} = G\Gamma\mathbf{y} = G\mathbf{t}$  has the pdf

$$f_p(v) = \frac{\Gamma\left[\frac{n-1}{2}\right]}{\pi^{p/2}\Gamma\left[\frac{n-1-p}{2}\right]} |M_p^{-1}|^{1/2} |G^{-1}| [1 - v'G'^{-1}M_p^{-1}G^{-1}v]^{(n-3-p)/2},$$

over the region  $v'G'^{-1}M_p^{-1}G^{-1}v < 1$ , and zero otherwise.  $\square$

Now it is time to prove the theorem.

*Proof of Theorem 2.10* Because

$$\begin{aligned} M &= QJQ' \\ &= \begin{pmatrix} \Gamma \\ \Gamma_0 \end{pmatrix} \begin{pmatrix} I_n - \frac{1}{n}\mathbf{1}\mathbf{1}' \end{pmatrix} \begin{pmatrix} \Gamma' & \Gamma_0' \end{pmatrix} \\ &= \begin{pmatrix} I_p - \frac{1}{n}\Gamma\mathbf{1}\mathbf{1}'\Gamma' & * \\ * & * \end{pmatrix}, \end{aligned}$$

we have  $M_p = I_p - (1/n)\Gamma\mathbf{1}\mathbf{1}'\Gamma'$ . But  $\Gamma\mathbf{1} = \mathbf{0}_p$  as  $B\mathbf{1} = \mathbf{0}_p$ , therefore  $M_p = I_p$ . From Lemma 2.12, we thus obtain that the pdf of  $V = B\mathbf{y}$  at  $v$  is given by

$$f_p(v) = \frac{\Gamma\left[\frac{n-1}{2}\right]}{\pi^{p/2}\Gamma\left[\frac{n-1-p}{2}\right]} |\Sigma|^{-1/2} [1 - v'\Sigma^{-1}v]^{(n-3-p)/2},$$

for  $v'\Sigma^{-1}v < 1$ , and zero otherwise, where  $\Sigma = GM_p^{-1}G' = GG' = G\Gamma\Gamma'G' = BB'$ .  $\square$

Consequently, based on the  $t$ -distribution with  $n-2$  degrees of freedom and Theorem 2.10, the null probability function of  $V$  can be calculated through (2.9).

*Remark 2.13* Ignoring the higher-order terms in (2.9), one can obtain Bonferroni approximations for the distribution of  $V$ . Worsley (1979) obtained the percentage points for the Bonferroni approximations.

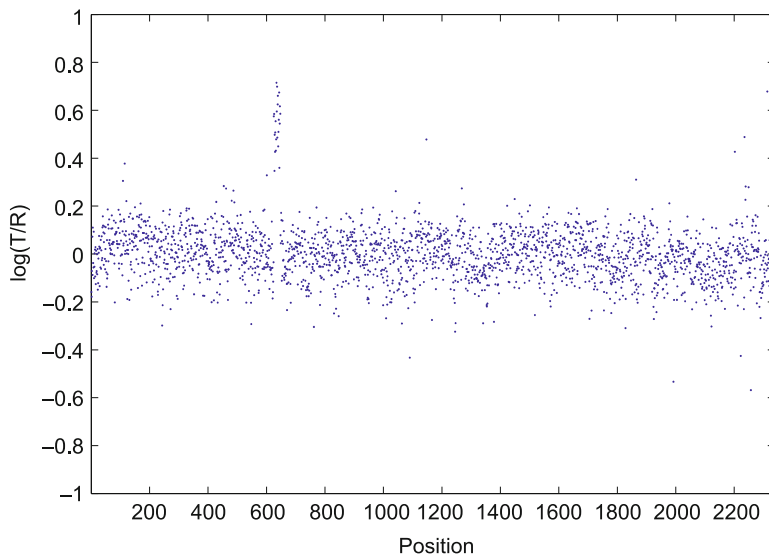
### 2.1.3 Application to Biomedical Data

In the last decade or so, life science research has been advanced by fast developing biotechnologies. Microarray technology, among the many well-developed biotechnologies, is a breakthrough that makes it possible to quantify the expression patterns of thousands or tens of thousands of genes in various tissues, cell lines, and conditions simultaneously. Biologists, geneticists, and medical researchers now routinely use microarray technology in their specified research projects thus resulting in voluminous numerical data related to expression of each gene encoded in the genome, the content of proteins and other classes of molecules in cells and tissues, and cellular responses to stimuli, treatments, and environmental factors.

Biological and medical research (e.g., see Lucito et al. 2000) reveals that some forms of cancer are caused by somatic or inherited mutations in oncogenes and tumor suppressor genes; cancer development and genetic disorders often result in chromosomal DNA copy number changes or copy number variations (CNVs). Consequently, identification of these loci where the DNA copy number changes or CNVs have taken place will (at least partially) facilitate the development of medical diagnostic tools and treatment regimes for cancer and other genetic diseases. Due to the advancement in array technology, the array Comparative Genomic Hybridization (aCGH) technique (see Kallioniemi et al. 1992 and Pinkel et al. 1998) or single nucleotide polymorphism (SNP) arrays (see Nannya et al. 2005) are often used in experiments that are deemed to study DNA copy numbers. The resulting data are typically called aCGH and SNP array data, respectively. However, because of the random noise inherited in the imaging and hybridization process in the DNA copy number experiments, identifying statistically significant CNVs or DNA copy number changes in aCGH data and in SNP array data is challenging.

In DNA copy number experiments such as aCGH copy number experiments, differentially labeled sample and reference DNA are hybridized to DNA microarrays, and the sample intensities of the test and reference samples are obtained (Pinkel et al. 1998, Pollack et al. 1999, and Myers et al. 2004). As the reference sample is assumed or chosen to have no copy number changes, markers whose test sample intensities are significantly higher (or lower) than the reference sample intensities are corresponding to DNA copy number gains (or losses) in the test sample at those locations (Olshen et al. 2004).

Concretely, the test sample intensity at locus  $i$  on the genome is usually denoted by  $T_i$  and the corresponding reference sample intensity by  $R_i$ , and the



**Fig. 2.4** The genome of the fibroblast cell line GM01524 of Snijders et al. (2001)

normalized log base 2 ratio of the sample and reference intensities,  $\log_2 T_i/R_i$ , at the  $i$ th biomarker, is one of the default outputs after the DNA copy number experiment is conducted using the aCGH technique. Here,  $\log_2 T_i/R_i = 0$  indicates no DNA copy number change at locus  $i$ ,  $\log_2 T_i/R_i < 0$  reveals a deletion at locus  $i$ , and  $\log_2 T_i/R_i > 0$  signifies duplication in the test sample at that locus. Due to various random noise, which occurs largely during the experimental and image processing stages, the  $\log_2 T_i/R_i$  becomes a random variable. Ideally, this random variable is assumed to follow a Gaussian distribution of mean 0 and constant variance  $\sigma^2$ . Then, deviations from the constant parameters (mean and variance) presented in  $\log_2 T_i/R_i$  data may indicate a copy number change. Hence, the key to identifying true DNA copy number changes becomes the problem of how to identify changes in the parameters of a normal distribution based on the observed sequence of  $\log_2 T_i/R_i$ . [Figure 2.4](#) is the scatterplot of the log base 2 ratio intensities of the genome of the fibroblast cell line GM01524 obtained by Snijders et al. (2001).

Since the publication of aCGH data by many research laboratories on copy number studies for different cell lines or diseases, analyzing aCGH data has become an active research topic for scientists, data analysts, and biostatisticians among others. A recent survey on the methods developed for analyzing aCGH data can be found in Chen (2010).

Among the many methods used for aCGH data, some methods are rooted in statistical change point analysis. Olshen et al. (2004) proposed a circular binary segmentation (CBS) method to identify DNA copy number changes in an aCGH database on the mean change point model proposed in Sen and Srivastava (1975a). This CBS method is mainly the combination of the

likelihood-ratio based test for testing no change in the mean against exactly one change in the mean with the BSP (Vostrikova, 1981) for searching multiple change points in the mean, assuming that the variance is unchanged. The idea of the CBS method (Olshen et al. 2004) can be summarized as follows.

Let  $X_i$  denote the normalized  $\log_2 R_i/G_i$  at the  $i$ th locus along the chromosome; then  $\{X_i\}$  is considered as a sequence of normal random variables taken from  $N(\mu_i, \sigma_i^2)$ , respectively, for  $i = 1, \dots, n$ . Consider any segment of the sequence of the log ratio intensities  $\{X_i\}$  (assumed to follow normal distributions) to be spliced at the two ends to form a circle; the test statistic  $Z_c$  of the CBS is based on the modified likelihood-ratio test and is specifically,

$$Z_c = \max_{1 \leq i < j \leq n} |Z_{ij}|, \quad (2.10)$$

where  $Z_{ij}$  is the likelihood-ratio test statistic given in Sen and Srivastava (1975a) for testing the hypothesis that the arc from  $i + 1$  to  $j$  and its complement have different means (i.e., there is a change point in the mean of the assumed normal distribution for the  $X_i$ s) and is given by:

$$Z_{ij} = \frac{1}{\{1/(j-i) + 1/(n-j+i)\}^{1/2}} \left\{ \frac{S_j - S_i}{j-i} - \frac{S_n - S_j + S_i}{n-j+i} \right\}, \quad (2.11)$$

with

$$S_i = X_1 + X_2 + \dots + X_i, \quad 1 \leq i < j \leq n.$$

Note that  $Z_c$  allows for both a single change ( $j = n$ ) and the epidemic alternative ( $j < n$ ). A change is claimed if the statistic exceeds an appropriate critical value at a given significant level based on the null distribution. However, the null distribution of the test statistic  $Z_c$  is not attainable so far in the literature of change point analysis. Then, as suggested in Olshen et al. (2004), the critical value when the  $X_i$ s are normal needs to be computed using Monte Carlo simulations or the approximation given by Siegmund (1986) for the tail probability. Once the null hypothesis of no change is rejected the changepoint(s) is (are) estimated to be  $i$  (and  $j$ ) such that  $Z_c = |Z_{ij}|$  and the procedure is applied recursively to identify all the changes in the whole sequence of the log ratio intensities of a chromosome (usually of hundreds to thousands of observations). The CBS algorithm is written as an *R* package and is available from the *R* Project website.

The influence of the CBS to the analyses of aCGH data is tremendous as the CBS method provided a statistical framework to the analysis of DNA copy number analysis. The  $p$ -value given by the CBS for a specific locus being a change point, however, is only obtained by a permutation method and the calculation of such a  $p$ -value takes a long computation time when the sequence is long, which is the case for high-density array data. Hence the CBS method has the slowest computational speed as pointed out in Picard et al. (2005). A recent result in Venkatraman and Olshen (2007) has improved the

**Table 2.1** Computational Output for Cell Line GM07408 Using the *R* Package DNACopy (Venkatraman and Olshen 2007)

Chr.			Number			
ID	Start Locus	End Locus	Mark	Seg Mean	Statistic	<i>p</i> -Value
1	468.3075	72188.15	46	0.0005	6.798636	$3.974580 \times 10^{-10}$
1	77228.4545	91439.90	9	-0.1613	5.206164	$6.896233 \times 10^{-6}$
1	93058.0760	209867.40	63	0.0002	4.205344	$7.806819 \times 10^{-4}$
1	211009.1200	240000.00	15	-0.1129	NA	NA
2	0.0000	245000.00	65	-0.0470	NA	NA
3	0.0000	218000.00	83	-0.0831	NA	NA
4	0.0000	15439.24	11	0.0549	4.091054	$1.604376 \times 10^{-3}$
4	22245.2500	169000.00	126	-0.0728	2.753006	$1.323652 \times 10^{-1}$
4	170170.0000	177387.29	8	0.0242	4.995123	$4.854205 \times 10^{-6}$
4	178400.0000	179200.0	4	0.1962	7.467238	$1.040446 \times 10^{-12}$
4	179490.1710	184000.00	12	0.0193	NA	NA
5	0.0000	198500.00	110	-0.0259	NA	NA
6	0.0000	65990.01	43	-0.0087	5.582021	$7.693862 \times 10^{-7}$
6	65990.0145	103277.13	12	-0.1887	3.581999	$6.122494 \times 10^{-3}$
6	104186.6075	188000.00	29	-0.0630	NA	NA
7	0.0000	161500.00	173	0.0166	NA	NA
8	0.0000	13266.95	23	0.0091	5.427554	$1.043974 \times 10^{-6}$
8	16299.9040	16815.62	3	-0.2023	5.065677	$7.349159 \times 10^{-6}$
8	17992.9110	41655.39	24	0.0483	6.296318	$1.294907 \times 10^{-8}$
8	41932.6780	86296.08	48	-0.0553	5.116016	$1.016467 \times 10^{-5}$
8	87774.9320	100534.00	15	-0.1495	7.786791	$2.741869 \times 10^{-13}$
8	107253.3650	147000.00	38	0.0241	NA	NA
9	0.0000	33134.86	29	-0.1002	5.493857	$1.828478 \times 10^{-6}$
9	33856.8540	115000.00	77	0.0187	NA	NA
10	0.0000	14349.90	16	0.0105	2.317172	$2.252879 \times 10^{-1}$
10	15126.2185	47803.26	23	-0.0401	5.230861	$3.958791 \times 10^{-6}$
10	49954.2190	69209.44	14	-0.1412	8.040106	$1.713658 \times 10^{-14}$
10	69549.0280	73067.45	9	0.0718	2.372705	$1.860309 \times 10^{-1}$
10	74000.0000	108902.62	25	0.0277	4.739140	$4.486729 \times 10^{-5}$
10	110000.0000	117000.00	10	-0.0562	5.814313	$1.408531 \times 10^{-7}$
10	118499.8495	142000.00	23	0.1067	NA	NA

computational speed of CBS. If there is an analytic formula for calculating the *p*-value of the change point hypothesis, the computational speed will undoubtedly be faster and more convenient.

For the application of the CBS method to the analysis of 15 fibroblast cell lines obtained in Snijders et al. (2001), readers are referred to Olshen et al. (2004) and Venkatraman and Olshen (2007). As a complete example, we present here the use of the *R*-package, DNACopy (Venkatraman and Olshen 2007), on the fibroblast cell line GM07408 (Snijders et al. 2001). After the data are read into the *R*-package, we obtained the *p*-values (based on the test statistics  $Z_{ij}$ , given in (2.11)) for each segment being a change along the chromosome. These results directly output from DNACopy are listed in [Table 2.1](#) of this chapter.

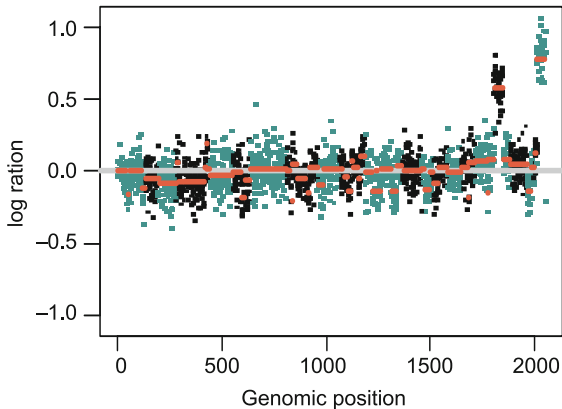
**Table 2.2** Computational Output for Cell Line GM07408 Using DNAcopy: [Table 2.1](#) Continued

Chr.		Number		Seg Mean	Statistic	<i>p</i> -Value
ID	Start Locus	End Locus	Mark			
11	0.0000	20607.03	34	−0.0076	4.551992	$1.319627 \times 10^{-4}$
11	20719.3180	34420.00	14	−0.1402	5.978688	$4.620736 \times 10^{-8}$
11	34420.0000	39388.78	14	0.0011	5.925536	$6.168566 \times 10^{-8}$
11	39623.3960	48010.94	13	−0.1385	6.550567	$2.330282 \times 10^{-9}$
11	48923.4020	87856.91	51	0.0187	7.290779	$1.511135 \times 10^{-11}$
11	88570.8380	117616.94	24	−0.1376	7.615232	$1.051503 \times 10^{-12}$
11	117817.0610	145000.00	30	0.0338	NA	NA
12	315.6600	142000.00	91	0.0091	NA	NA
13	5653.9480	28325.81	14	0.0122	6.231862	$1.191981 \times 10^{-8}$
13	28469.3085	78470.20	22	−0.1277	5.223176	$4.356758 \times 10^{-6}$
13	80645.9990	100500.00	18	−0.0093	NA	NA
14	769.5125	40288.14	30	−0.0847	6.140269	$3.336396 \times 10^{-8}$
14	42901.6730	97000.00	39	0.0301	NA	NA
15	0.0000	79000.00	67	−0.0065	NA	NA
16	0.0000	52092.53	31	0.0312	5.063312	$9.064410 \times 10^{-6}$
16	53000.0000	55905.26	4	−0.1804	4.543496	$1.095742 \times 10^{-4}$
16	57000.0000	84000.00	30	0.0549	NA	NA
17	0.0000	52738.37	58	0.0686	5.153190	$8.319006 \times 10^{-6}$
17	52804.6060	56190.77	4	−0.1529	4.598118	$8.167274 \times 10^{-5}$
17	56313.4240	86000.00	28	0.0868	NA	NA
18	0.0000	86000.00	44	0.5821	NA	NA
19	0.0000	70000.00	35	0.0828	NA	NA
20	0.0000	73000.00	82	0.0459	NA	NA
21	3130.9400	19079.91	20	−0.1420	6.288501	$7.767927 \times 10^{-9}$
21	19247.3920	30000.00	13	0.0274	NA	NA
22	1100.0000	33000.00	13	0.1213	NA	NA
23	0.0000	155000.00	42	0.7744	NA	NA

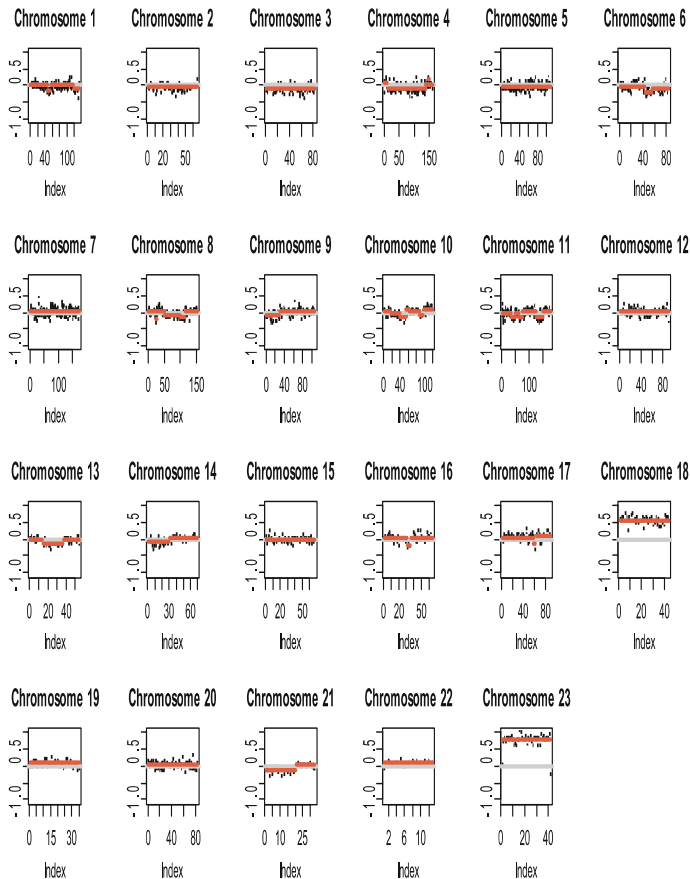
The plot that indicated the whole genome of GM07408 is also obtained and is given as [Figure 2.5](#). In [Figure 2.5](#), the adjacent chromosomes are indicated by green and black colors alternately, and the red line segment represents the sample mean for each segment.

[Figure 2.6](#) gives the collection of all 23 plots for the 23 chromosomes of the cell line GM07408 with changes identified. It is customary to also obtain a plot of the log ratio intensities chromosome by chromosome with mean changes identified by the CBS method. More plots, say [Figures 2.7–2.11](#), are also obtained by using the same *R*-package for chromosomes 4, 5, 8–11, 22, and 23 of the cell line GM07408 with changes identified.

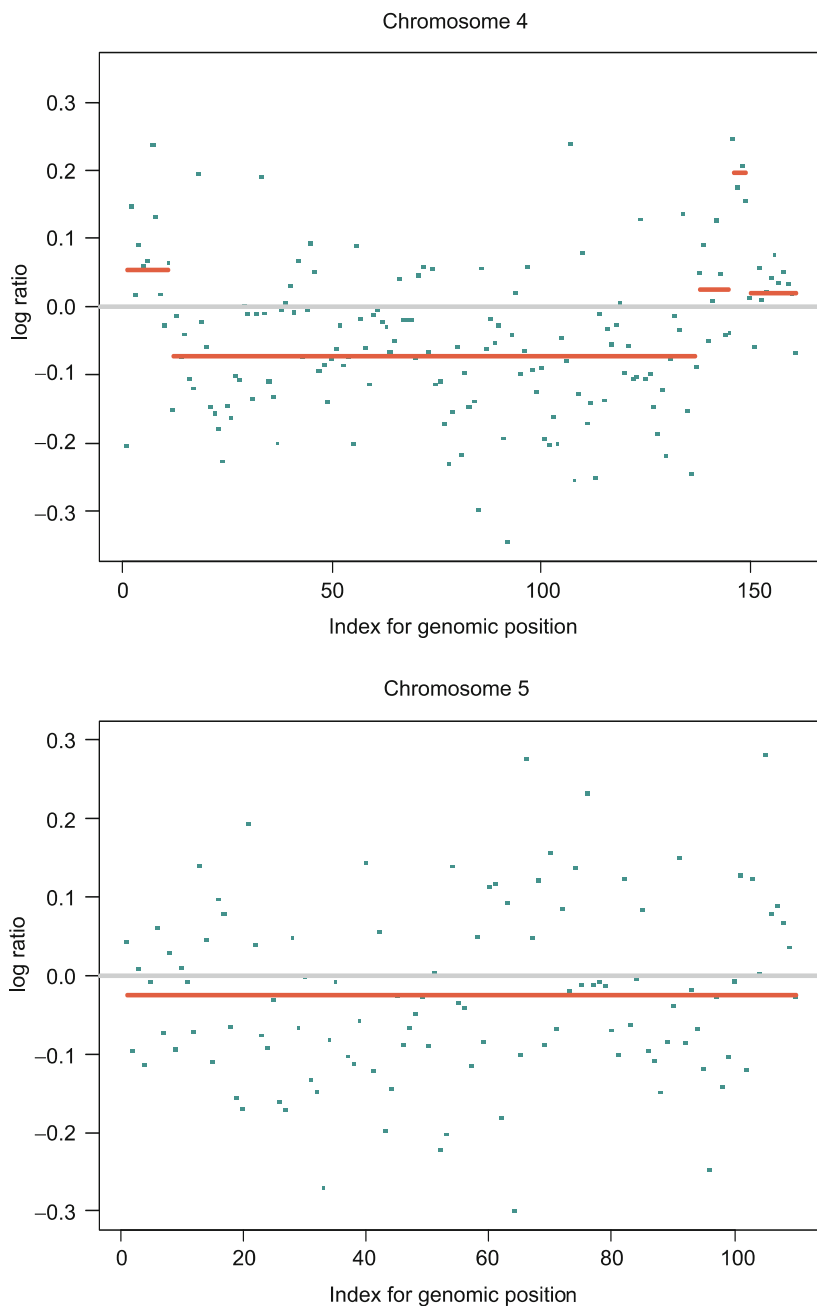
There are changes identified by using the CBS method on each chromosome.



**Fig. 2.5** The genome plot of the fibroblast cell line GM07408 of Snijders et al. (2001) using the *R*-package DNAcopy

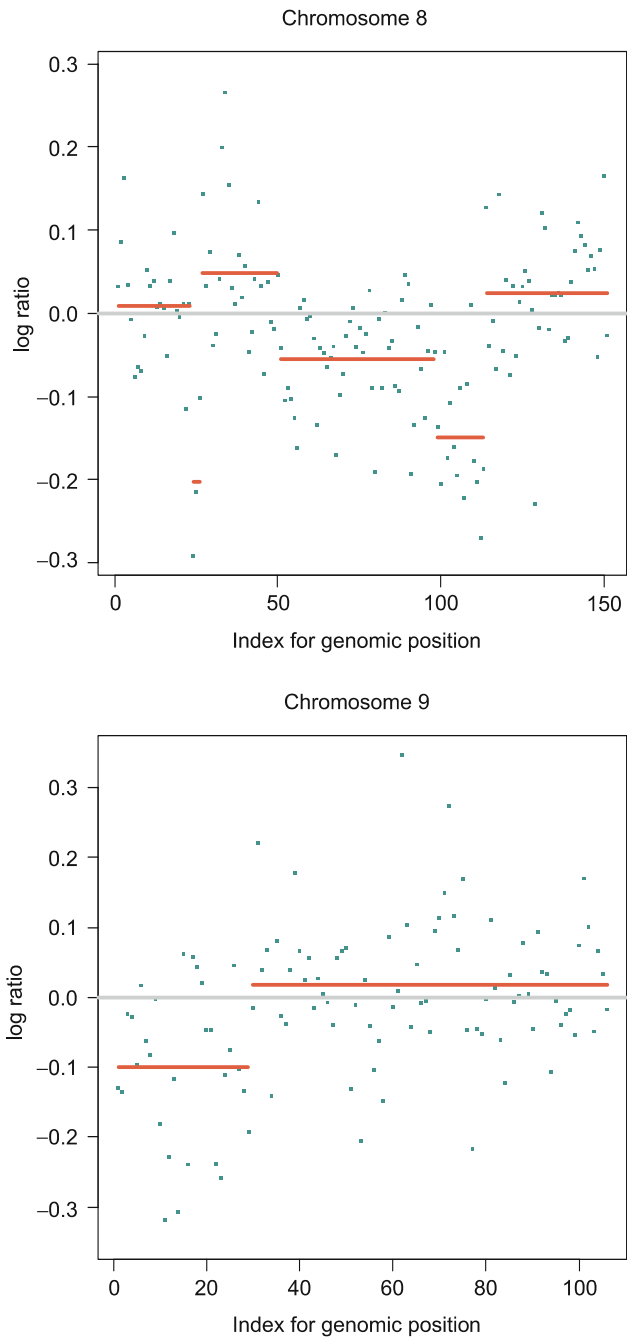


**Fig. 2.6** The plots of all chromosomes on the fibroblast cell line GM07408 of Snijders et al. (2001) with mean changes indicated

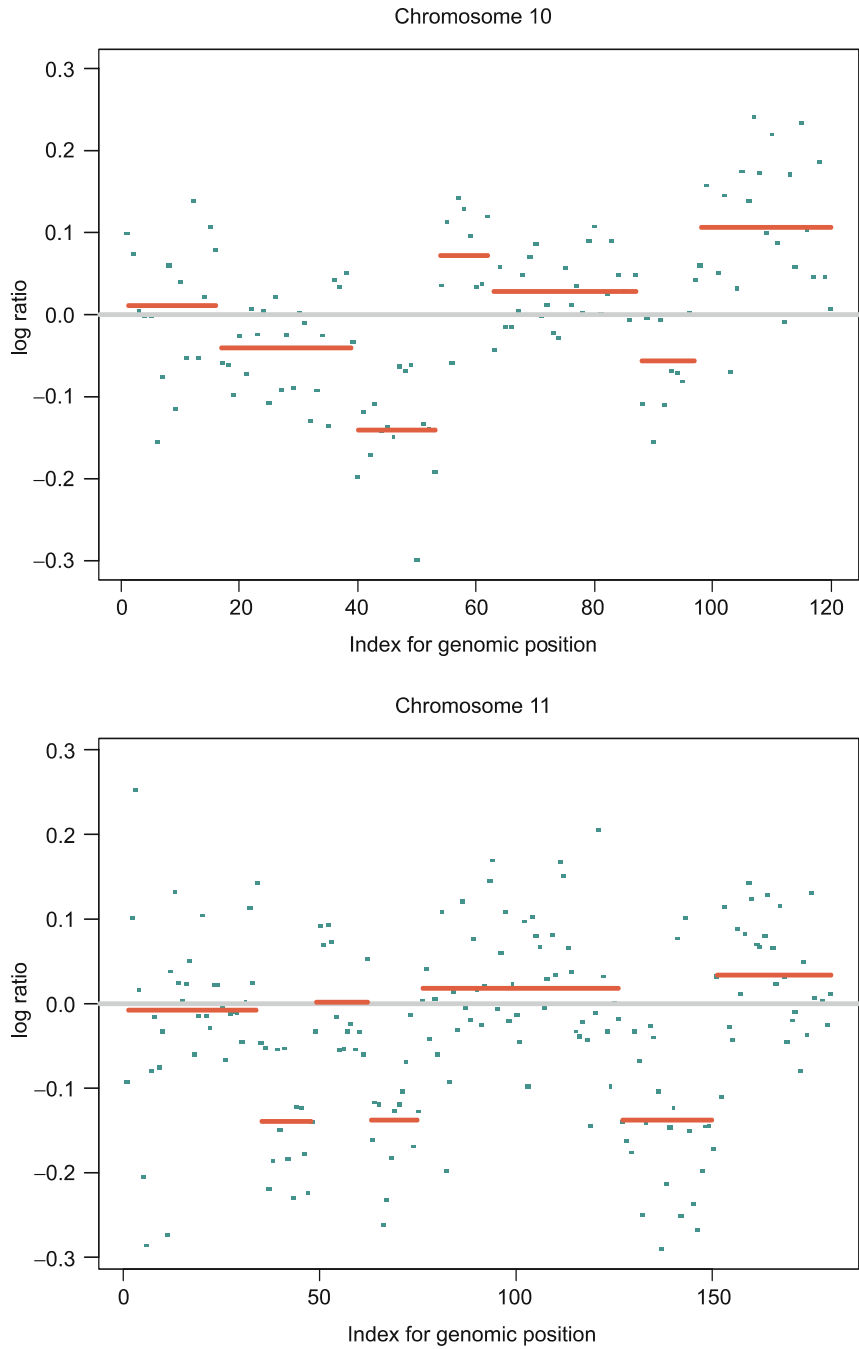


**Fig. 2.7** The plot of chromosomes 4 and 5 on the fibroblast cell line GM07408 of Snijders et al. (2001) with mean changes indicated

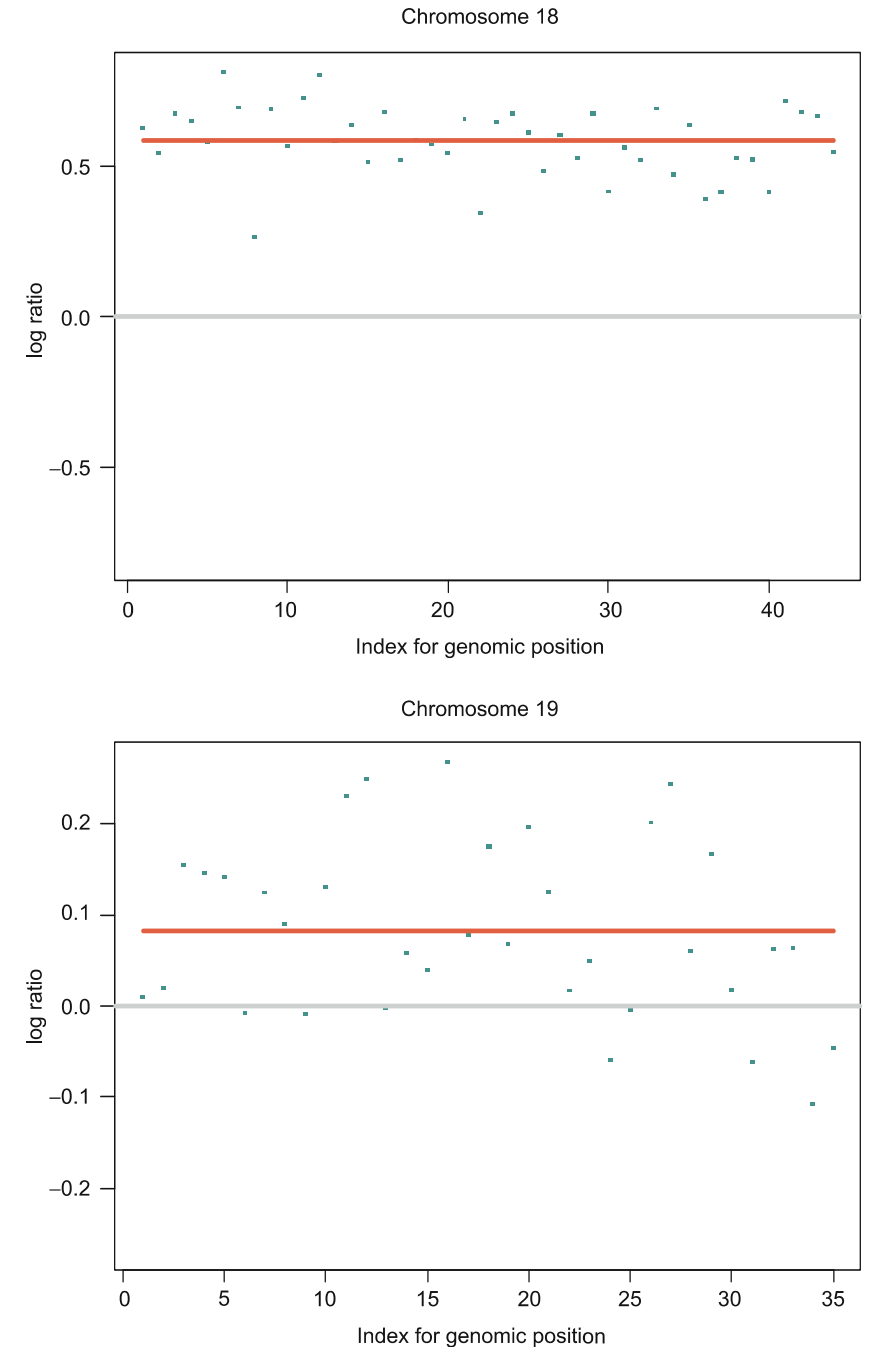




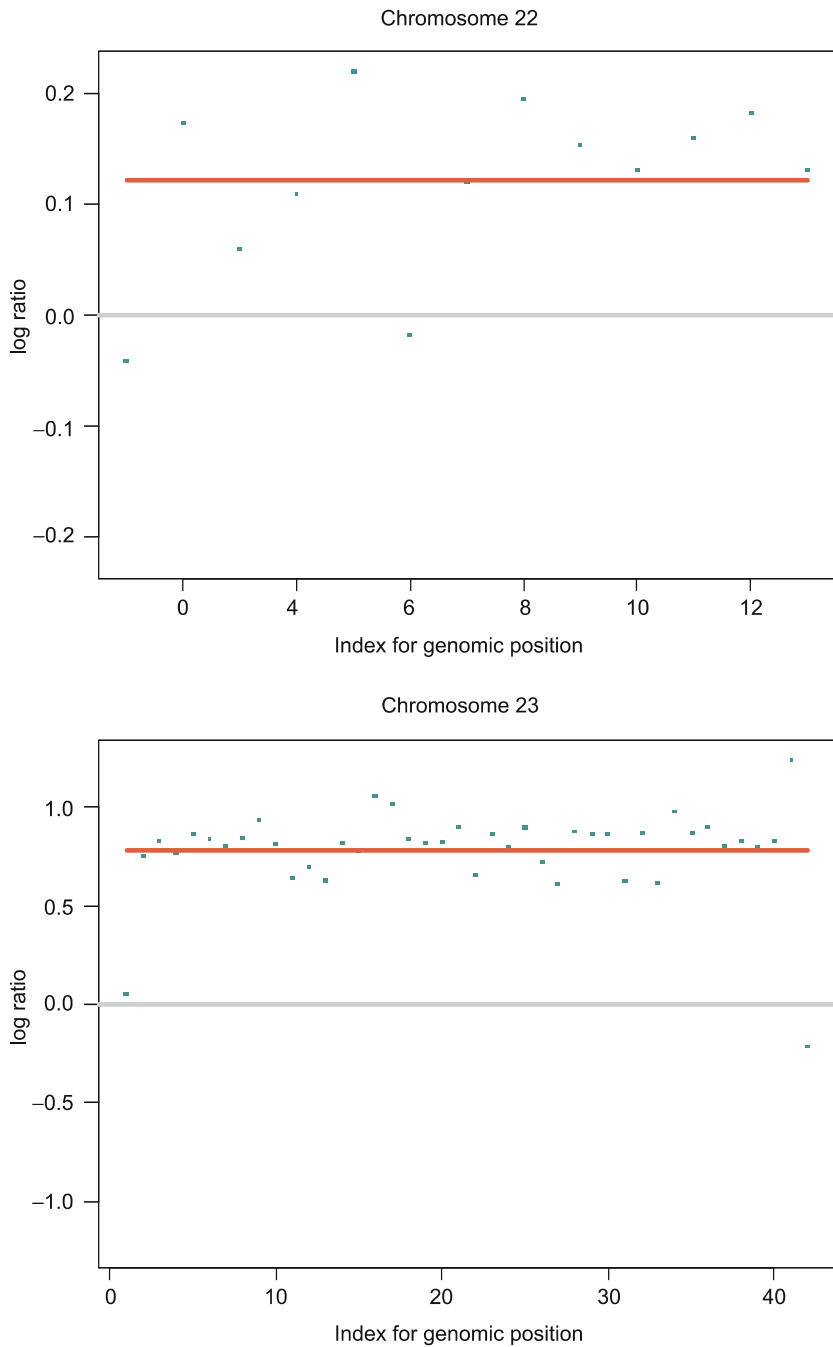
**Fig. 2.8** The plot of chromosomes 8 and 9 on the fibroblast cell line GM07408 of Snijders et al. (2001) with mean changes indicated



**Fig. 2.9** The plot of chromosomes 10 and 11 on the fibroblast cell line GM07408 of Snijders et al. (2001) with mean changes indicated



**Fig. 2.10** The plot of chromosomes 18 and 19 on the fibroblast cell line GM07408 of Snijders et al. (2001) with mean changes indicated



**Fig. 2.11** The plot of chromosomes 22 and 23 on the fibroblast cell line GM07408 of Snijders et al. (2001) with mean changes indicated

## 2.2 Variance Change

Testing and estimation about mean change in a Gaussian model has been studied in Section 2.1. The corresponding problem of changes in the regression model is studied in Chapter 4. Inference about variance changes while the mean remains common has been studied by Wichern, Miller, and Hsu (1976), Hsu (1977), Davis (1979), Abraham and Wei (1984), Inclán (1993), and Chen and Gupta (1997). In this section, the variance change problem for the univariate Gaussian model using different methods is considered

Let  $x_1, x_2, \dots, x_n$  be a sequence of independent normal random variables with parameters  $(\mu, \sigma_1^2), (\mu, \sigma_2^2), \dots, (\mu, \sigma_n^2)$ , respectively. Assume that  $\mu$  is known. The interest here is to test the hypothesis (see Gupta and Tang, 1987):

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = \sigma^2 (\text{unknown}), \quad (2.12)$$

versus the alternative:

$$H_A : \sigma_1^2 = \dots = \sigma_{k_1}^2 \neq \sigma_{k_1+1}^2 = \dots = \sigma_{k_2}^2 \neq \dots \neq \sigma_{k_q+1}^2 = \dots = \sigma_n^2,$$

where  $q$  is the unknown number of change points, and  $1 \leq k_1 < k_2 < \dots < k_q < n$ , are the unknown positions of the change points, respectively. Using the binary segmentation procedure, as described in Chapter 1, it suffices to test and estimate the position of a single change point at each stage, that is, to test  $H_0$  defined by (2.10) against the following alternative:

$$H_1 : \sigma_1^2 = \dots = \sigma_k^2 \neq \sigma_{k+1}^2 = \dots = \sigma_n^2, \quad (2.13)$$

where  $1 < k < n$ , is the unknown position of the single change point.

### 2.2.1 Likelihood-Ratio Procedure

Under  $H_0$ , the log likelihood function is:

$$\log L_0(\sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}.$$

Let  $\hat{\sigma}^2$  be the MLE of  $\sigma^2$  under  $H_0$ , Then

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n},$$

and the maximum likelihood is

$$\log L_0(\hat{\sigma}^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2}.$$

Under  $H_1$ , the log likelihood function is:

$$\begin{aligned} \log L_1(\sigma_1^2, \sigma_n^2) = & -\frac{n}{2} \log 2\pi - \frac{k}{2} \log \sigma_1^2 - \frac{n-k}{2} \log \sigma_n^2 \\ & - \frac{\sum_{i=1}^k (x_i - \mu)^2}{2\sigma_1^2} - \frac{\sum_{i=k+1}^n (x_i - \mu)^2}{2\sigma_n^2}. \end{aligned}$$

Let  $\hat{\sigma}_1^2, \hat{\sigma}_n^2$  be the MLEs of  $\sigma_1^2, \sigma_n^2$ , respectively; then

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^k (x_i - \mu)^2}{k}, \quad \hat{\sigma}_n^2 = \frac{\sum_{i=k+1}^n (x_i - \mu)^2}{n-k},$$

and the maximum log likelihood is:

$$\log L_1(\hat{\sigma}_1^2, \hat{\sigma}_n^2) = -\frac{n}{2} \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{n-k}{2} \log \hat{\sigma}_n^2 - \frac{n}{2}.$$

Then the likelihood-ratio procedure statistic is

$$\lambda_n = \left\{ \max_{1 < k < n-1} [n \log \hat{\sigma}^2 - k \log \hat{\sigma}_1^2 - (n-k) \log \hat{\sigma}_n^2] \right\}^{1/2}. \quad (2.14)$$

Notice that, to be able to obtain the MLEs, we can only detect changes for  $2 \leq k \leq n-2$ . According to the principle of the maximum likelihood procedure, we estimate the position  $k$  of the change point by  $\hat{k}$  such that (2.12) attains its maximum at  $\hat{k}$ .

Next, we derive the asymptotic null distribution of  $\lambda_n$ . Note that, for large  $n$ ,

$$\begin{aligned} \lambda_n &= \left\{ \max_{1 < k < n-1} [n \log \hat{\sigma}^2 - k \log \hat{\sigma}_1^2 - (n-k) \log \hat{\sigma}_n^2] \right\}^{1/2} \\ &= \max_{1 < k < n-1} \left[ n \log \frac{\sum_{i=1}^n (x_i - \mu)^2}{n} - k \log \frac{\sum_{i=1}^k (x_i - \mu)^2}{k} \right. \\ &\quad \left. - (n-k) \log \frac{\sum_{i=k+1}^n (x_i - \mu)^2}{n-k} \right]^{1/2}. \end{aligned}$$

Under  $H_0$ ,  $y_i = (x_i - \mu)/\sigma \sim N(0, 1)$ ; then

$$\lambda_n \stackrel{D}{=} \max_{1 < k < n-1} \left[ n \log \frac{\sum_{i=1}^n y_i^2}{n} - k \log \frac{\sum_{i=1}^k y_i^2}{k} - (n-k) \log \frac{\sum_{i=k+1}^n y_i^2}{n-k} \right]^{1/2}.$$

Now, using the three-term Taylor expansion, we write

$$\begin{aligned}
\xi_k &= n \log \frac{\sum_{i=1}^n y_i^2}{n} - k \log \frac{\sum_{i=1}^k y_i^2}{k} - (n-k) \log \frac{\sum_{i=k+1}^n y_i^2}{n-k} \\
&= n \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right) - \frac{n}{2} \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right)^2 + \frac{n}{3} (\theta_n^{(1)})^{-3} \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right)^3 \\
&\quad - k \left( \frac{\sum_{i=1}^k y_i^2}{k} - 1 \right) + \frac{k}{2} \left( \frac{\sum_{i=1}^k y_i^2}{k} - 1 \right)^2 - \frac{k}{3} (\theta_n^{(2)})^{-3} \left( \frac{\sum_{i=1}^k y_i^2}{k} - 1 \right)^3 \\
&\quad - (n-k) \left( \frac{\sum_{i=k+1}^n y_i^2}{n-k} - 1 \right) + \frac{n-k}{2} \left( \frac{\sum_{i=k+1}^n y_i^2}{n-k} - 1 \right)^2 \\
&\quad - \frac{n-k}{3} (\theta_n^{(3)})^{-3} \left( \frac{\sum_{i=k+1}^n y_i^2}{n-k} - 1 \right)^3 \\
&= -\frac{1}{2n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 \\
&\quad + \frac{n}{3} (\theta_n^{(1)})^{-3} \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right)^3 - \frac{k}{3} (\theta_n^{(2)})^{-3} \left( \frac{\sum_{i=1}^k y_i^2}{k} - 1 \right)^3 \\
&\quad - \frac{n-k}{3} (\theta_n^{(3)})^{-3} \left( \frac{\sum_{i=k+1}^n y_i^2}{n-k} - 1 \right)^3,
\end{aligned}$$

where  $|\theta_n^{(1)} - 1| < |\sum_{i=1}^n y_i^2/n - 1|$ ,  $|\theta_n^{(2)} - 1| < |\sum_{i=1}^k y_i^2/k - 1|$ , and  $|\theta_n^{(3)} - 1| < |\sum_{i=k+1}^n y_i^2/(n-k) - 1|$ . Denote  $\xi_k = W_k + Q_k + R_k$ , where

$$\begin{aligned}
W_k &= -\frac{1}{2n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2, \\
Q_k &= \frac{n}{3} (\theta_n^{(1)})^{-3} \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right)^3 - \frac{k}{3} (\theta_n^{(2)})^{-3} \left( \frac{\sum_{i=1}^k y_i^2}{k} - 1 \right)^3, \\
R_k &= -\frac{n-k}{3} (\theta_n^{(3)})^{-3} \left( \frac{\sum_{i=k+1}^n y_i^2}{n-k} - 1 \right)^3.
\end{aligned}$$

Before having our main theorem, we first give the following lemmas which are needed in the sequel.

**Lemma 2.14**

- (i)  $\max_{1 < k < n} k^{1/2}(\log \log k)^{-3/2} |Q_k| = O_p(1)$ ,  
(ii)  $\max_{1 < k < n} (n - k)^{1/2}(\log \log(n - k))^{-3/2} |R_k| = O_p(1)$ .

*Proof.* (i) From the law of iterated logarithm, we obtain:

$$\frac{\left| (\theta_n^{(1)})^{-1} n \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right) \right|}{(n \log \log n)^{1/2}} = O_p(1).$$

Then,

$$\frac{\left| (\theta_n^{(1)})^{-3} n^3 \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right)^3 \right|}{(n \log \log n)^{3/2}} = O_p(1);$$

that is,

$$n^{1/2}(\log \log n)^{-3/2} \left| (\theta_n^{(1)})^{-3} n \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right)^3 \right| = O_p(1).$$

But,  $k^{1/2}(\log \log k)^{-3/2} \leq n^{1/2}(\log \log n)^{-3/2}$  for  $1 < k < n$ , hence

$$\max_{1 < k < n} k^{1/2}(\log \log k)^{-3/2} \left| (\theta_n^{(1)})^{-3} n \left( \frac{\sum_{i=1}^n y_i^2}{n} - 1 \right)^3 \right| = O_p(1).$$

Similarly, we obtain

$$\max_{1 < k < n} k^{1/2}(\log \log k)^{-3/2} \left| (\theta_n^{(2)})^{-3} k \left( \frac{\sum_{i=1}^k y_i^2}{k} - 1 \right)^3 \right| = O_p(1).$$

That is,  $\max_{1 < k < n} k^{1/2}(\log \log k)^{-3/2} |Q_k| = O_p(1)$ , which completes the proof of (i).

(ii) Similar to the proof of (i).  $\square$

**Lemma 2.15** Let  $a(\log n) = (2 \log \log n)^{1/2}$ , and  $b(\log n) = 2 \log \log n + \frac{1}{2} \log \log \log n - \log \Gamma\left(\frac{1}{2}\right)$ ; then for all  $x \in R$ , as  $n \rightarrow \infty$ , the following hold.

- (i)  $a^2(\log n) \max_{1 < k < \log n} W_k - (x + b(\log n))^2 \xrightarrow{P} -\infty$ .  
(ii)  $a^2(\log n) \max_{1 < k < \log n} \xi_k - (x + b(\log n))^2 \xrightarrow{P} -\infty$ .  
(iii)  $a^2(\log n) \max_{n - \log n < k < n} W_k - (x + b(\log n))^2 \xrightarrow{P} -\infty$ .  
(iv)  $a^2(\log n) \max_{n - \log n < k < n} \xi_k - (x + b(\log n))^2 \xrightarrow{P} -\infty$ .



*Proof.* (i) Because as  $k \longrightarrow \infty$ ,

$$\frac{1}{k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 \xrightarrow{P} \frac{1}{2},$$

then

$$\frac{[\sum_{i=1}^k (y_i^2 - 1)]^2}{k \log k} \xrightarrow{P} 0,$$

as  $k \longrightarrow \infty$ . Thus, there exists a constant  $c$ ,  $0 < c < 1$ , such that for large  $k$ ,

$$0 < \frac{[\sum_{i=1}^k (y_i^2 - 1)]^2}{k \log k} < 1 - c.$$

Now,

$$\begin{aligned} & \frac{a^2(\log n) \max_{1 < k < \log n} \frac{1}{k} [\sum_{i=1}^k (y_i^2 - 1)]^2}{[b(\log n)]^2} \\ & \leq \frac{(2 \log \log n) \max_{1 < k < \log n} \frac{1}{k} [\sum_{i=1}^k (y_i^2 - 1)]^2}{[2 \log \log n]^2} \\ & \leq \max_{1 < k < \log n} \frac{[\sum_{i=1}^k (y_i^2 - 1)]^2}{k \log \log n} \\ & \leq \max_{1 < k < \log n} \frac{[\sum_{i=1}^k (y_i^2 - 1)]^2}{k \log k} \\ & < 1 - c. \end{aligned}$$

Hence, as  $n \longrightarrow \infty$ ,

$$a^2(\log n) \max_{1 < k < \log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 - (x + b(\log n))^2 \xrightarrow{P} -\infty.$$

Similarly, as  $n \longrightarrow \infty$ , we obtain

$$a^2(\log n) \max_{1 < k < \log n} \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - (x + b(\log n))^2 \xrightarrow{P} -\infty.$$

Moreover, because as  $n \longrightarrow \infty$ ,

$$\frac{1}{2n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \xrightarrow{P} 1 \quad \text{and} \quad a^2(\log n) \longrightarrow \infty,$$

as  $n \longrightarrow \infty$ ,

$$\frac{a^2(\log n)}{2n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \xrightarrow{P} \infty.$$

Consequently, as  $n \longrightarrow \infty$ ,

$$a^2(\log n) - \frac{1}{2n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 - (x + b(\log n))^2 \xrightarrow{P} -\infty.$$

Therefore, as  $n \longrightarrow \infty$ ,

$$a^2(\log n) \max_{1 < k < \log n} W_k - (x + b(\log n))^2 \xrightarrow{P} -\infty.$$

Proceeding similarly as above, we can obtain (ii)–(iv).  $\square$

**Lemma 2.16** *As  $n \longrightarrow \infty$ , the following hold.*

- (i)  $a^2(\log n) \max_{\log n < k < n - \log n} |\xi_k - W_k| = o_p(1)$ .
- (ii)  $a^2(\log n) \max_{1 < k < n / \log n} \left| \frac{1}{(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \right| = o_p(1)$ .

*Proof.* (i) Clearly,  $\xi_k - W_k = Q_k + R_k$ . Now,

$$\begin{aligned} 0 &\leq a^2(\log n) \max_{\log n < k < n - \log n} |Q_k + R_k| \\ &\leq 2 \log \log n \max_{\log n < k < n - \log n} |Q_k| + 2 \log \log n \max_{\log n < k < n - \log n} |R_k| \\ &= 2 \log \log n \max_{\log n < k < n - \log n} \frac{(\log \log k)^{3/2}}{k^{1/2}} (\log \log k)^{-3/2} |Q_k| \\ &\quad + 2 \log \log n \max_{\log n < k < n - \log n} \frac{(\log \log(n-k))^{3/2}}{(n-k)^{1/2}} (\log \log(n-k))^{-3/2} |R_k| \\ &\leq \frac{2(\log \log n)^{5/2}}{(\log n)^{1/2}} \max_{\log n < k < n - \log n} k^{1/2} (\log \log k)^{-3/2} |Q_k| \\ &\quad + \frac{2(\log \log n)^{5/2}}{(\log n)^{1/2}} \max_{\log n < k < n - \log n} (n-k)^{1/2} (\log \log(n-k))^{-3/2} |R_k| \\ &\xrightarrow{P} 0, \end{aligned}$$

as  $n \longrightarrow \infty$ , hence

$$\lim_{n \rightarrow \infty} a^2(\log n) \max_{\log n < k < n - \log n} |\xi_k - W_k| \stackrel{P}{=} 0;$$

that is, (i) holds. (ii) For all  $n$  and  $k$ ,

$$E \left\{ \frac{1}{(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \right\} = 0,$$

therefore we have,

$$E \left\{ a^2(\log n) \left[ \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \right] \right\} = 0$$

for all  $n$  and  $k$ . Hence, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \left| \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \right| \xrightarrow{P} 0,$$

for all  $k$  and  $1 < k < n/\log n$ . That is,

$$a^2(\log n) \max_{1 < k < n/\log n} \left| \frac{1}{(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \right| = o_p(1).$$

□

**Lemma 2.17** For all  $x \in R$ , as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n - n/\log n} W_k - (x + b(\log n))^2 \xrightarrow{P} -\infty.$$

*Proof.* Recall that

$$W_k = \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - \frac{1}{2n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2.$$

Let's consider the first term of  $W_k$ .

From Theorem 2 of Darling and Erdős (1956), we have for  $x \in R$ , as  $n \rightarrow \infty$ ,

$$\begin{aligned} & P \left[ \max_{n/\log n < k < n - n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 \right. \\ & \quad \left. < (2 \log \log n)^{1/2} + \frac{\log \log \log n}{2(\log \log n)^{1/2}} + \frac{x}{2(\log \log n)^{1/2}} \right] \rightarrow \exp \left( -\frac{1}{\sqrt{\pi}} e^{-x} \right). \end{aligned}$$

Therefore, as  $n \longrightarrow \infty$ ,

$$\begin{aligned}
 P & \left[ \frac{a^2(\log n) \max_{n/\log n < k < n-n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2}{(x + b(\log n))^2} \right. \\
 & < \left. \frac{1}{2} \left( \frac{2 \log \log n + \frac{1}{2} \log \log \log n + x}{x + b(\log n)} \right)^2 \right] \\
 & \longrightarrow \exp \left( -\frac{1}{\sqrt{\pi}} e^{-x} \right).
 \end{aligned}$$

Because  $b(\log n) = 2 \log \log n + \frac{1}{2} \log \log \log n - \log \Gamma(\frac{1}{2})$ , it follows that as  $n \longrightarrow \infty$ ,

$$\begin{aligned}
 P & \left[ \frac{a^2(\log n) \max_{n/\log n < k < n-n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2}{(x + b(\log n))^2} \right. \\
 & < \left. \left( \frac{1}{\sqrt{2}} + \frac{\log \Gamma(1/2)}{\sqrt{2}(x + b(\log n))} \right)^2 \right] \\
 & \longrightarrow \exp \left( -\frac{1}{\sqrt{\pi}} e^{-x} \right).
 \end{aligned}$$

Choose  $n$  sufficiently large, such that

$$\left( \frac{1}{\sqrt{2}} + \frac{\log \Gamma(1/2)}{\sqrt{2}(x + b(\log n))} \right)^2 < 1 - M, \quad \text{for } 0 < M < 1,$$

therefore, as  $n \longrightarrow \infty$ ,

$$\begin{aligned}
 P & \left[ \frac{a^2(\log n) \max_{n/\log n < k < n-n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2}{(x + b(\log n))^2} < 1 - M \right] \\
 & \longrightarrow \exp \left( -\frac{1}{\sqrt{\pi}} e^{-x} \right).
 \end{aligned}$$

Now, letting  $x \longrightarrow \infty$ , as  $n \longrightarrow \infty$ , we obtain:

$$a^2(\log n) \max_{n/\log n < k < n-n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 - (x + b(\log n))^2 \xrightarrow{P} -\infty,$$

and similarly,

$$a^2(\log n) \max_{n/\log n < k < n - n/\log n} \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 - (x + b(\log n))^2 \xrightarrow{P} -\infty.$$

For the third term of  $W_k$  (i.e.,  $-(1/2n) [\sum_{i=1}^n (y_i^2 - 1)]^2$ ) because as  $n \rightarrow \infty$ ,  $-(1/2n) [\sum_{i=1}^n (y_i^2 - 1)]^2 \xrightarrow{P} -1$ ; therefore, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n - n/\log n} \left( -\frac{1}{2n} \left[ \sum_{i=1}^n (y_i^2 - 1) \right]^2 \right) - (x + b(\log n))^2 \xrightarrow{P} -\infty.$$

This completes the proof of the lemma.  $\square$

**Lemma 2.18** Let  $\{(z_i^{(1)}, \dots, z_i^{(d)}), 1 \leq i < \infty\}$  be independently and identically distributed random vectors, and define  $S^{(j)}(k) = \sum_{i=1}^k z_i^{(j)}, 1 \leq j \leq d$ . Assume that  $E[z_i^{(1)}] = E[z_i^{(2)}] = \dots = E[z_i^{(d)}] = 0$ , the covariance matrix of  $(z_i^{(1)}, \dots, z_i^{(d)})$  is the identity matrix, and  $\max_{1 \leq j \leq d} E|z_i^{(j)}|^r < \infty$  for some  $r > 2$ . Then as  $n \rightarrow \infty$ ,

$$a(\log n) \max_{1 \leq j \leq d} \left( \sum_{j=1}^d [k^{-1/2} S^{(j)}(k)]^2 \right)^{1/2} - b_d(\log n) \xrightarrow{D} y^*,$$

where  $y^*$  has cdf  $F_{y^*}(x) = \exp\{-e^{-x}\}$ ,  $a(x) = (2 \log x)^{1/2}$ ,  $b_d(x) = 2 \log x + (d/2) \log \log x - \log \Gamma(d/2)$ , and “ $\xrightarrow{D}$ ” means “convergence in distribution”.

*Proof.* See Horváth (1993).  $\square$

**Theorem 2.19** Under the null hypothesis  $H_0$ , as  $n \rightarrow \infty, k \rightarrow \infty$ , such that  $(k/n) \rightarrow \infty$ ; then for all  $x \in R$ ,

$$\lim_{n \rightarrow \infty} P[a(\log n)\lambda_n - b(\log n) \leq x] = \exp\{-2e^{-x}\},$$

where  $a(\log n)$  and  $b(\log n)$  are defined in Lemma 2.15.

*Proof.* From Lemma 2.14 (i) and (ii), we have

$$\max_{1 < k < n} \xi_k \stackrel{D}{=} \max_{\log n \leq k \leq n - \log n} \xi_k.$$

From Lemma 2.15 (i), it is seen that

$$\max_{1 < k < n} \xi_k \stackrel{D}{=} \max_{\log n \leq k \leq n - \log n} W_k.$$

From Lemma 2.15 (ii), we thus have

$$\max_{\log n \leq k \leq n/\log n} W_k \stackrel{D}{=} \max_{\log n \leq k \leq n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2. \quad (2.15)$$

In view of Lemma 2.16,

$$\max_{\log n \leq k \leq n-\log n} W_k \stackrel{D}{=} \max \left\{ \left[ \max_{\log n \leq k \leq n/\log n} W_k \right], \left[ \max_{n-n/\log n \leq k \leq n-\log n} W_k \right] \right\}. \quad (2.16)$$

Next, applying Lemma 2.16 to (2.16), we obtain

$$\max_{n-n/\log n \leq k \leq n-\log n} W_k \stackrel{D}{=} \max_{n-n/\log n \leq k \leq n-\log n} \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2. \quad (2.17)$$

Combining (2.15) through (2.17), we obtain

$$\begin{aligned} \max_{\log n \leq k \leq n/\log n} W_k \stackrel{D}{=} \max & \left\{ \max_{1 < k \leq n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2, \right. \\ & \left. \max_{n-n/\log n \leq k < n} \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 \right\}. \end{aligned}$$

Therefore,

$$\begin{aligned} & \lim_{n \rightarrow \infty} P[a(\log n)\lambda_n - b(\log n) \leq x] \\ &= \lim_{n \rightarrow \infty} P \left[ a(\log n) \max_{1 < k < n-1} \xi_k^{1/2} - b(\log n) \leq x \right] \\ &= \lim_{n \rightarrow \infty} P \left[ a^2(\log n) \max_{1 < k < n-1} \xi_k \leq [x + b(\log n)]^2 \right] \\ &= \lim_{n \rightarrow \infty} P \left[ a^2(\log n) \max_{\log n \leq k \leq n-\log n} W_k \leq [x + b(\log n)]^2 \right] \\ &= \lim_{n \rightarrow \infty} P \left[ a^2(\log n) \max \left\{ \max_{1 < k \leq n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2, \right. \right. \\ & \quad \left. \left. \max_{n-n/\log n \leq k < n} \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 \right\} \leq [x + b(\log n)]^2 \right]. \quad (2.18) \end{aligned}$$

Inasmuch as  $\{y_i, 1 \leq i \leq k, 1 \leq k \leq n/\log n\}$  and  $\{y_j, k+1 \leq j \leq n, n - n/\log n \leq k \leq n\}$  are independent, (2.18) simplifies to

$$\begin{aligned}
& \lim_{n \rightarrow \infty} P \left[ a^2(\log n) \max_{1 < k \leq n/\log n} \frac{1}{2k} \left[ \sum_{i=1}^k (y_i^2 - 1) \right]^2 \leq [x + b(\log n)]^2 \right] * \\
& \lim_{n \rightarrow \infty} P \left[ \max_{n - n/\log n \leq k < n} \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (y_i^2 - 1) \right]^2 \leq [x + b(\log n)]^2 \right] \\
& = \lim_{n \rightarrow \infty} P \left[ a(\log n) \max_{1 < k \leq n/\log n} \frac{1}{\sqrt{2k}} \left| \sum_{i=1}^k (y_i^2 - 1) \right| - b(\log n) \leq x \right] * \\
& \lim_{n \rightarrow \infty} P \left[ a(\log n) \max_{1 < k \leq n/\log n} \frac{1}{\sqrt{2(n-k)}} \left| \sum_{i=k+1}^n (y_i^2 - 1) \right| - b(\log n) \leq x \right].
\end{aligned} \tag{2.19}$$

Denote the first term of (2.19) by (a) and the second by (b). Let's consider (a) first. Let  $v_i = ((y_i^2 - 1)/\sqrt{2})$ ,  $1 \leq i < \infty$ ; we see that  $\{v_i, 1 \leq i < \infty\}$  is a sequence of iid random variables, with  $E[v_i] = 0$ ,  $\text{Var}[v_i] = 1$ , and  $E|v_i|^r < \infty$  for  $r > 2$ . Let  $S(k) = \sum_{i=1}^k v_i$ ; it is easy to see that  $k^{-1/2}S(k) = (1/2k)[\sum_{i=1}^k (y_i^2 - 1)]^2$ . Then from Lemma 2.18, as  $n \rightarrow \infty$ ,

$$a(\log n) \max_{1 < k \leq n/\log n} \frac{1}{\sqrt{2k}} \left| \sum_{i=1}^k (y_i^2 - 1) \right| - b(\log n) \xrightarrow{D} y^*,$$

where  $y^*$  has cdf  $F_{y^*}(x) = \exp\{-e^{-x}\}$ ; therefore, (a)  $= \exp\{-e^{-x}\}$ . Similarly, we can obtain: (b)  $= \exp\{-e^{-x}\}$ . Then, (2.19)  $= \exp\{-2e^{-x}\}$ , which completes the proof of the theorem.  $\square$

*Remark 2.20* In many real situations, it is more likely that  $\mu$  remains common but unknown instead of being known. Under these circumstances, the likelihood procedure can still be applied. Under  $H_0$ , the maximum log likelihood is easily obtained as

$$\log L_0(\hat{\sigma}^2, \hat{\mu}) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2},$$

where  $\hat{\sigma}^2 = (\sum_{i=1}^n (x_i - \bar{x})^2)/n$  and  $\hat{\mu} = \bar{x}$  are the MLEs of  $\sigma^2$  and  $\mu$ , respectively. Under  $H_1$ , the log likelihood function is

$$\begin{aligned} \log L_1(\mu, \sigma_1^2, \sigma_n^2) = & -\frac{n}{2} \log 2\pi - \frac{k}{2} \log \sigma_1^2 - \frac{n-k}{2} \log \sigma_n^2 \\ & - \frac{\sum_{i=1}^k (x_i - \mu)^2}{2\sigma_1^2} - \frac{\sum_{i=k+1}^n (x_i - \mu)^2}{2\sigma_n^2}, \end{aligned}$$

and the likelihood equations are:

$$\begin{cases} \sigma_n^2 \sum_{i=1}^k (x_i - \mu)^2 + \sigma_1^2 \sum_{i=k+1}^n (x_i - \mu)^2 = 0 \\ \sigma_1^2 = \frac{1}{k} \sum_{i=1}^k (x_i - \mu)^2 \\ \sigma_n^2 = \frac{1}{n-k} \sum_{i=k+1}^n (x_i - \mu)^2 \end{cases}$$

where the solutions of  $\mu, \sigma_1^2$ , and  $\sigma_n^2$  are the MLEs  $\hat{\mu}, \hat{\sigma}_1^2$ , and  $\hat{\sigma}_n^2$ , respectively. Unfortunately, solving this system of equations will not give us the closed forms for  $\hat{\mu}, \hat{\sigma}_1^2$ , and  $\hat{\sigma}_n^2$ . However, we can use Newton's iteration method, or some other iteration methods, to obtain an approximate solution. Under the regularity conditions (Dennis and Schnable, 1983), the solution will yield the unique MLE. Then the log maximum likelihood under  $H_1$  can be expressed as

$$\log L_1(\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_n^2) = -\frac{n}{2} \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{n-k}{2} \log \hat{\sigma}_n^2 - \frac{n}{2},$$

where  $\hat{\mu}, \hat{\sigma}_1^2$ , and  $\hat{\sigma}_n^2$  are the numerical solutions of the above system of equations, and  $2 \leq k \leq n-2$ .

### 2.2.2 Informational Approach

In 1973, Hirotugu Akaike introduced the Akaike Information Criterion (AIC) for model selection in statistics (Akaike, 1973). Since then, this criterion has profoundly influenced developments in statistical analysis, particularly in time series, analysis of outliers (Kitagawa, 1979), robustness, regression analysis, multivariate analysis (e.g., see Bozdogan, Sclove, and Gupta, 1994), and so on. On the basis of Akaike's work, many authors have further introduced various information criteria and used them in many other fields such as econometrics, psychometrics, control theory, and decision theory.

Suppose  $x_1, x_2, \dots, x_n$  is a sequence of independent and identically distributed random variables with probability density function  $f(\cdot)$ , where  $f$  is a model with  $K$  parameters; that is,

$$Model(K) : \{f(\cdot|\theta) : \theta = (\theta_1, \theta_2, \dots, \theta_K), \theta \in \Theta_K\}.$$

It is assumed that there are no constraints on the parameters and hence the number of free parameters in the model is  $K$ . The restricted parameter space is given by



$$\Theta_k = \{\theta \in \Theta_K | \theta_{k+1} = \theta_{k+2} = \dots = \theta_K = 0\}$$

and the corresponding model is denoted by model  $(k)$ .

To view the change point hypothesis testing of the null hypothesis given by (1.1) against the alternative hypothesis given by (1.2) in a model selection context, we target to select a “best” model from a collection of models corresponding to (1.1) and (1.2). Specifically, corresponding to the alternative hypothesis (1.2) of  $q$  change points, it is equivalent to state that:  $X_1, \dots, X_{k_1} \sim \text{iid } f(\theta_1), X_{k_1+1}, \dots, X_{k_2} \sim \text{iid } f(\theta_2), \dots, X_{k_{q-1}+1}, \dots, X_{k_q} \sim \text{iid } f(\theta_{q-1}), X_{k_q+1}, \dots, X_{k_n} \sim \text{iid } f(\theta_q)$ , where  $1 < k_1 < k_2 < \dots < k_q < n$ ,  $q$  is assumed to be the unknown number of change points and  $k_1, k_2, \dots, k_q$  are the respective unknown change point positions.

Akaike (1973) proposed the following information criterion,

$$\text{AIC}(k) = -2 \log L(\hat{\Theta}_k) + 2k, k = 1, 2, \dots, K,$$

where  $L(\hat{\Theta}_k)$  is the maximum likelihood for model  $(k)$ , as a measure of model evaluation. A model that minimizes the AIC (Minimum AIC estimate, MAICE) is considered to be the most appropriate model. However, the MAICE is not an asymptotically consistent estimator of model order (e.g., see Schwarz, 1978). Some authors made efforts to modify the information criterion without violating Akaike’s original principles. For more details of the various kinds of modifications, the reader is referred to Bozdogan (1987), Hannan and Quinn (1979), Zhao, Krishnaiah and Bai (1986a, 1986b), and Rao and Wu (1989).

One of the modifications is the Schwarz Information Criterion, denoted as SIC, and proposed by Schwarz in 1978. It is expressed as

$$\text{SIC}(k) = -2 \log L(\hat{\Theta}_k) + k \log n, k = 1, 2, \dots, K.$$

Apparently, the difference between AIC and SIC is in the penalty term, instead of  $2k$ , it is  $k \log n$ . However, SIC gives an asymptotically consistent estimate of the order of the true model. The SIC has been applied to change point analysis for different underlying models by many authors in the literature. Recently, Chen and Gupta (2003) and Pan and Chen (2006) proposed a new information criterion named the modified information criterion (MIC) for studying change point models, which demonstrated that the penalty term in SIC for change point problems should be defined according to the nature of change point problems. Here, for historical reasons, the SIC is employed to find the change point.

### (i) SICs of the Change Point Inference

According to the information criterion principle, we are going to estimate the position of the change point  $k$  by  $\hat{k}$  such that  $\text{SIC}(\hat{k})$  is the minimal. To be

specific, corresponding to the  $H_0$  defined by (2.10), is one SIC, denoted by  $\text{SIC}(n)$ , which is found as

$$\text{SIC}(n) = n \log 2\pi + n \log \hat{\sigma}^2 + n + \log n, \quad (2.20)$$

where  $\hat{\sigma}^2 = (\sum_{i=1}^n (x_i - \mu)^2)/n$  is the MLE of  $\sigma^2$  under  $H_0$ . Corresponding to the  $H_1$  defined by (2.11), are the  $n - 3$  SICs, denoted by  $\text{SIC}(k)$  for  $2 \leq k \leq n - 2$ , which are found as

$$\text{SIC}(k) = n \log 2\pi + k \log \hat{\sigma}_1^2 + (n - k) \log \hat{\sigma}_n^2 + n + 2 \log n, \quad (2.21)$$

where  $\hat{\sigma}_1^2 = (\sum_{i=1}^k (x_i - \mu)^2)/k$  and  $\hat{\sigma}_n^2 = ((\sum_{i=k+1}^n (x_i - \mu)^2)/(n - k))$  are the MLEs of  $\sigma_1^2$  and  $\sigma_n^2$ , respectively, under  $H_1$ .

Notice that to be able to obtain the MLEs, we can only detect changes that are located between the second and  $(n - 2)$  positions. According to the information criterion principle, we accept  $H_0$  if

$$\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k),$$

and accept  $H_1$  if

$$\text{SIC}(n) > \text{SIC}(k)$$

for some  $k$ , and estimate the position of the change point by  $\hat{k}$  such that

$$\text{SIC}(\hat{k}) = \min_{2 \leq k \leq n-2} \text{SIC}(k). \quad (2.22)$$

On the one hand, we point out (see Gupta and Chen, 1996) that information criteria, such as SIC, provide a remarkable way for exploratory data analysis with no need to resort to either the distribution or the significance level  $\alpha$ . However, when the SICs are very close, one may question that the small difference among the SICs might be caused by the fluctuation of the data, and therefore there may be no change at all. To make the conclusion about change point statistically convincing, we introduce the significance level  $\alpha$  and its associated critical value  $c_\alpha$ , where  $c_\alpha \geq 0$ . Instead of accepting  $H_0$  when  $\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k)$ , we now accept  $H_0$  if

$$\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k) + c_\alpha,$$

where  $c_\alpha$  is determined from

$$1 - \alpha = P[\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k) + c_\alpha | H_0 \text{ holds}].$$

By using Theorem 2.19, the approximate  $c_\alpha$  values can be obtained as follows.

$$\begin{aligned}
1 - \alpha &= P[\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k) + c_\alpha | H_0 \text{ holds}] \\
&= P[\lambda_n^2 < \log n + c_\alpha | H_0 \text{ holds}] \\
&= P[0 < \lambda_n < (\log n + c_\alpha)^{1/2} | H_0 \text{ holds}] \\
&= P[-b(\log n) < a(\log n)\lambda_n - b(\log n) \\
&\quad < a(\log n)(\log n + c_\alpha)^{1/2} - b(\log n) | H_0 \text{ holds}] \\
&\cong \exp\{-2 \exp[-a(\log n)(\log n + c_\alpha)^{1/2} + b(\log n)]\} \\
&\quad - \exp\{-2 \exp[b(\log n)]\},
\end{aligned}$$

and solving for  $c_\alpha$ , we obtain:

$$\begin{aligned}
c_\alpha &\cong \left\{ -\frac{1}{a(\log n)} \log \log[1 - \alpha + \exp(-2e^{b(\log n)})]^{-1/2} + \frac{b(\log n)}{a(\log n)} \right\}^2 \\
&\quad - \log n.
\end{aligned} \tag{2.23}$$

For a different significance level  $\alpha$  ( $\alpha = 0.01, 0.025, 0.05, 0.1$ ), and sample sizes  $n$  ( $n = 13, 14, \dots, 200$ ), the approximate values of  $c_\alpha$  have been calculated according to (2.23) and tabulated in [Table 2.3](#).

## (ii) Unbiased SICs

To derive the information criterion AIC, Akaike (1973) used  $\log L(\hat{\theta})$  as an estimate of  $J = E_{\hat{\theta}}[\int f(\mathbf{y}|\theta_0) \log f(\mathbf{y}|\hat{\theta}) d\mathbf{y}]$ , where  $f(\mathbf{y}|\theta_0)$  is the probability density of the future observations  $\mathbf{y} = (y_1, \dots, y_n)$  of the same size and distribution as the  $x$ s,  $\mathbf{x} = (x_1, \dots, x_n)$ , and  $x$  and  $y$  are independent. The expectation is taken under the distribution of  $x$  when  $H_0$  is true; that is,  $\theta_0 \in \Theta_{H_0}$ . Unfortunately,  $\log L(\hat{\theta})$  is not an unbiased estimator of  $J$ . When the sample size  $n$  is finite, Sugiura (1978) proposed unbiased versions, the finite corrections of AIC, for different model selection problems.

In this section, we derive the unbiased versions of our SIC under our  $H_0$  defined by (2.10) and  $H_1$  defined by (2.11), denoted by  $u - \text{SIC}(n)$ , and  $u - \text{SIC}(k)$ , respectively.

### (1) Unbiased SIC under $H_0 : u - \text{SIC}(n)$

Under  $H_0$ , let  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  be a sample of the same size and distributions as  $\mathbf{x}$ ,  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , and that  $\mathbf{y}$  be independent of  $\mathbf{x}$ .

**Table 2.3** Approximate Critical Values of SIC

$n/\alpha$	0.010	0.025	0.050	0.100
13	20.927	14.570	10.496	6.946
14	20.431	14.340	10.375	6.895
15	20.077	14.165	10.279	6.852
16	19.807	14.023	10.199	6.816
17	19.589	13.903	10.130	6.783
18	19.405	13.799	10.068	6.753
19	19.247	13.706	10.012	6.725
20	19.106	13.623	9.961	6.698
21	18.980	13.546	9.914	6.673
22	18.860	13.476	9.870	6.649
23	18.759	13.411	9.829	6.626
24	18.661	13.350	9.790	6.605
25	18.569	13.293	9.753	6.583
26	18.484	13.239	9.718	6.563
27	18.404	13.188	9.685	6.543
28	18.328	13.140	9.653	6.524
29	18.257	13.094	9.622	6.506
30	18.189	13.050	9.593	6.488
35	17.895	12.858	9.463	6.406
40	17.656	12.699	9.352	6.333
45	17.456	12.564	9.256	6.268
50	17.284	12.446	9.171	6.208
55	17.134	12.342	9.095	6.154
60	17.001	12.249	9.026	6.104
70	16.773	12.088	8.904	6.014
80	16.584	11.951	8.800	5.934
90	16.422	11.832	8.708	5.863
100	16.280	11.728	8.626	5.799
120	16.043	11.550	8.484	5.686
140	15.848	11.402	8.365	5.589
160	15.684	11.276	8.261	5.504
180	15.542	11.165	8.170	5.428
200	15.416	11.067	8.088	5.359

$$\begin{aligned}
J &= E_{\hat{\theta}} \left[ \int f(\mathbf{y}|\theta_0) \log f(\mathbf{y}|\hat{\theta}) d\mathbf{y} \right] \\
&= E_{\hat{\theta}} \left[ E_y \left\{ -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (y_i - \mu)^2 \right\} \right] \\
&= E_{\hat{\theta}} \left[ -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2} + \frac{n}{2} - E_y \left\{ \frac{1}{2\hat{\sigma}^2} \sum_{i=1}^n (y_i - \mu)^2 \right\} \right],
\end{aligned}$$

where  $\hat{\sigma}^2 = \sum_{i=1}^n (x_i - \mu)^2 / n$ , and  $\mu$  is known. Because  $\sum_{i=1}^n (y_i - \mu)^2 \sim \sigma^2 \chi_n^2$ , that is,  $n\hat{\sigma}^2 / \sigma^2 \sim \chi_n^2$ , we get:

$$\begin{aligned}
J &= E_{\hat{\theta}} \left[ \log L_0(\hat{\sigma}^2) + \frac{n}{2} - \frac{n\sigma^2}{2\hat{\sigma}^2} \right] \\
&= E_{\hat{\theta}}[\log L_0(\hat{\sigma}^2)] + \frac{n}{2} - \frac{n^2}{2} \frac{1}{n-2} \\
&= E_{\hat{\theta}}[\log L_0(\hat{\sigma}^2)] - \frac{n}{n-2}.
\end{aligned}$$

Clearly,  $-2 \log L_0(\hat{\sigma}^2) + 2n/(n-2)$  is unbiased for  $-2J$ . Therefore, the unbiased  $u - \text{SIC}(n)$  is obtained as

$$\begin{aligned}
u - \text{SIC}(n) &= -2 \log L_0(\hat{\sigma}^2) + \frac{2n}{n-2} \\
&= \text{SIC}(n) + \frac{2n}{n-2} - \log n.
\end{aligned}$$

## (2) Unbiased SIC under $H_1 : u - \text{SIC}(k)$

Under  $H_1$ , let  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  be a sample of the same size and distributions as  $\mathbf{x}$ ,  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , and that  $\mathbf{y}$  be independent of  $\mathbf{x}$ . That is,  $y_1, y, \dots, y_k$  are iid  $N(\mu, \sigma_1^2)$ , and  $y_{k+1}, y_{k+2}, \dots, y_n$  are iid  $N(\mu, \sigma_n^2)$ .

$$\begin{aligned}
J &= E_{\hat{\theta}}[E_y\{\log L_1(\hat{\sigma}_1^2, \hat{\sigma}_n^2, Y)\}] \\
&= E_{\hat{\theta}} \left[ E_y \left\{ -\frac{n}{2} \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{n-k}{2} \log \hat{\sigma}_n^2 \right. \right. \\
&\quad \left. \left. - \frac{1}{2\hat{\sigma}_1^2} \sum_{i=1}^k (y_i - \mu)^2 - \frac{1}{2\hat{\sigma}_n^2} \sum_{i=k+1}^n (y_i - \mu)^2 \right\} \right],
\end{aligned}$$

where

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^k (x_i - \mu)^2}{k}, \quad \hat{\sigma}_n^2 = \frac{\sum_{i=k+1}^n (x_i - \mu)^2}{n-k},$$

and  $\mu$  is known.

$\sum_{i=1}^k (y_i - \mu)^2 \sim \sigma_1^2 \chi_k^2$ ,  $\sum_{i=k+1}^n (y_i - \mu)^2 \sim \sigma_n^2 \chi_{n-k}^2$ ,  $k\hat{\sigma}_1^2/\sigma_1^2 \sim \chi_k^2$ , and  $(n-k)\hat{\sigma}_n^2/\sigma_n^2 \sim \chi_{n-k}^2$  therefore we get

$$\begin{aligned}
J &= E_{\hat{\theta}}[\log L_1(\hat{\sigma}_1^2, \hat{\sigma}_n^2)] + \frac{n}{2} - E_{\hat{\theta}} \left[ \frac{k\sigma_1^2}{2\hat{\sigma}_1^2} + \frac{(n-k)\sigma_n^2}{2\hat{\sigma}_n^2} \right] \\
&= E_{\hat{\theta}}[\log L_1(\hat{\sigma}_1^2, \hat{\sigma}_n^2)] + \frac{n}{2} - \frac{k^2}{2} \frac{1}{k-2} - \frac{(n-k)^2}{2} \frac{1}{n-k-2} \\
&= E_{\hat{\theta}}[\log L_1(\hat{\sigma}_1^2, \hat{\sigma}_n^2)] - \frac{2(nk - k^2 - n)}{(k-2)(n-k-2)}.
\end{aligned}$$

Hence, the unbiased SIC under  $H_1$  is

$$u - \text{SIC}(k) = -2 \log L_1(\hat{\sigma}_1^2, \hat{\sigma}_n^2) + \frac{4(nk - k^2 - n)}{(k-2)(n-k-2)},$$

for  $2 \leq k \leq n-2$ .

### 2.2.3 Other Methods

In addition to the likelihood procedure and information approach to the variance change point problem, there are several other methods available in the literature; see Hsu (1977), Davis (1979), and Abraham and Wei (1984) for more details. Here, a Bayesian approach based on the work of Inclán (1993) is presented.

Let  $x_1, x_2, \dots, x_n$  be a sequence of independent normal random variables with parameters  $(0, \sigma_1^2), (0, \sigma_2^2), \dots, (0, \sigma_n^2)$ , respectively. It is desired to test the hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_n^2 = \sigma^2 \text{ (unknown)},$$

against the alternative:

$$H_1 : \sigma_1^2 = \dots = \sigma_{k_1}^2 = \eta_0^2 \neq \sigma_{k_1+1}^2 = \dots = \sigma_{k_2}^2 = \eta_1^2 \neq \dots \\ \neq \sigma_{k_{q-1}+1}^2 = \dots = \sigma_{k_q}^2 = \eta_{q-1}^2 \neq \sigma_{k_q+1}^2 = \dots = \sigma_n^2 = \eta_q^2,$$

where  $q$  is the unknown number of change points, and  $1 < k_1 < k_2 < \dots < k_q < n$ , are the unknown positions of the change points, respectively.

Let  $K_{r,m}$  denote the posterior odds of  $r$  changes versus  $m$  changes. A systematic way of using the posterior odds to determine  $q$  is to calculate  $K_{r,r-1}$  for  $r = 1, 2, \dots, n$ . Starting with  $r = 1$ , where  $K_{1,0}$  means one change versus no change, if  $K_{1,0} > 1$ , then there is at least one change. Next, compute  $K_{2,1}$ ; if  $K_{2,1} > 1$ ; then there are at least two changes. Keep calculating  $K_{r,r-1}$  as long as  $K_{r,r-1} > 1$ . If  $K_{r+1,r} \leq 1$ , stop the process and conclude that there are  $r$  changes and estimate  $q$  by  $\hat{q} = r$ .

In the following, the derivation of  $K_{r,m}$  is given. Let  $k_0 = 0, k_{q+1} = n$ ; then there are  $d_j = k_{j+1} - k_j$  observations with variances  $\eta_j^2, j = 0, 1, \dots, q$ . Let  $\sigma = (\sigma_1, \dots, \sigma_q)'$ ,  $\eta = (\eta_0, \dots, \eta_q)'$ ,  $\mathbf{k} = (k_1, \dots, k_q)'$ , and  $\mathbf{x} = (x_1, \dots, x_n)'$ . The joint density of  $x$  given  $\eta, \mathbf{k}, q$  can be written as

$$\begin{aligned}
f(\mathbf{x}|\eta, \mathbf{k}, q) &= \frac{1}{(2\pi)^{n/2}} \prod_{i=1}^n \sigma_i^{-1} \exp \left\{ -\frac{1}{2\sigma_i^2} x_i^2 \right\} \\
&= \frac{1}{(2\pi)^{n/2}} \prod_{j=0}^q \eta_j^{-d_j} \exp \left\{ -\frac{1}{2\eta_j^2} \sum_{i=k_j+1}^{k_{j+1}} x_i^2 \right\}. \quad (2.24)
\end{aligned}$$

Now let the prior distributions be as follows,

$$q|\theta \sim \text{Binomial}(n-1, \theta),$$

where  $\theta$  is the probability of observing a change; that is,

$$f(q|\theta) = \binom{n-1}{q} \theta^q (1-\theta)^{n-1-q}, \quad q = 0, 1, \dots, n-1.$$

Assume that  $\eta_j$ 's are conditionally independent drawn from the inverted gamma density and independent of  $q$  and  $\mathbf{k}$ :

$$f(\eta_j|c, \nu) = \frac{2c^{\nu/2}}{\Gamma(\nu/2)} \eta_j^{-(\nu+1)} \exp(-c\eta_j^2), \quad (2.25)$$

where  $0 < \eta_j < \infty$ ,  $j = 0, \dots, q$ . Assume  $k_1 < k_2 < \dots < k_q$  are equally likely:

$$f(\mathbf{k}|q) = \frac{1}{\binom{n-1}{q}}. \quad (2.26)$$

Then from (2.22) through (2.24) the joint probability density function of  $\mathbf{x}$ ,  $\eta$ ,  $\mathbf{k}$  given  $q$  is obtained as

$$\begin{aligned}
f(\mathbf{x}, \eta, \mathbf{k}|q) &= f(\mathbf{k}|q) f(\eta|\mathbf{k}, q) f(\mathbf{x}|\eta, \mathbf{k}, q) \\
&= f(\mathbf{k}, q) f(\eta|c, \nu) f(\mathbf{x}|\eta, \mathbf{k}, q) \\
&= \frac{(2\pi)^{-n/2}}{\binom{n-1}{q}} \left( \frac{2c^{\nu/2}}{\Gamma(\nu/2)} \right)^{q+1} \\
&\quad \cdot \prod_{j=0}^q \left( \eta_j^{-(d_j+\nu+1)} \exp \left\{ -\frac{1}{2\eta_j^2} \left[ \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + 2c \right] \right\} \right). \quad (2.27)
\end{aligned}$$

Inasmuch as

$$\begin{aligned} & \int_0^\infty \eta_j^{-(d_j+\nu+1)} \exp \left\{ -\frac{1}{2\eta_j^2} \left[ \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + 2c \right] \right\} d\eta_j \\ &= \frac{\Gamma\left(\frac{d_j+\nu}{2}\right)}{2 \left[ \frac{1}{2} \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + c \right]^{(d_j+\nu)/2}}, \end{aligned}$$

and  $\eta_j$ 's are independent,

$$\begin{aligned} f(\mathbf{x}, \mathbf{k}|q) &= \int \cdots \int f(\mathbf{x}, \eta, \mathbf{k}|q) d\eta \\ &= \frac{(2\pi)^{-n/2}}{\binom{n-1}{q}} \left( \frac{c^{\nu/2}}{\Gamma(\nu/2)} \right)^{q+1} \\ &\quad \cdot \prod_{j=0}^q \left\{ \Gamma\left(\frac{d_j+\nu}{2}\right) \left[ \frac{1}{2} \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + c \right]^{-((d_j+\nu)/2)} \right\}. \end{aligned} \quad (2.28)$$

Now,

$$f(\mathbf{x}|q) = \sum_{k_1} \cdots \sum_{k_q} f(\mathbf{x}, \mathbf{k}|q), \quad (2.29)$$

where the sums are over all possible values of  $\mathbf{k}$  :  $k_1 = 1, 2, \dots, n-q$ ;  $k_2 = k_1 + 1, \dots, n-q+1$ ;  $\dots$ ;  $k_{j+1} = k_j + 1, \dots, n-q+j$ ;  $\dots$ ; and  $k_q = k_{q-1} + 1, \dots, n-1$ .

Therefore,

$$f(\mathbf{k}|\mathbf{x}, q) = \frac{f(\mathbf{x}, \mathbf{k}|q)}{f(\mathbf{x}|q)}. \quad (2.30)$$

Because

$$f(q|\mathbf{x}) = \frac{f(q)f(\mathbf{x}|q)}{f(\mathbf{x})} \propto f(q)f(\mathbf{x}|q),$$

the posterior odds  $K_{r,m}$  are given by

$$\begin{aligned} K_{r,m} &= \frac{P(q=r|\mathbf{x})}{P(q=m|\mathbf{x})} \\ &= \frac{P(q=r)f(\mathbf{x}|q=r)}{P(q=m)f(\mathbf{x}|q=m)} \\ &= \frac{\binom{n-1}{r} \theta^r (1-\theta)^{n-1-r}}{\binom{n-1}{m} \theta^m (1-\theta)^{n-1-m}} \cdot \frac{\sum_{k_1} \cdots \sum_{k_r} f(\mathbf{x}, \mathbf{k}|q=r)}{\sum_{k_1} \cdots \sum_{k_m} f(\mathbf{x}, \mathbf{k}|q=m)} \end{aligned}$$



$$\begin{aligned}
&= \frac{\binom{n-1}{r}}{\binom{n-1}{m}} \left( \frac{\theta}{1-\theta} \right)^{r-m} \cdot \frac{\sum_{k_1} \cdots \sum_{k_r} \frac{(2\pi)^{-n/2}}{\binom{n-1}{r}} \left( \frac{c^{\nu/2}}{\Gamma(\nu/2)} \right)^{r+1}}{\sum_{k_1} \cdots \sum_{k_m} \frac{(2\pi)^{-n/2}}{\binom{n-1}{m}} \left( \frac{c^{\nu/2}}{\Gamma(\nu/2)} \right)^{m+1}} \\
&\quad \frac{\prod_{j=0}^r \left\{ \Gamma\left(\frac{d_j+\nu}{2}\right) \left[ \frac{1}{2} \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + c \right]^{-((d_j+\nu)/2)} \right\}}{\prod_{j=0}^m \left\{ \Gamma\left(\frac{d_j+\nu}{2}\right) \left[ \frac{1}{2} \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + c \right]^{-((d_j+\nu)/2)} \right\}} \\
&= \left( \frac{\theta}{1-\theta} \right)^{r-m} \left( \frac{c^{\nu/2}}{\Gamma(\nu/2)} \right)^{r-m} \\
&\quad \cdot \frac{\sum_{k_1} \cdots \sum_{k_r} \prod_{j=0}^r \left\{ \Gamma\left(\frac{d_j+\nu}{2}\right) \left[ \frac{1}{2} \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + c \right]^{-((d_j+\nu)/2)} \right\}}{\sum_{k_1} \cdots \sum_{k_m} \prod_{j=0}^m \left\{ \Gamma\left(\frac{d_j+\nu}{2}\right) \left[ \frac{1}{2} \sum_{i=k_j+1}^{k_{j+1}} x_i^2 + c \right]^{-((d_j+\nu)/2)} \right\}}. \quad (2.31)
\end{aligned}$$

It may be noted that  $\lambda, c, \nu$  are hyperparameters that will not be modeled. However, the values of  $\lambda, c, \nu$  can be assigned thoughtfully according to experience. A typical assignment is:  $\lambda = 1/n, \nu = 1$  or  $\nu = 2$ , and  $c = (\nu + 1)/2$ . For a discussion of this matter, the reader is referred to Inclán (1993).

After estimating the number of change points  $q$ , the next step is to locate the change points. One way to do it is to obtain the posterior pdf of  $\mathbf{k}$  given  $\mathbf{x}$  and  $q = \hat{q}$ . From (2.28),

$$f(\mathbf{k}|\mathbf{x}, q = \hat{q}) = \frac{f(\mathbf{x}, \mathbf{k}|q = \hat{q})}{f(\mathbf{x}|q = \hat{q})}.$$

Then, obtain the marginal distributions of each  $k_j$ , for  $j = 1, 2, \dots, \hat{q}$ . Finally, the joint mode ( $\text{mode}(k_1), \text{mode}(k_2), \dots, \text{mode}(k_{\hat{q}})$ ) gives the locations of the change points; that is,

$$\hat{k}_1 = \text{mode}(k_1), \hat{k}_2 = \text{mode}(k_2), \dots, \hat{k}_{\hat{q}} = \text{mode}(k_{\hat{q}}).$$

### 2.2.4 Application to Stock Market Data

We give an application of the SIC test procedure to searching a change point in stock prices (Chen and Gupta, 1997). Hsu (1977) analyzed the U.S. stock market return series during the period 1971–1974 using  $T$ - and  $G$ -statistics, and found that there was one variance change point which is suspected to have occurred in conjunction with the Watergate events. Later, he (Hsu,

1979) reanalyzed the stock market return series data by considering a gamma sequence and came up with the same conclusion.

Here we take the same stock market price data as in Hsu (1979), and perform the change point analysis by using the SIC procedure. Let  $P_t$  be the stock price; we first transform the data into  $R_t = (P_{t+1} - P_t)/P_t$ ,  $t = 1, \dots, 161$ . According to Hsu (1977),  $\{R_t\}$  is a sequence of independent normal random variables with mean zero. We then test the following hypothesis based on the  $R_t$  series,

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_{161}^2 = \sigma^2(\text{unknown}),$$

versus the alternative:

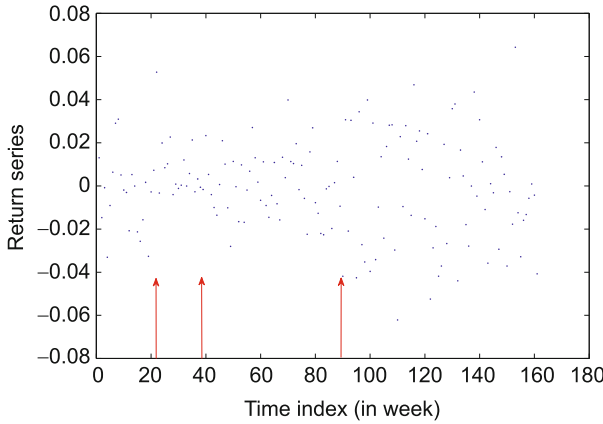
$$H_1 : \sigma_1^2 = \dots = \sigma_{k_1}^2 \neq \sigma_{k_1+1}^2 = \dots = \sigma_{k_2}^2 \neq \dots \neq \sigma_{k_q+1}^2 = \dots = \sigma_{161}^2,$$

where  $q$  is the unknown number of change points, and  $1 \leq k_1 < k_2 < \dots < k_q < 161$ , are the unknown positions of the change points, respectively.

Using the binary segmentation procedure along with the SIC, we are able to detect all the changes in the  $R_t$  series. According to our computations, at the first stage  $\min_{1 < k < 160} \text{SIC}(k) = \text{SIC}(89) = -787.5745 < \text{SIC}(161) = -765.6242$ . If we use the  $c_\alpha$  in Tables 2.3–2.5, we still have  $\text{SIC}(89) + c_\alpha < \text{SIC}(161)$ . Hence,  $t = 89$  is a variance change point for the  $R_t$  series. Transferring to the price  $P_t$ ,  $t + 1 = 90$  is the location of the variance change point. In other words, the stock price started to change at the 91st time point, which corresponds to the calendar week of March 19–23, 1973. Our conclusion matches Hsu's (1977, 1979) conclusion at this point.

Moreover, we continue to test the two subsequences:  $t$  from 1 to 88, and  $t$  from 89 to 161. Our computational results show that there are no further changes in the subsequence of  $t$  from 89 to 161, but there are at least two more changes in the subsequence of  $t$  from 1 to 88. One of the changes occurred during the period July 19 to August 8, 1971, and the other change occurred during the period November 15 to December 12, 1971. Going back to some historical records (e.g., Leonard, Crippen and Aronson, 1988), and looking at what happened to the U.S. economy and environment during those two periods we find that: from July 19 to August 8, 1971, several union strikes influenced the changes of the U.S. stock markets. The wage increases, resulting from several union–company negotiations, caused grave concerns about market prices. Among those strikes, the one organized by the United Transportation Union on July 26 was the biggest, and the negotiations were suspended indefinitely over a dispute on work rule changes. Responding to the suspension, U.S. gold stocks fell dramatically. The rate of wage increases in steel companies, the U.S. Postal Service, and some others was as high as 30 percent. On August 4, President Nixon said he would consider establishment of wage–price review bonds to examine the situation of American markets.

From November 15 to December 12, the most eye-catching economic event was the price increases of some important industrial products. Although the



**Fig. 2.12** Return series  $R_t$  of the weekly stock prices from 1971 to 1974

Nixon administration established the price commission to stabilize prices, economic conditions forced the administration to approve price increases. For example, the price commission approved steel price increases that were about triple the commission's target. Also, the commission approved price increases for the Big Three automakers, averaging nearly 3 percent.

The scatterplot of the return series  $R_t$  is given in Figure 2.12 with the identified changes indicated by arrows.

## 2.3 Mean and Variance Change

Let  $x_1, x_2, \dots, x_n$  be a sequence of independent normal random variables with parameters  $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \dots, (\mu_n, \sigma_n^2)$ , respectively. Testing and estimation about multiple mean changes in a Gaussian model have been studied in Section 2.1, and inference about multiple variance changes has been studied in Section 2.2. In this section, inference about the multiple mean and variance changes is discussed. To be specific, the interest here is to test the hypothesis (Chen and Gupta, 1999):

$$H_0 : \mu_1 = \dots = \mu_n = \mu \quad \text{and} \quad \sigma_1^2 = \dots = \sigma_n^2 = \sigma^2 (\mu, \sigma^2 \text{ unknown}) \quad (2.32)$$

versus the alternative:

$$H_A : \mu_1 = \dots = \mu_{k_1} \neq \mu_{k_1+1} = \dots = \mu_{k_2} \neq \dots \neq \mu_{k_q+1} = \dots = \mu_n$$

and

$$\sigma_1^2 = \dots = \sigma_{k_1}^2 \neq \sigma_{k_1+1}^2 = \dots = \sigma_{k_2}^2 \neq \dots \neq \sigma_{k_q+1}^2 = \dots = \sigma_n^2.$$

As discussed in previous sections, the binary segmentation procedure can be applied to this situation. Then it suffices to test (2.30) versus the alternative:

$$H_1 : \mu_1 = \cdots = \mu_k \neq \mu_{k+1} = \cdots = \mu_n$$

and

$$\sigma_1^2 = \cdots = \sigma_k^2 \neq \sigma_{k+1}^2 = \cdots = \sigma_n^2. \quad (2.33)$$

### 2.3.1 Likelihood-Ratio Procedure

Under  $H_0$ , the log likelihood function is

$$\log L_0(\mu, \sigma^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Denote the MLEs of  $\mu$  and  $\sigma^2$  by  $\hat{\mu}$  and  $\hat{\sigma}^2$ ; then

$$\begin{aligned} \hat{\mu} &= \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \end{aligned}$$

and the maximum log likelihood is:

$$\log L_0(\hat{\mu}, \hat{\sigma}^2) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2}.$$

Under  $H_1$ , the log likelihood function is:

$$\begin{aligned} \log L_1(\mu_1, \mu_n, \sigma_1^2, \sigma_n^2) &= -\frac{n}{2} \log 2\pi - \frac{k}{2} \log \sigma_1^2 - \frac{(n-k)}{2} \log \sigma_n^2 \\ &\quad - \frac{1}{2\sigma_1^2} \sum_{i=1}^k (x_i - \mu_1)^2 - \frac{1}{2\sigma_n^2} \sum_{i=k+1}^n (x_i - \mu_n)^2. \end{aligned}$$

Let  $\hat{\mu}_1$ ,  $\hat{\mu}_n$ ,  $\hat{\sigma}_1^2$ , and  $\hat{\sigma}_n^2$  be the MLEs under  $H_1$  of  $\mu_1, \mu_n, \sigma_1^2$ , and  $\sigma_n^2$ , respectively. Then

$$\begin{aligned} \hat{\mu}_1 &= \bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i, & \hat{\sigma}_1^2 &= \frac{1}{k} \sum_{i=1}^k (x_i - \bar{x}_k)^2, \\ \hat{\mu}_n &= \bar{x}_{n-k}, & \hat{\sigma}_n^2 &= \frac{1}{n-k} \sum_{i=k+1}^n (x_i - \bar{x}_{n-k})^2, \end{aligned}$$

and the maximum log likelihood is

$$\log L_1(\hat{\mu}_1, \hat{\mu}_n, \hat{\sigma}_1^2, \hat{\sigma}_n^2) = -\frac{n}{2} \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{n-k}{2} \log \hat{\sigma}_n^2 - \frac{n}{2}.$$

The likelihood-ratio procedure (Lehmann, 1986, p. 16) statistic is

$$A_n = \max_{2 \leq k \leq n-2} \frac{\hat{\sigma}^n}{\hat{\sigma}_1^k \hat{\sigma}_n^{n-k}}.$$

Horváth (1993) derived the asymptotic null distribution of a function of  $A_n$ . The exact null distribution of  $A_n$  is not yet available in the literature. Therefore, in the following, Horváth's main theorem is presented and its detailed proof is given.

For large  $n$ , the asymptotic null distribution of  $\lambda_n$ , where

$$\lambda_n = (2 \log A_n)^{1/2} = \left[ \max_{2 \leq k \leq n-2} (n \log \hat{\sigma}^2 - k \log \hat{\sigma}_k^2 - (n-k) \log \hat{\sigma}_{n-k}^2) \right]^{1/2}$$

is derived. However, it is convenient to simplify  $\lambda_n$  and to prove the following results first. Throughout, “ $\stackrel{AD}{=}$ ” means “asymptotically distributed as”. Under  $H_0$ ;

$$\begin{aligned} \lambda_n^2 &= \max_{1 < k < n-1} \left[ n \log \frac{\hat{\sigma}^2}{\sigma^2} - k \log \frac{\hat{\sigma}_k^2}{\sigma^2} - (n-k) \log \frac{\hat{\sigma}_{n-k}^2}{\sigma^2} \right] \\ &\stackrel{D}{=} \max_{1 < k < n-1} \left[ n \log \frac{1}{n} \chi_{n-1}^2 - k \log \frac{1}{k} \chi_{k-1}^2 - (n-k) \log \frac{1}{n-k} \chi_{n-k-1}^2 \right], \end{aligned}$$

where  $\chi_j^2$  denote the chi-square random variable with  $j$  degrees of freedom.

Let

$$\begin{aligned} \chi_{n-1}^2 &= \sum_{i=1}^n (z_i - \bar{z})^2, \\ \chi_{k-1}^2 &= \sum_{i=1}^k (z_i - \bar{z}_k)^2, \\ \chi_{n-k-1}^2 &= \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2, \end{aligned}$$

where  $z_1, \dots, z_n$  are iid  $N(0, 1)$  random variables, and

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i, \quad \bar{z}_k = \frac{1}{k} \sum_{i=1}^k z_i \quad \text{and} \quad \bar{z}_{n-k} = \frac{1}{n-k} \sum_{i=k+1}^n z_i.$$

Then

$$\lambda_n^2 \stackrel{D}{=} \max_{1 < k < n-1} \left[ n \log \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 - k \log \frac{1}{k} \sum_{i=1}^k (z_i - \bar{z}_k)^2 \right. \\ \left. - (n-k) \log \frac{1}{n-k} \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2 \right].$$

Let

$$\xi_k = n \log \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 - k \log \frac{1}{k} \sum_{i=1}^k (z_i - \bar{z}_k)^2 \\ - (n-k) \log \frac{1}{n-k} \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2.$$

Using the three-term Taylor expansion, we have

$$\xi_k = n \left[ \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 - 1 \right] - \frac{n}{2} \left[ \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 - 1 \right]^2 \\ + \frac{n}{3} (Q_n^{(1)})^{-3} \left[ \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 - 1 \right]^3 - k \left[ \frac{1}{k} \sum_{i=1}^k (z_i - \bar{z}_k)^2 - 1 \right] \\ + \frac{k}{2} \left[ \frac{1}{k} \sum_{i=1}^k (z_i - \bar{z}_k)^2 - 1 \right]^2 - \frac{k}{3} (Q_k^{(2)})^{-2} \left[ \frac{1}{k} \sum_{i=1}^k (z_i - \bar{z}_k)^2 - 1 \right]^3 \\ - (n-k) \left[ \frac{1}{n-k} \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2 - 1 \right] \\ + \frac{n-k}{2} \left[ \frac{1}{n-k} \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2 - 1 \right]^2 \\ - \frac{n-k}{3} (Q_{n-k}^{(2)})^{-3} \left[ \frac{1}{n-k} \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2 - 1 \right]^3,$$

where  $|Q_n^{(1)} - 1| \leq |(1/n) \sum_{i=1}^n (z_i - \bar{z})^2 - 1|$ ,  $|Q_k^{(2)} - 1| \leq |(1/k) \sum_{i=1}^k (z_i - \bar{z}_k)^2 - 1|$ , and  $|Q_{n-k}^{(2)} - 1| \leq |(1/(n-k)) \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2 - 1|$ . After some algebraic simplification,

$$\begin{aligned}
\xi_k &= k\bar{z}_k^2 + (n-k)\bar{z}_{n-k}^2 - n\bar{z}^2 - \frac{1}{2n} \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 \\
&\quad + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 + \bar{z}^2 \sum_{i=1}^n (z_i^2 - 1) - \frac{n}{2} \bar{z}^4 - \bar{z}_k^2 \sum_{i=1}^k (z_i^2 - 1) \\
&\quad + \frac{k}{2} \bar{z}_k^4 + \frac{n}{3} (Q_n^{(1)})^{-3} (t_{n-1} - 1)^3 - \frac{k}{3} (Q_k^{(2)})^{-3} (t_{k-1})^3 + \frac{n-k}{2} \bar{z}_{n-k}^4 \\
&\quad - \bar{z}_{n-k}^2 \sum_{i=k+1}^n (z_i^2 - 1) + \frac{n-k}{3} (Q_{n-k}^{(2)})^{-3} (t_{n-k-1} - 1)^3 \\
&= W_k^{(1)} + W_k^{(2)} + Q_k^{(1)} + Q_k^{(2)},
\end{aligned}$$

where

$$\begin{aligned}
W_k^{(1)} &= k\bar{z}_k^2 + (n-k)\bar{z}_{n-k}^2 - n\bar{z}^2, \\
W_k^{(2)} &= -\frac{1}{2n} \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2, \\
Q_k^{(1)} &= \bar{z} \sum_{i=1}^n (z_i^2 - 1) - \frac{n}{2} \bar{z}^4 - \bar{z}_k^2 \sum_{i=1}^k (z_i^2 - 1) + \frac{k}{2} \bar{z}_k^4 \\
&\quad + \frac{n}{3} (\theta_n^{(1)})^{-3} (t_{n-1} - 1)^3 - \frac{k}{3} (\theta_k^{(2)})^{-3} (t_{k-1} - 1)^3, \\
Q_k^{(2)} &= \frac{n-k}{2} \bar{z}_{n-k}^4 - \bar{z}_{n-k}^2 \sum_{i=k+1}^n (z_i^2 - 1) + \frac{n-k}{3} (\theta_{n-k}^{(3)})^{-3} (t_{n-k-1} - 1)^3, \\
t_{n-1} &= \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2, \\
t_{k-1} &= \frac{1}{k} \sum_{i=1}^k (z_i - \bar{z}_k)^2, \quad \text{and} \\
t_{n-k-1} &= \frac{1}{n-k} \sum_{i=k+1}^n (z_i - \bar{z}_{n-k})^2.
\end{aligned}$$

Next, we propose the following lemmas for the properties of the above-listed quantities.

**Lemma 2.21**

- (i)  $\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} |Q_k^{(1)}| = O_p(1)$ .  
(ii)  $\max_{1 < k < n} (n - k)^{1/2} [\log \log(n - k)]^{-(3/2)} |Q_k^{(2)}| = O_p(1)$ .

*Proof.* (i) Because  $kt_{k-1} = \sum_{i=1}^k (z_i^{(j)} - \bar{z}_k^{(j)})^2 \stackrel{AD}{=} \chi_k^2$  for large  $k$ ,  $E(kt_{k-1}) = k$ , or  $E[k(t_{k-1} - 1)] = 0$ , and  $\text{Var}[k(t_{k-1} - 1)] = 2k$ . From the law of iterated logarithm,

$$\max_{1 < k < n} \frac{|(Q_k^{(2)})^{-1} k(t_{k-1} - 1)|}{(k \log \log k)^{1/2}} = O_p(1).$$

Hence,

$$\max_{1 < k < n} \frac{|(Q_k^{(2)})^{-3} k^3 (t_{k-1} - 1)^3|}{(k \log \log k)^{3/2}} = O_p(1);$$

that is,

$$\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} |(Q_k^{(2)})^{-3} k(t_{k-1} - 1)^3| = O_p(1) \quad (2.34)$$

Inasmuch as  $\bar{z}_k$  is distributed as  $N(0, 1/k)$ ,  $E(k\bar{z}_k) = 0$ ,  $\text{Var}(k\bar{z}_k) = k$ . From the law of iterated logarithm,

$$\max_{1 < k < n} \frac{k\bar{z}_k}{(k \log \log k)^{1/2}} = O_p(1).$$

Therefore,

$$\max_{1 < k < n} \frac{k^2 \bar{z}_k^2}{(k \log \log k)} = O_p(1), \quad (2.35)$$

and

$$\max_{1 < k < n} \frac{k^4 \bar{z}_k^4}{(k \log \log k)^2} = O_p(1). \quad (2.36)$$

From (2.35),

$$\max_{1 < k < n} \left( \frac{k}{(\log \log k)} \right) \bar{z}_k^2 = O_p(1). \quad (2.37)$$

The law of iterated logarithm also implies

$$\max_{1 < k < n} \frac{\sum_{i=1}^k (z_i^2 - 1)}{(k \log \log k)^{1/2}} = O_p(1), \quad (2.38)$$

therefore combining (2.37) and (2.38), we obtain

$$\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} \bar{z}_k^2 \sum_{i=1}^k (z_i^2 - 1) = O_p(1). \quad (2.39)$$



Considering the fact that  $\lim_{n \rightarrow \infty} (\log \log k/k)^{1/2} = 0$  and combining it with (2.36), we thus obtain:

$$\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} k \bar{z}_k^4 = O_p(1). \quad (2.40)$$

Similar to (2.35), (2.39), and (2.40), from the law of iterated logarithm, we can show that

$$n^{1/2} (\log \log n)^{-(3/2)} |(Q_n^{(1)})^{-3} n(t_{n-1} - 1)^3| = O_p(1),$$

$$n^{1/2} (\log \log n)^{-(3/2)} \bar{z}_n^2 \sum_{i=1}^n (z_i^2 - 1) = O_p(1),$$

and

$$n^{1/2} (\log \log n)^{-(3/2)} n \bar{z}_n^4 = O_p(1).$$

Due to the inequality:  $k^{1/2} (\log \log k)^{-(3/2)} \leq n^{1/2} (\log \log n)^{-(3/2)}$ , for  $1 < k < n$ , we thus conclude:

$$\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} |(\theta_n^{(1)})^{-3} n(t_{n-1} - 1)^3| = O_p(1), \quad (2.41)$$

$$\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} z_n^{(j)2} \sum_{i=1}^n (z_i^2 - 1) = O_p(1), \quad (2.42)$$

and

$$\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} n \bar{z}_n^4 = O_p(1). \quad (2.43)$$

Also, (2.34) and (2.39) through (2.43) together give us

$$\max_{1 < k < n} k^{1/2} (\log \log k)^{-(3/2)} |\theta_k^{(1)}| = O_p(1).$$

(ii) Proceeding as in (i), we obtain

$$\max_{1 < k < n} (n-k)^{1/2} [\log \log (n-k)]^{-(3/2)} |(\theta_{n-k}^{(3)})^{-3} (n-k)(t_{n-k-1}^{(j)} - 1)^3| = O_p(1)$$

$$\max_{1 < k < n} (n-k)^{1/2} [\log \log (n-k)]^{-(3/2)} z_{n-k}^{(5)2} \sum_{i=k+1}^n (z_i^{(j)2}) = O_p(1)$$

$$\max_{1 < k < n} (n-k)^{1/2} [\log \log (n-k)]^{-(3/2)} (n-k) \bar{z}_{n-k}^{(j)4} = O_p(1).$$

Hence,

$$\max_{1 < k < n} (n-k)^{1/2} [\log \log (n-k)]^{-(3/2)} |\theta_k^{(2)}| = O_p(1).$$

□

**Lemma 2.22** For all  $x \in R$ , as  $n \rightarrow \infty$ ,

- (i)  $a^2(\log n) \max_{1 < k < \log n} (W_k^{(1)} + W_k^{(2)}) - [x + b(\log n)]^2 \xrightarrow{P} -\infty$ ,
- (ii)  $a^2(\log n) \max_{1 < k < \log n} \xi_k - [x + b(\log n)]^2 \xrightarrow{P} -\infty$ ,
- (iii)  $a^2(\log n) \max_{n - \log n < k < n} (W_k^{(1)} + W_k^{(2)}) - [x + b(\log n)]^2 \xrightarrow{P} -\infty$ ,
- (iv)  $a^2(\log n) \max_{n - \log n < k < n} \xi_k - [x + b(\log n)]^2 \xrightarrow{P} -\infty$ ,

where

$$a(\log n) = (2 \log \log n)^{1/2}, \quad (2.44)$$

$$b(\log n) = 2 \log \log n + \log \log \log n. \quad (2.45)$$

*Proof.* (i) Recall  $W_k^{(1)} = k\bar{z}_k^2 + (n - k)\bar{z}_{n-k}^2 - n\bar{z}_n^2$ . Because  $k\bar{z}_k^2 \sim x_1^2$ ,  $E(k\bar{z}_k^2) = 1$ , we have  $k\bar{z}_k^2 \xrightarrow{P} 1$  as  $k \rightarrow \infty$ . But  $1/\log k \rightarrow 0$  as  $k \rightarrow \infty$ , hence  $k\bar{z}_k^2/\log k \xrightarrow{P} 0$  as  $k \rightarrow \infty$ . There exists a constant  $c$ ,  $0 < c < 1$ , such that  $k\bar{z}_k^2/\log k \stackrel{P}{<} c$  for large  $k$ . Meanwhile,

$$\begin{aligned} \frac{a^2(\log n) \max_{1 < k < \log n} k\bar{z}_k^2}{[b(\log n)]^2} &\leq \frac{2 \log \log n \cdot \max_{1 < k < \log n} k\bar{z}_k^2}{(2 \log \log n)^2} \\ &\leq \max_{1 < k < \log n} \frac{k\bar{z}_k^2}{\log \log n} \\ &\leq \max_{1 < k < \log n} \frac{k\bar{z}_k^2}{\log k} \\ &\stackrel{P}{<} c = 1 - M, \quad 0 < M < 1. \end{aligned}$$

Hence,

$$a^2(\log n) \max_{1 < k < \log n} k\bar{z}_k^2 - [x + b(\log n)]^2 \stackrel{P}{<} -M[x + b(\log n)]^2;$$

that is,

$$a^2(\log n) \max_{1 < k < \log n} k\bar{z}_k^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty, \quad \text{as } n \rightarrow \infty. \quad (2.46)$$

Similarly, we can show that

$$a^2(\log n) \max_{1 < k < \log n} (n - k)\bar{z}_{n-k}^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty \quad \text{as } n \rightarrow \infty. \quad (2.47)$$

Because  $n\bar{z}_n^2 \sim \chi_1^2$ ,  $E(n\bar{z}_n^2) = 1$ . Then  $n\bar{z}_n^2 \xrightarrow{P} 1$  as  $n \rightarrow \infty$ ,  $-n\bar{z}_n^2 \xrightarrow{P} -1$  as  $n \rightarrow \infty$ . But  $a^2(\log n) \rightarrow \infty$  and  $[x + b(\log n)]^2 \rightarrow \infty$  as  $n \rightarrow \infty$ . Hence,

$$a^2(\log n)(-n\bar{z}_n^2) - [x + b(\log n)]^2 \xrightarrow{P} -\infty \quad \text{as } n \rightarrow \infty. \quad (2.48)$$

Combining (2.46) through (2.48), we thus obtain

$$a^2(\log n) \max_{1 < k < \log n} (W_k^{(1)} - [x + b(\log n)]^2) \xrightarrow{P} -\infty \quad \text{as } n \rightarrow \infty. \quad (2.49)$$

Recall again

$$W_k^{(2)} = -\frac{1}{2n} \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2.$$

Inasmuch as  $z_i^2 \sim \chi_1^2$ ,  $E \left[ \sum_{i=1}^k (z_i^2 - 1) \right] = 0$ . Then

$$E \left\{ \frac{1}{k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 \right\} = \text{Var} \left\{ \frac{1}{k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 \right\} = 2$$

as  $k \rightarrow \infty$ , and

$$\frac{[\sum_{i=1}^k (z_i^2 - 1)]^2}{k \log k} \xrightarrow{P} 0$$

as  $n \rightarrow \infty$ . Therefore, there exists a constant  $c$ ,  $0 < c < 1$ , such that

$$0 < \frac{[\sum_{i=1}^k (z_i^2 - 1)]^2}{k \log k} < 1 - c$$

for large  $k$ . Now,

$$\begin{aligned} & \frac{a^2(\log n) \max_{1 < k < \log n} \frac{1}{k} [\sum_{i=1}^k (z_i^2 - 1)]^2}{[b(\log n)]^2} \\ & \leq \frac{2 \log \log n \max_{1 < k < \log n} \frac{1}{k} [\sum_{i=1}^k (z_i^2 - 1)]^2}{(2 \log \log n)^2} \\ & < \max_{1 < k < \log n} \frac{[\sum_{i=1}^k (z_i^2 - 1)]^2}{k \log \log n} \\ & < \max_{1 < k < \log n} \frac{[\sum_{i=1}^k (z_i^2 - 1)]^2}{k \log k} \\ & \stackrel{P}{<} 1 - c. \end{aligned}$$

Hence, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{1 < k < \log n} \frac{1}{k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty. \quad (2.50)$$

Similarly, as  $n \rightarrow \infty$ , we have

$$a^2(\log n) \max_{1 < k < \log n} \frac{1}{n-k} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty. \quad (2.51)$$

Moreover,  $\left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 \xrightarrow{P} 0$ , as  $n \rightarrow \infty$ , therefore  $a^2(\log n)(1/n) \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 \xrightarrow{P} 0$  as  $n \rightarrow \infty$ . Hence, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \left\{ -\frac{1}{2n} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 \right\} - [x + b(\log n)]^2 \xrightarrow{P} -\infty. \quad (2.52)$$

Then, (2.50) through (2.52) together give us, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{1 < k < \log n} W_k^{(2)} - [x + b(\log n)]^2 \xrightarrow{P} -\infty. \quad (2.53)$$

From (2.49) and (2.53), as  $n \rightarrow \infty$ , we obtain:

$$a^2(\log n) \max_{1 < k < \log n} (W_k^{(1)} + W_k^{(2)}) - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

This completes the proof of (i).

Now recall  $\xi_k = W_k^{(1)} + W_k^{(2)} + Q_k^{(1)} + Q_k^{(2)}$  and from Lemma 2.22,

$$\max_{1 < k < \log n} k^{1/2} (\log \log k)^{-(3/2)} |Q_k^{(1)}| = O_p(1).$$

Then

$$\begin{aligned} & \frac{a^2(\log n) \max_{1 < k < \log n} |Q_k^{(1)}|}{[b(\log n)]^2} \\ & \leq \frac{2 \log \log n \max_{1 < k < \log n} |Q_k^{(1)}|}{(2 \log \log n)^2} \\ & \leq \frac{1}{\log \log n} \cdot \max_{1 < k < \log n} \frac{(\log \log k)^{3/2}}{k^{1/2}} \cdot k^{1/2} (\log \log k)^{-(3/2)} |Q_k^{(1)}| \\ & \leq \frac{(\log \log \log n)^{(3/2)}}{\log \log n} \cdot \max_{1 < k < \log n} k^{1/2} (\log \log k)^{-(3/2)} |Q_k^{(1)}|. \end{aligned}$$

Notice that

$$\lim_{n \rightarrow \infty} \frac{(\log \log \log n)^{3/2}}{\log \log n} = 0;$$

then, there exists a constant  $M$ ,  $0 < M < 1$ , such that

$$0 < \frac{a^2(\log n) \max_{1 < k < \log n} |Q_k^{(1)}|}{[b(\log n)]^2} < 1 - M \quad \text{for large } n.$$

Hence, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{1 < k < \log n} |Q_k^{(1)}| - [x + b(\log n)]^2 \xrightarrow{P} -\infty. \quad (2.54)$$

From Lemma 2.22,

$$\max_{1 < k < \log n} (n - k)^{1/2} [\log \log(n - k)]^{-(3/2)} |Q_k^{(2)}| = O_p(1).$$

Hence,

$$\begin{aligned} & \frac{a^2(\log n) \max_{1 < k < \log n} |Q_k^{(2)}|}{[b(\log n)]^2} \\ & \leq \frac{1}{\log \log n} \max_{1 < k < \log n} |Q_k^{(2)}| \\ & = \frac{1}{\log \log n} \max_{1 < k < \log n} \frac{[\log \log(n - k)]^{3/2}}{(n - k)^{1/2}} \cdot (n - k)^{1/2} [\log \log(n - k)]^{-(3/2)} |Q_k^{(2)}| \\ & \leq \frac{(\log \log n)^{3/2}}{(\log \log n)(n - \log n)^{1/2}} \max_{1 < k < \log n} (n - k)^{1/2} [\log \log(n - k)]^{-(3/2)} |Q_k^{(2)}| \\ & = \left( \frac{\log \log n}{n - \log n} \right)^{1/2} \max_{1 < k < \log n} (n - k)^{1/2} [\log \log(n - k)]^{-(3/2)} |Q_k^{(2)}|. \end{aligned}$$

Because  $\lim_{n \rightarrow \infty} (\log \log n)/(n - \log n) = 0$ , there exists a constant  $M$ ,  $0 < M < 1$ , such that for large  $n$ ,

$$0 < \frac{a^2(\log n) \max_{1 < k < \log n} |Q_k^{(2)}|}{[b(\log n)]^2} < 1 - M.$$

Then, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{1 < k < \log n} |Q_k^{(2)}| - [x + b(\log n)]^2 \xrightarrow{P} -\infty. \quad (2.55)$$

Combining (2.54), (2.55), and (i), we thus conclude that (ii) holds.

Next, recall that

$$W_k^{(1)} = k\bar{z}_k^2 + (n - k) - \bar{z}_{n-k}^2 - n\bar{z}_n^2.$$

Because  $k\bar{z}_k^2 \sim \chi_1^2$ ,  $E(k\bar{z}_k^2)^2 = 1$ , and  $k\bar{z}_k^2 \xrightarrow[k \rightarrow \infty]{P} 1$ , as  $k \rightarrow \infty$ , then

$$\max_{n-\log n < k < n} k\bar{z}_k^2 \xrightarrow{P} 1$$

as  $n \rightarrow \infty$ . But  $\lim_{n \rightarrow \infty} (1/\log \log k) = 0$  and  $k \rightarrow \infty$ , as  $n \rightarrow \infty$ , hence there exists a constant  $M$ ,  $0 < M < 1$ , such that for large  $n$

$$\max_{n-\log n < k < n} \frac{k\bar{z}_k^2}{\log \log k} < 1 - M.$$

Now,

$$\begin{aligned} \frac{a^2(\log n) \max_{n-\log n < k < n} k\bar{z}_k^2}{[b(\log n)]^2} &< \frac{1}{\log \log n} \max_{n-\log n < k < n} k\bar{z}_k^2 \\ &= \max_{n-\log n < k < n} \frac{k\bar{z}_k^2}{\log \log n} \\ &< \max_{n-\log n < k < n} \frac{k\bar{z}_k^2}{\log \log k} \\ &< 1 - M; \end{aligned}$$

then, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n-\log n < k < n} k\bar{z}_k^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

Similarly, we have as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n-\log n < k < n} (n-k)\bar{z}_{n-k}^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty,$$

and

$$a^2(\log n) \max_{n-\log n < k < n} (-n\bar{z}_n^2) - [x + b(\log n)]^2 \xrightarrow{P} -\infty \quad \text{as } n \rightarrow \infty.$$

Therefore, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n-\log n < k < n} W_k^{(1)} - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

Similarly, we can show that as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n-\log n < k < n} W_k^{(2)} - [x + b(\log n)]^2 \xrightarrow{P} -\infty,$$

and (iii) is established.

To prove (iv), we start with Lemma 2.22,

$$\max_{n-\log n < k < n} k^{1/2}(\log \log k)^{-(3/2)} |Q_k^{(1)}| = O_p(1).$$

Then,

$$\begin{aligned} & \frac{a^2(\log n) \max_{n-\log n < k < n} |Q_k^{(1)}|}{[b(\log n)]^2} \\ & \leq \frac{1}{\log \log n} \max_{n-\log n < k < n} |Q_k^{(1)}| \\ & = \frac{1}{\log \log n} \cdot \max_{n-\log n < k < n} \frac{(\log \log k)^{3/2}}{k^{1/2}} \cdot k^{1/2}(\log \log k)^{-(3/2)} |Q_k^{(1)}| \\ & \leq \left( \frac{\log \log n}{n - \log n} \right)^{1/2} \max_{n-\log n < k < n} k^{1/2}(\log \log k)^{-(3/2)} |Q_k^{(1)}|. \end{aligned}$$

There exists a constant  $M$ ,  $0 < M < 1$ , such that

$$\frac{a^2(\log n) \max_{n-\log n < k < n} |Q_k^{(1)}|}{[b(\log n)]^2} < 1 - M;$$

therefore, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n-\log n < k < n} |Q_k^{(1)}| - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

Starting with Lemma 2.22, we obtain

$$a^2(\log n) \max_{n-\log n < k < n} |Q_k^{(2)}| - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

In view of (iii), we thus conclude that (iv) holds.  $\square$

**Lemma 2.23** *As  $n \rightarrow \infty$ , the following hold.*

- (i)  $a^2(\log n) \max_{\log n \leq k \leq n-\log n} |\xi_k - (W_k^{(1)} + W_k^{(2)})| = o_p(1).$
- (ii)  $a^2(\log n) \max_{1 < k < n/\log n} |(n-k)\bar{z}_{n-k}^2 - n\bar{z}_n^2| = o_p(1), j = 1, \dots, m.$
- (iii)  $a^2(\log n) \max_{1 < k < n/\log n} \left| \frac{1}{n-k} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 \right| = o_p(1), j = 1, \dots, m.$

*Proof.* (i) Clearly,  $\xi_k - (W_k^{(1)} + W_k^{(2)}) = Q_k^{(1)} + Q_k^{(2)}.$

$$\begin{aligned} 0 & \leq a^2(\log n) \max_{\log n \leq k \leq n-\log n} |Q_k^{(1)} + Q_k^{(2)}| \\ & \leq 2 \log \log n \max_{\log n \leq k \leq n-\log n} |Q_k^{(1)}| + 2 \log \log n \cdot \max_{\log n \leq k \leq n-\log n} |Q_k^{(2)}| \end{aligned}$$

$$\begin{aligned}
&= 2 \log \log n \cdot \max_{\log n \leq k \leq n - \log n} \frac{(\log \log k)^{3/2}}{k^{1/2}} \cdot k^{1/2} (\log \log k)^{-(3/2)} |Q_k^{(1)}| \\
&\quad + 2 \log \log n \cdot \max_{\log n \leq k \leq n - \log n} \frac{[\log \log(n - k)]^{3/2}}{(n - k)^{1/2}} \cdot (n - k)^{1/2} \\
&\quad \cdot [\log \log(n - k)]^{-(3/2)} |Q_k^{(2)}| \\
&\leq \frac{2(\log \log n)^{5/2}}{(\log n)^{1/2}} \cdot \max_{\log n \leq k \leq n - \log n} k^{1/2} (\log \log k)^{-(3/2)} |Q_k^{(1)}| \\
&\quad + \frac{2(\log \log n)^{5/2}}{(\log n)^{1/2}} \max_{\log n \leq k \leq n - \log n} (n - k)^{1/2} [\log \log(n - k)]^{-(3/2)} |Q_k^{(2)}|.
\end{aligned}$$

Because  $\lim_{n \rightarrow \infty} (\log \log n)^{5/2} / (\ln n)^{1/2} = 0$ , in view of Lemma 2.22, we then obtain

$$\lim_{n \rightarrow \infty} a^2(\log n) \max_{\log n \leq k \leq n - \log n} |\xi_k - (W_k^{(1)} + W_k^{(2)})| = 0 \text{ in probability.}$$

Therefore (i) holds.

(ii) First, observe that

$$\begin{aligned}
(n - k)\bar{z}_{n-k}^2 - n\bar{z}_n^2 &= \frac{k}{n(n - k)} \left( \sum_{i=1}^n z_i \right)^2 \\
&\quad - \frac{2}{n - k} \left( \sum_{i=1}^n z_i \right) \left( \sum_{i=1}^k z_i \right) + \frac{1}{n - k} \left( \sum_{i=1}^k z_i \right)^2.
\end{aligned}$$

From the law of iterated logarithm,

$$\frac{\sum_{i=1}^n z_i}{(n \log \log n)^{1/2}} = O_p(1), \quad (2.56)$$

and hence,

$$\frac{(\sum_{i=1}^n z_i)^2}{n \log \log n} = O_p(1).$$

Furthermore,

$$\begin{aligned}
0 &< a^2(\log n) \max_{1 < k < n / \log n} \frac{k}{n(n - k)} \left( \sum_{i=1}^n z_i \right)^2 \\
&\leq 2 \log \log n \cdot \frac{\frac{n}{\log n}}{n \left( n - \frac{n}{\log n} \right)} \left( \sum_{i=1}^n z_i \right)^2 \\
&\leq \frac{2(\log \log n)^2}{\log n} \cdot \frac{\sum_{i=1}^n z_i^2}{n \log \log n}.
\end{aligned}$$



Because  $\lim_{n \rightarrow \infty} ((\log \log n)^2 / \log n) = 0$ , we obtain that

$$\lim_{n \rightarrow \infty} a^2(\log n) \max_{1 < k < (n/\log n)} (k/n(n-k)) \left( \sum_{i=1}^n z_i \right)^2 = 0 \text{ in probability.}$$

From the law of iterated logarithm again, we have

$$\max_{1 < k < n/\log n} \frac{\sum_{i=1}^k z_i}{(k \log \log k)^{1/2}} = O_p(1). \quad (2.57)$$

Then

$$\begin{aligned} 0 &\leq a^2(\log n) \max_{1 < k < n/\log n} \frac{2}{n-k} \sum_{i=1}^n |z_i| \cdot \left| \sum_{i=1}^k z_i \right| \\ &\leq \frac{4 \log \log n}{n - \frac{n}{\log n}} \max_{1 < k < n/\log n} \left| \sum_{i=1}^n z_i \right| \cdot \max_{1 < k < n/\log n} \left| \sum_{i=1}^k z_i \right| \\ &= \frac{4n^{\frac{1}{2}} (\log \log n)^{3/2}}{n - \frac{n}{\log n}} \frac{|\sum_{i=1}^n z_i|}{(n \log \log n)^{1/2}} \\ &\quad \cdot \max_{1 < k < (n/\log n)} (k \log \log k)^{1/2} \frac{|\sum_{i=1}^k z_i|}{(k \log \log k)} \\ &\leq \frac{4(\log \log n)^2 (\log n)^{1/2}}{\log n - 1} \frac{|\sum_{i=1}^n z_i|}{(n \log \log n)^{1/2}} \cdot \max_{1 < k < n/\log n} \frac{|\sum_{i=1}^k z_i|}{(k \log \log k)}. \end{aligned}$$

Combining  $\lim_{n \rightarrow \infty} (\log \log n)^2 (\log n)^{\frac{1}{2}} / (\log n - 1) = 0$  with (2.56) and (2.57), we obtain:

$$\lim_{n \rightarrow \infty} a^2(\log n) \max_{1 < k < (n/\log n)} |(n-k)\bar{z}_{n-k}^{(j)^2} - n\bar{z}_n^{(j)^2}| = 0 \text{ in probability;}$$

that is, (ii) holds.

(iii) Because

$$\begin{aligned} &E \left\{ \frac{1}{n-k} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 \right\} \\ &= E \left\{ \frac{1}{n-k} \sum_{i=k+1}^n (z_i^2 - 1)^2 + \frac{1}{n-k} \sum_{i \neq \iota} (z_i^2 - 1)(z_\iota^2 - 1) - \frac{1}{n} \sum_{i=1}^n (z_i^2 - 1) \right\}^2 \\ &\quad - \frac{1}{n} \sum_{i \neq \iota} (z_i^2 - 1)(z_\iota^2 - 1) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{n-k} \sum_{i=k+1}^n \text{Var}(z_i^2) + \frac{1}{n-k} \sum_{i \neq \ell} E(z_i^2 - 1)E(z_\ell^2 - 1) - \frac{1}{n} \sum_{i=1}^n \text{Var}(z_i^2) \\
&\quad - \frac{1}{n} \sum_{i \neq \ell} E(z_i^2 - 1)E(z_\ell^2 - 1) \\
&= \frac{1}{n-k} \sum_{i=k+1}^n 2 - \frac{1}{n} \sum_{i=1}^n 2 = 0 \quad \text{for all } n \text{ and all } k,
\end{aligned}$$

we have,  $E\langle a^2(\log n) \{ (1/(n-k)) [\sum_{i=k+1}^n (z_i^2 - 1)]^2 - (1/n) [\sum_{i=1}^n (z_i^2 - 1)]^2 \} \rangle = 0$  for all  $n$  and all  $k$ . Hence,

$$a^2(\log n) \left| \frac{1}{n-k} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 \right| \xrightarrow{P} 0$$

as  $n \rightarrow \infty$  for all  $k$ ,  $1 < k < n/\log n$ . That is,

$$a^2(\log n) \max_{1 < k < (n/\log n)} \left| \frac{1}{n-k} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 - \frac{1}{n} \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2 \right| = o_p(1).$$

□

**Lemma 2.24** For all  $x \in R$ , as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n - n/\log n} (W_k^{(1)} + W_k^{(2)}) - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

*Proof.* Note that  $W_k^{(1)} = k\bar{z}_k^2 + (n-k)\bar{z}_{n-k}^2 - n\bar{z}_n^2$ . Let's consider term by term:

$$k\bar{z}_k^2 = \left( \frac{\left| \sum_{i=1}^k z_i \right|}{k^{1/2}} \right)^2.$$

From Theorem 2 of Darling and Erdős (1956) we have

$$\begin{aligned}
&P \left[ \max_{n/\log n < k < n - (n/\log n)} k\bar{z}_k^2 \right. \\
&\quad \left. < \left[ (2 \log \log n)^{1/2} + \frac{\log \log \log n}{2(2 \log \log n)^{1/2}} + \frac{x}{(2 \log \log n)^{1/2}} \right]^2 \right] \\
&= e^{-(1/\sqrt{\pi})e^{-x}} \cdot x \in R.
\end{aligned}$$

Then,

$$\begin{aligned}
P & \left[ \frac{a^2(\log n) \max_{(n/\log n) < k < n - (n/\log n)} k \bar{z}_k^2}{[b(\log n) + x]^2} \right. \\
& \left. < \left[ \frac{2 \log \log n + \frac{1}{2} \log \log \log n + x}{b(\log n) + x} \right]^2 \right] \\
& = e^{-(1/\sqrt{\pi})e^{-x}}.
\end{aligned}$$

Because  $b(\log n) = 2 \log \log n + \log \log \log n$ , we can choose  $n$  large enough, such that

$$\left[ \frac{2 \log \log n + \frac{1}{2} \log \log \log n + x}{b(\log n) + x} \right]^2 < 1 - M, \quad 0 < M < 1.$$

Therefore,

$$P \left\{ \frac{a^2(\log n) \max_{(n/\log n) < k < n - (n/\log n)} k \bar{z}_k^2}{[x + b(\log n)]^2} < 1 - M \right\} = e^{-(1/\sqrt{\pi})e^{-x}}.$$

Letting  $x \rightarrow \infty$ , we then obtain

$$\begin{aligned}
P & \left[ a^2(\log n) \max_{n/\log n < k < n - n/\log n} k \bar{z}_k^2 \right. \\
& \left. - [x + b(\log n)]^2 < -M[x + b(\log n)]^2 \right] = 1.
\end{aligned}$$

Hence, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n - n/\log n} k \bar{z}_k^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

For the next term  $(n - k) \bar{z}_{n-k}^2$ , observe that

$$(n - k) \bar{z}_{n-k}^2 = \left[ \frac{|\sum_{i=k+1}^n z_i|}{(n - k)^{1/2}} \right]^2 \quad \text{and} \quad \frac{n}{\log n} < n - k < n - \frac{n}{\log n};$$

then, proceeding in the same manner as above, we can show that, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n - n/\log n} (n - k) \bar{z}_{n-k}^2 - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

For the last term  $-n \bar{z}_n^2$ , applying the law of iterated logarithm, we have  $n \bar{z}_n / (n \log \log n)^{1/2} = O_p(1)$ ; that is,  $n \bar{z}_n^2 / \log \log n = O_p(1)$ . Therefore, as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n-n/\log n} (-n\bar{z}_n^2) \xrightarrow{P} -\infty,$$

and

$$a^2(\log n) \max_{n/\log n < k < n-n/\log n} (-n\bar{z}_n^2) - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

Then we conclude from all of the above that as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n-n/\log n} W_k^{(1)} - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

Similarly, we can show that as  $n \rightarrow \infty$ ,

$$a^2(\log n) \max_{n/\log n < k < n-n/\log n} W_k^{(2)} - [x + b(\log n)]^2 \xrightarrow{P} -\infty.$$

Thus the lemma is proved.  $\square$

Similar to Lemma 2.22(ii) and (iii), we obtain the following results.

**Lemma 2.25**

- (i)  $a^2(\log n) \max_{n-(n/\log n) < k < n} |k\bar{z}_k^2 - n\bar{z}_n^2| = O_p(1), j = 1, \dots, m.$   
(ii)  $a^2(\log n) \max_{n-(n/\log n) < k < n} |(1/k) \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 - (1/n) \left[ \sum_{i=1}^n (z_i^2 - 1) \right]^2| = O_p(1), j = 1, \dots, m.$

*Proof.* (i) Start with the identity:

$$k\bar{z}_k^2 - n\bar{z}_n^2 = \frac{n-k}{kn} \left( \sum_{i=1}^n z_i \right)^2 - \frac{2}{k} \left( \sum_{i=1}^n z_i \right) \left( \sum_{i=k+1}^n z_i \right) + \frac{1}{k} \left( \sum_{i=k+1}^n z_i \right)^2.$$

The law of iterated logarithm yields  $\sum_{i=1}^n z_i / (n \log \log n)^{1/2} = o_p(1)$ . Then,  $(\sum_{i=1}^n z_i)^2 / (n \log \log n) = O_p(1)$ . Moreover,

$$\begin{aligned} 0 &< a^2(\log n) \max_{n-\frac{n}{\log n} < k < n} \frac{n-k}{kn} \left( \sum_{i=1}^n z_i \right)^2 \\ &< 2 \log \log n \cdot \frac{\frac{n}{\log n}}{(n - \frac{n}{\log n})n} \left( \sum_{i=1}^n z_i \right)^2 \\ &= \frac{2(\log \log n)^2}{\log n - 1} \cdot \frac{(\sum_{i=1}^n z_i)^2}{n \log \log n} \xrightarrow{P} 0, \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Hence,

$$a^2(\log n) \max_{n-(n/\log n) < k < n} \frac{n-k}{kn} \left( \sum_{i=1}^n z_i \right)^2 = o_p(1).$$

Proceeding similarly, we can show that

$$a^2(\log n) \max_{n-(n/\log n) < k < n} \frac{2}{k} \left| \sum_{i=1}^n z_i \right| \left| \sum_{i=k+1}^n z_i \right| = o_p(1),$$

and

$$a^2(\log n) \max_{n-(n/\log n) < k < n} \frac{1}{k} \left( \sum_{i=k+1}^n z_i \right)^2 = o_p(1).$$

Thus, (i) holds.

(ii) Similar to the proof of Lemma 2.22(iii), one can easily obtain (ii) here.  $\square$

Finally, we state without proof Lemma 2.2 of Horváth (1993).

**Theorem 2.26** *Under the null hypothesis  $H_0$ , when  $n \rightarrow \infty$ ,*

$$\lim_{n \rightarrow \infty} P[a(\log n)\lambda_n - b(\log n) \leq x] = \exp\{-2e^{-x}\}$$

for  $x \in R$ , where  $a(\log n)$  and  $b(\log n)$  are defined in (2.44) and (2.45).

*Proof.* First, observe that

$$\{1 < k < n\} = \{1 < k \leq \ln n\} \cup \{\log \leq k \leq n - \log n\} \cup \{n - \log n < k < n\}.$$

From Lemma 2.21(ii) and (iii), we obtain:

$$\max_{1 < k < n} \xi_k \stackrel{D}{=} \max_{\log n \leq k \leq n - \log n} \xi_k.$$

From Lemma 2.22(i), then

$$\max_{1 < k < n} \xi_k \stackrel{D}{=} \max_{\log n \leq k \leq n - \log n} (W_k^{(1)} + W_k^{(2)}).$$

But, for large  $n$ , we have

$$\begin{aligned} |\log n \leq k \leq n - \log n| &= \left\{ \log n \leq k \leq \frac{n}{\log n} \right\} \\ &\cup \left\{ \frac{n}{\log n} < k \leq n - \frac{n}{\log n} \right\} \\ &\cup \left\{ n - \frac{n}{\log n} < k \leq n - \log n \right\}, \end{aligned}$$

and

$$\left\{ \log n \leq k \leq \frac{n}{\log n} \right\} \subseteq \left\{ 1 \leq k \leq \frac{n}{\log n} \right\}.$$

From Lemma 2.22(ii) and (iii), we have

$$\begin{aligned} & \max_{\log n \leq k \leq (n/\log n)} (W_k^{(1)} + W_k^{(2)}) \\ & \stackrel{D}{=} \max_{\log n \leq k \leq (n/\log n)} \left\{ k z_k^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 \right\}. \end{aligned} \quad (2.58)$$

In view of Lemma 2.23, we have

$$\begin{aligned} & \max_{\log n \leq k \leq n - \log n} (W_k^{(1)} + W_k^{(2)}) \stackrel{D}{=} \left[ \max_{\log n \leq k \leq (n/\log n)} (W_k^{(1)} + W_k^{(2)}) \right] \\ & \vee \left[ \max_{n - (n/\log n) \leq k \leq n - \log n} (W_k^{(1)} + W_k^{(2)}) \right], \end{aligned} \quad (2.59)$$

where  $a \vee b \equiv \max\{a, b\}$ . Because

$$\left\{ n - \frac{n}{\log n} \leq k \leq n - \log n \right\} \subseteq \left\{ n - \frac{n}{\log n} \leq k \leq n \right\},$$

applying Lemma 2.24(i) and (ii), we obtain:

$$\begin{aligned} & \max_{n - (n/\log n) \leq k \leq n - \log n} (W_k^{(1)} + W_k^{(2)}) \\ & \stackrel{D}{=} \max_{n - (n/\log n) \leq k \leq n - \log n} \left\{ (n - k) \bar{z}_{n-k}^2 + \frac{1}{2(n - k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 \right\}. \end{aligned} \quad (2.60)$$

Combining (2.58) through (2.60), we thus have

$$\begin{aligned} & \max_{\log n \leq k \leq n - \log n} (W_k^{(1)} + W_k^{(2)}) \stackrel{D}{=} \max \left\{ \max_{1 \leq k < (n/\log n)} \left[ k \bar{z}_k^2 + \frac{1}{2k} \sum_{i=1}^k (z_i^2 - 1) \right], \right. \\ & \left. \max_{n - (n/\log n) \leq k < n} \left[ (n - k) \bar{z}_{n-k}^2 + \frac{1}{2(n - k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 \right] \right\}. \end{aligned}$$

Then,

$$\begin{aligned} & \lim_{n \rightarrow \infty} P\{a(\log n)\lambda_n - b(\log n) \leq x\} \\ & = \lim_{n \rightarrow \infty} P\left\{a(\log n) \max_{1 < k < n-1} \xi_k^{1/2} - b(\log n) \leq x\right\} \end{aligned}$$

$$\begin{aligned}
&= \lim_{n \rightarrow \infty} P \left\{ a^2(\log n) \max_{1 \leq k < n} \xi_k \leq [x + b(\log n)]^2 \right\} \\
&= \lim_{n \rightarrow \infty} P \left\{ a^2(\log n) \max_{\log n \leq k < n - \log n} (W_k^{(1)} + W_k^{(2)}) \leq [x + b(\log n)]^2 \right\} \\
&= \lim_{n \rightarrow \infty} P \left\{ a^2(\log n) \max \left\{ \max_{1 \leq k < (n/\log n)} [k\bar{z}_k^2 + \frac{1}{2k} \sum_{i=1}^k (z_i^2 - 1)], \right. \right. \\
&\quad \left. \left. \max_{n - (n/\log n) \leq k < n} \left[ (n - k)\bar{z}_{n-k}^2 + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 \right] \right\} \right. \\
&\quad \left. \leq [x + b(\log n)]^2 \right\}. \tag{2.61}
\end{aligned}$$

Because  $\{z_i, 1 \leq i < (n/\log n)\}$  and  $\{z_i, n - (n/\log n) \leq i \leq n\}$  are independent, (2.61) reduces to

$$\begin{aligned}
&\lim_{n \rightarrow \infty} P \left\{ a^2(\log n) \max_{1 \leq k < (n/\log n)} \left( k\bar{z}_k^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 \right) \leq [x + b(\log n)]^2 \right\} \\
&\cdot \lim_{n \rightarrow \infty} P \left\{ \left[ a^2(\log n) \max_{n - (n/\log n) \leq k < n} \left[ (n - k)\bar{z}_{n-k}^2 \right. \right. \right. \\
&\quad \left. \left. \left. + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 \right] \leq [x + b(\log n)]^2 \right\} \right. \\
&= \lim_{n \rightarrow \infty} P \left\{ a(\log n) \max_{1 \leq k < (n/\log n)} \left( k\bar{z}_k^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 \right)^{1/2} - b(\log n) \leq x \right\} \\
&\cdot \lim_{n \rightarrow \infty} P \left\{ a(\log n) \max_{n - (n/\log n) \leq k < n} \{ (n - k)\bar{z}_{n-k}^2 \right. \\
&\quad \left. + \frac{1}{2(n-k)} \left[ \sum_{i=k+1}^n (z_i^2 - 1) \right]^2 \}^{1/2} - b(\log n) \leq x \right\}. \tag{2.62}
\end{aligned}$$

Denote the first term of (2.62) by (a) and the second by (b). Let's consider (a) first. Note that

$$k\bar{z}_k^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2 = \left( \sum_{i=1}^k \frac{z_i}{\sqrt{k}} \right)^2 + \left( \sum_{i=1}^k \frac{z_i^2 - 1}{\sqrt{2k}} \right)^2.$$

Let  $\mathbf{v}_i = (z_i, ((z_i^2 - 1)/\sqrt{2}))$ ,  $1 \leq i < \infty$ , then  $\{\mathbf{v}_i, 1 \leq i < \infty\}$  is a sequence of iid  $d$ -dimensional random vectors with  $d = 2$ , and  $\mathbf{v}_i^{(1)} = z_i$ ,  $\mathbf{v}_i^{(2)} = (z_i^2 - 1)/\sqrt{2}$ . Now,  $E[\mathbf{v}_i^{(1)}] = E(z_i) = o$  for all  $i$ ,  $E[\mathbf{v}_i^{(2)}] = E((z_i^2 - 1)/\sqrt{2}) = (1 - 1)/\sqrt{2} = o$  for all  $i$ . Hence,  $E[\mathbf{v}_i^{(j)}] = o$  for  $j = 1, 2$  and all  $i$ .  $\text{Cov}(\mathbf{v}_i^{(j)}, \mathbf{v}_i^{(\iota)}) = o$  for  $1 \leq j \neq \iota \leq 2$ . If  $j = 1, \iota = 2$ ,  $\text{Cov}(\mathbf{v}_i^{(j)}, \mathbf{v}_i^{(\iota)}) = E(z_i, (z_i^2 - 1)/\sqrt{2}) = E((z_i^3 - z_i)/\sqrt{2}) = o$ . Therefore, the covariance matrix of  $\mathbf{v}_i$  is the  $2 \times 2$  identity matrix. And clearly,  $E|\mathbf{v}_i^{(j)}|^r < \infty$  for  $j = 1, 2$  and  $r > 2$ ; Let  $S_i^{(j)} = \sum_{i=1}^k \mathbf{v}_i^{(j)}$  for  $j = 1, 2$ ; then

$$\sum_{j=1}^2 \left( \frac{S_i^{(j)}}{\sqrt{k}} \right)^2 = k\bar{z}_k^2 + \frac{1}{2k} \left[ \sum_{i=1}^k (z_i^2 - 1) \right]^2.$$

In view of Lemma 2.18, we thus obtain  $(a) = \exp\{-e^{-x}\}$ . Similarly,  $(b) = \exp\{-e^{-x}\}$ . This completes the proof of the theorem.  $\square$

### 2.3.2 Informational Approach

#### (i) SICs

Under  $H_0$ , the MLEs for  $\mu$  and  $\sigma^2$  are

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

respectively. Then denoting that SIC under  $H_0$  by  $\text{SIC}(n)$ , we have:

$$\text{SIC}(n) = n \log 2\pi + n \log \hat{\sigma}^2 + n + 2 \log n. \quad (2.63)$$

Under  $H_1$ , we use  $\text{SIC}(k)$  to denote the SICs, for  $2 \leq k \leq n - 2$ . Then after some simple computations, we have:

$$\text{SIC}(k) = n \log 2\pi + k \log \hat{\sigma}_1^2 + (n - k) \log \hat{\sigma}_n^2 + n + 4 \log n, \quad (2.64)$$

where  $\hat{\sigma}_1^2 = (1/k) \sum_{i=1}^k (x_i - \bar{x}_k)^2$ ,  $\bar{x}_k = (1/k) \sum_{i=1}^k x_i$ ,  $\hat{\sigma}_n^2 = (1/(n - k)) \sum_{i=k+1}^n (x_i - \bar{x}_{n-k})^2$ , and  $\bar{x}_{n-k} = (1/(n - k)) \sum_{i=k+1}^n x_i$  are the MLEs for  $\hat{\sigma}_1^2$ ,  $\mu_1$ ,  $\sigma_n^2$ , and  $\mu_n$ , respectively. Now, we estimate  $k$  by  $\hat{k}$  such that

$$\text{SIC}(\hat{k}) = \min_{2 \leq k \leq n-2} \{\text{SIC}(k)\}. \quad (2.65)$$

It is noted that in order to obtain the MLEs, we can only detect change that is located at  $k$  for  $2 \leq k \leq n - 2$ .



**(ii) Asymptotic Null Distribution**

Let  $\Delta_n = \min_{2 \leq k \leq n-2} [\text{SIC}(k) - \text{SIC}(n)]$ . The asymptotic distribution of a function of  $\Delta_n$  is given in the following theorem. Note that

$$\begin{aligned}\Delta_n &= - \max_{2 \leq k \leq n-2} [\text{SIC}(k) - \text{SIC}(n)] \\ &= \lambda_n^2 + 2 \log n,\end{aligned}$$

where

$$\lambda_n^2 = \left[ \max_{2 \leq k \leq n-2} \langle n \log \hat{\sigma}^2 - k \log \hat{\sigma}_1^2 - (n-k) \log \hat{\sigma}_n^2 \rangle \right]^{1/2}.$$

$\lambda_n = (2 \log n - \Delta_n)^{1/2}$ , thus we have the following.

**Theorem 2.27** *Under  $H_0$ , for all  $x \in R$ ,*

$$\lim_{n \rightarrow \infty} P[a(\log n)(2 \log n - \Delta_n)^{1/2} - b(\log n) \leq x] = \exp(-2e^{-x}), \quad (2.66)$$

where  $a(\log n) = (2 \log \log n)^{1/2}$ , and  $b(\log n) = 2 \log \log n + \log \log \log n$ .

*Proof.* This is an immediate corollary of Theorem 2.26.  $\square$

We point out (see Gupta and Chen, 1996) that information criteria, such as SIC, provide a remarkable way for exploratory data analysis with no need to resort to either the distribution or the significant level  $\alpha$ . On the other hand, when the SICs are very close, one may question that the small difference among the SICs might be caused by the fluctuation of the data, and therefore there may be no change at all. To make the conclusion about change point statistically convincing, we introduce the significant level  $\alpha$  and its associated critical value  $c_\alpha$ . Instead of accepting  $H_0$  when  $\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k)$ , we accept  $H_0$ , if  $\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k) + c_\alpha$ , where  $c_\alpha$  and  $\alpha$  have the relationship:

$$1 - \alpha = P[\text{SIC}(n) < \min_{2 \leq k \leq n-2} \text{SIC}(k) + c_\alpha | H_0 \text{ holds}]. \quad (2.67)$$

From (2.66) and (2.67)

$$\begin{aligned}1 - \alpha &= P \left\{ - \max_{2 \leq k \leq n-2} [\text{SIC}(n) - \text{SIC}(k)] > -c_\alpha | H_0 \text{ holds} \right\} \\ &= P[\Delta_n > -c_\alpha | H_0 \text{ holds}] \\ &= P[-\lambda_n^2 + 2 \log n > -c_\alpha | H_0 \text{ holds}] \\ &= P[0 < \lambda_n < (c_\alpha + 2 \log n)^{1/2} | H_0 \text{ holds}]\end{aligned}$$

$$\begin{aligned}
&= P[-b(\log n) < a(\log n)\lambda_n - b(\log n) \\
&< a(\log n)(c_\alpha + 2 \log n)^{1/2} - b(\log n) | H_0 \text{ holds}] \\
&\cong \exp\{-2 \exp[b(\log n) - a(\log n)(c_\alpha + 2 \log n)^{1/2}]\} \\
&\quad - \exp\{-2 \exp(b \log n)\}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\exp\{-2 \exp[b(\log n) - a(\log n)(c_\alpha + 2 \log n)^{1/2}]\} \\
&\cong 1 - \alpha + \exp\{-2 \exp[b(\log n)]\}.
\end{aligned}$$

Solving for  $c_\alpha$ , we obtain:

$$\begin{aligned}
c_\alpha \cong &\left\{ -\frac{1}{a(\log n)} \log \log[1 - \alpha + \exp(-2 \exp(b(\log n)))]^{-(1/2)} \right. \\
&\left. + \frac{b(\log n)}{a(\log n)} \right\}^2 - 2 \log n.
\end{aligned}$$

For different significant levels  $\alpha = 0.01, 0.025, 0.05$ , and  $0.1$ , and different sample sizes  $n = 7, \dots, 200$ , we computed the critical values for SICs, and listed them in [Table 2.4](#).

### (iii) Unbiased SICs

Recall from previous sections, we mentioned that to derive the information criterion AIC, Akaike (1973) used  $\log L(\hat{\theta})$  as an estimate of  $J = E_{\hat{\theta}}[\int f(\mathbf{y}|\theta_0) \log f(\mathbf{y}|\hat{\theta}) d\mathbf{y}]$ , where  $f(\mathbf{y}|\theta_0)$  is the probability density of the future observations  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  of the same size and distribution as the  $\mathbf{x}$ s,  $\mathbf{x} = (x_1, x_2, \dots, x_n)$ , and  $\mathbf{x}$  and  $\mathbf{y}$  are independent. The expectation is taken under the distribution of  $\mathbf{x}$  when  $H_0$  is true; that is,  $\theta_0 \in H_0$ . Unfortunately,  $\log L(\hat{\theta})$  is not an unbiased estimator of  $J$ . When the sample size  $n$  is finite, Sugiura (1978) proposed unbiased versions, finite corrections of AIC, for different model selection problems.

In this section, we derive the unbiased version of SIC under our  $H_0$  and  $H_1$ , denoted by  $u - \text{SIC}(H_i)$ ,  $i = 0, 1$ .

(1)  $u - \text{SIC}(H_0)$

$$\begin{aligned}
J &= E_{\hat{\theta}}[E_{\theta_0}|y(\log L(\hat{\theta}))] \\
&= E_{\mathbf{x}} \left[ E_{\mathbf{y}} \left\{ -\frac{1}{2} n \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{1}{2} \sum_{i=1}^n \frac{(y_i - \bar{x})^2}{\hat{\sigma}^2} \right\} \right] \\
&= E_{\mathbf{x}} \left\{ -\frac{1}{2} n \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 - \frac{n}{2} + \frac{n}{2} - \frac{1}{2} E_{\mathbf{y}} \sum_{i=1}^n \frac{(y_i - \bar{x})^2}{\hat{\sigma}^2} \right\}.
\end{aligned}$$

**Table 2.4** Approximate Critical Values of SIC

$n/\alpha$	0.010	0.025	0.050	0.100
7	35.699	19.631	12.909	7.758
8	25.976	17.232	11.925	7.405
9	23.948	16.423	11.540	7.262
10	23.071	15.994	11.313	7.168
11	22.524	15.691	11.139	7.087
12	22.108	15.445	10.989	7.010
13	21.763	15.233	10.854	6.936
14	21.463	15.044	10.731	6.863
15	21.198	14.873	10.617	6.793
16	20.960	14.717	10.511	6.725
17	20.744	14.574	10.411	6.660
18	20.546	14.441	10.317	6.597
19	20.364	14.317	10.228	6.536
20	20.195	14.201	10.144	6.477
21	20.038	14.092	10.064	6.420
22	19.891	13.989	9.988	6.364
23	19.753	13.892	9.916	6.311
24	19.623	13.799	9.846	6.259
25	19.501	13.711	9.779	6.209
26	19.384	13.627	9.715	6.160
27	19.274	13.547	9.653	6.113
28	19.169	13.470	9.593	6.067
29	19.069	13.397	9.536	6.023
30	18.973	13.326	9.480	5.979
35	18.548	13.008	9.227	5.778
40	18.193	12.737	9.008	5.600
45	17.888	12.501	8.814	5.439
50	17.622	12.292	8.640	5.293
55	17.386	12.104	8.482	5.160
60	17.173	11.934	8.338	5.036
70	16.804	11.635	8.082	4.815
80	16.490	11.377	7.859	4.620
90	16.218	11.151	7.662	4.446
100	15.977	10.950	7.486	4.289
120	15.567	10.604	7.179	4.015
140	15.225	10.313	6.919	3.780
160	14.933	10.061	6.693	3.574
180	14.678	9.840	6.493	3.391
200	14.451	9.643	6.313	3.227

Notice that, because the  $x_i$ s and  $y_i$ s are independent, and are all distributed as  $N(\mu, \sigma^2)$ , we have  $y_i - \bar{x} \sim N(0, ((n+1)/n)\sigma^2)$ . Therefore,

$$\frac{n}{n+1} \sum_{i=1}^n \frac{(y_i - \bar{x})^2}{\hat{\sigma}^2} \sim \chi_n^2,$$

and hence

$$\begin{aligned}
 J &= E_{\hat{\theta}} \left[ \log L(\hat{\theta}) + \frac{n}{2} - \frac{1}{2} \frac{(n+1)\sigma^2}{\hat{\sigma}^2} \right] \\
 &= E_{\hat{\theta}} \left[ \log L(\hat{\theta}) + \frac{n}{2} - \frac{n+1}{2} E_{\hat{\theta}} \left( \frac{\sigma^2}{\hat{\sigma}^2} \right) \right] \\
 &= E_{\hat{\theta}} [\log L(\hat{\theta})] + \frac{n}{2} - \frac{n+1}{2} \frac{n}{n-3} \\
 &= E_{\hat{\theta}} [\log L(\hat{\theta})] - \frac{2n}{n-3}.
 \end{aligned}$$

Therefore,  $\log L(\hat{\theta}) - 2n/(n-3)$  is unbiased for  $J$ , or  $-2 \log L(\hat{\theta}) + 4n/(n-3)$  is unbiased for  $-2J$ . We have

$$\begin{aligned}
 u - \text{SIC}(H_0) &= -2 \log L(\hat{\theta}) + \frac{4n}{n-3} \\
 &= \text{SIC}(n) + \frac{4n}{n-3} - 2 \log n.
 \end{aligned}$$

(2)  $u - \text{SIC}(H_1)$

$$\begin{aligned}
 J &= E_{\hat{\theta}} \left[ E_{\mathbf{y}} \left\{ -\frac{1}{2} n \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{n-k}{2} \log \hat{\sigma}_n^2 - \frac{1}{2} \sum_{i=1}^k \frac{(y_i - \bar{x}_k)^2}{\hat{\sigma}_1^2} \right. \right. \\
 &\quad \left. \left. - \frac{1}{2} \sum_{i=k+1}^n \frac{(y_i - \bar{x}_{n-k})^2}{\hat{\sigma}_1^2} \right\} \right] \\
 &= E_{\hat{\theta}} \left\{ -\frac{1}{2} n \log 2\pi - \frac{k}{2} \log \hat{\sigma}_1^2 - \frac{n-k}{2} \log \hat{\sigma}_n^2 - \frac{n}{2} + \frac{n}{2} \right. \\
 &\quad \left. - \frac{1}{2} E_{\mathbf{y}} \left[ \sum_{i=1}^k \frac{(y_i - \bar{x}_k)^2}{\hat{\sigma}_1^2} \right] - \frac{1}{2} E_{\mathbf{y}} \left[ \sum_{i=k+1}^n \frac{(y_i - \bar{x}_{n-k})^2}{\hat{\sigma}_1^2} \right] \right\}.
 \end{aligned}$$

Now,

$$y_i - \bar{x}_k \sim N(0, \frac{k+1}{k} \sigma_1^2) \quad \text{and} \quad \frac{k+1}{k} \sum_{i=1}^k \frac{(y_i - \bar{x}_k)^2}{\hat{\sigma}_1^2} \sim \chi_k^2,$$

and therefore,

$$E_{\mathbf{y}} \left[ \sum_{i=1}^k \frac{(y_i - \bar{x}_k)^2}{\hat{\sigma}_1^2} \right] = (k+1) \frac{\sigma_1^2}{\hat{\sigma}_1^2}.$$

Similarly,

$$E_{\mathbf{y}} \left[ \sum_{i=k+1}^n \frac{(y_i - \bar{x}_{n-k})^2}{\hat{\sigma}_1^2} \right] = (n - k + 1) \frac{\sigma_1^2}{\hat{\sigma}_1^2}.$$

Thus,

$$\begin{aligned} J &= E_{\hat{\theta}}[\log L(\hat{\theta})] + \frac{n}{2} - \frac{k+1}{2} E_{\hat{\theta}} \left[ \frac{\sigma_1^2}{\hat{\sigma}_1^2} \right] - \frac{n-k+1}{2} E_{\hat{\theta}} \left[ \frac{\sigma_1^2}{\hat{\sigma}_1^2} \right] \\ &= E_{\hat{\theta}}[\log L(\hat{\theta})] + \frac{n}{2} - \frac{k+1}{2} \frac{k}{k-3} - \frac{n-k+1}{2} \frac{n-k}{n-k-3} \\ &= E_{\hat{\theta}}[\log L(\hat{\theta})] - \frac{k(k+1)(n-k-3) + (k-3)(n-k)(n-k+1)}{(k-3)(n-k-3)} \\ &\quad - \frac{n(k-3)(n-k-3)}{(k-3)(n-k-3)}. \end{aligned}$$

Hence,

$$\begin{aligned} u - \text{SIC}(H_1) &= -2 \log L(\hat{\theta}) \\ &\quad + 2 \frac{k(k+1)(n-k-3) + (k-3)(n-k)(n-k+1)}{(k-3)(n-k-3)} \\ &\quad - 2 \frac{n(k-3)(n-k-3)}{(k-3)(n-k-3)}. \end{aligned}$$

#### (iv) Data Analysis

*Example 2.1* As an application of SIC for change point analysis, we analyze the tensile strength data given in Shewhart (1931). There are 60 observations. Assume that the data are normally distributed with means  $\mu_1, \mu_2, \dots, \mu_{60}$  and variances  $\sigma_1^2, \sigma_2^2, \dots, \sigma_{60}^2$ , respectively. Then we test the following hypothesis,

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 = \dots = \mu_{60} = \mu \quad \text{and} \\ \sigma_1^2 &= \sigma_2^2 = \dots = \sigma_{60}^2 = \sigma^2, \end{aligned}$$

versus the alternative hypothesis

$$\begin{aligned} H_1 : \mu_1 &= \dots = \mu_k \neq \mu_{k+1} = \dots = \mu_{60} \quad \text{and} \\ \sigma_1^2 &= \dots = \sigma_k^2 \neq \sigma_{k+1}^2 = \dots = \sigma_{60}^2. \end{aligned}$$

Using the method developed here, we obtain the  $\text{SIC}(n)$ , and  $\text{SIC}(k)$ , for  $2 \leq k \leq n-1$ , and list them in [Table 2.5](#) along with the original data values, where the starred value is the minimum SIC value. Clearly,

**Table 2.5** SIC Values for the Tensile Strength Data

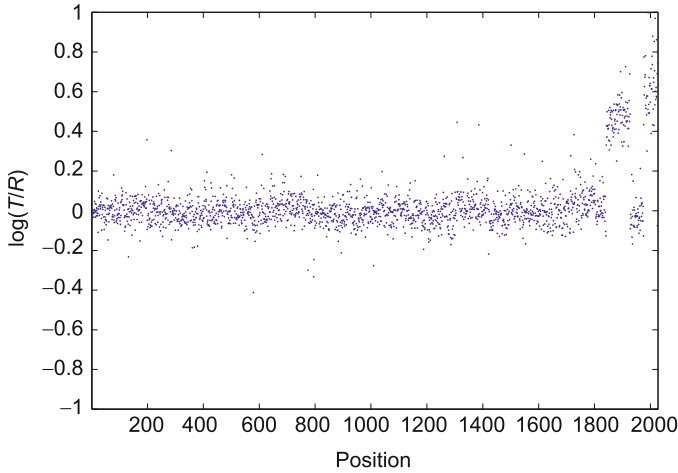
$x_k$	$k$	SIC( $k$ )	$x_k$	$k$	SIC( $k$ )	$x_k$	$k$	SIC( $k$ )
29314	1	—	25770	21	1180.2	29668	41	1177.7
34860	2	1181.7	23690	22	1177.3	32622	42	1179.1
36818	3	1181.4	28650	23	1177.4	32822	43	1179.6
30120	4	1181.2	32380	24	1178.6	30380	44	1178.5
34020	5	1180.7	28210	25	1177.3	38580	45	1183.2
30824	6	1180.2	34002	26	1178.9	28202	46	1178.7
35396	7	1179.8	34470	27	1179.9	29190	47	1177.9
31260	8	1179.1	29248	28	1178.8	35636	48	1182.0
32184	9	1178.2	28710	29	1178.0	34332	49	1182.6
33424	10	1177.1	29830	30	1177.9	34750	50	1183.9
37694	11	1177.8	29250	31	1177.1	40578	51	1189.3
34876	12	1176.3	27992	32	1175.6	28900	52	1184.4
24660	13	1180.2	31852	33	1176.5	34648	53	1188.1
34760	14	1180.3	27646	34	1174.1	31244	54	1186.9
38020	15	1179.9	31698	35	1174.9	33802	55	1188.1
25680	16	1180.6	30844	36	1174.3	34850	56	1188.7
25810	17	1181.5	31988	37	1174.6	36690	57	1189.6
26460	18	1181.8	36640	38	1177.6	32344	58	1187.5
28070	19	1182.0	41578	39	1181.6	34440	59	—
24640	20	1181.0	30496	40	1178.5	34650	60	1172.6*

\*Indicates the minimum SIC value

$\min_{2 \leq k \leq 58} \text{SIC}(k) = \text{SIC}(34) = 1174.1$ . Then  $\text{SIC}(n) = \text{SIC}(60) = 1172.6 < \min_{2 \leq k \leq 58} \text{SIC}(k)$ , and these two values are very close. What decision should we make? Use our Table 2.4, for any  $\alpha$ , because  $c_\alpha > 0$ , then  $\text{SIC}(n) < \min_{2 \leq k \leq 58} \text{SIC}(k) + c_\alpha$ . Therefore, we fail to reject  $H_0$ , and conclude that there is no change in both mean and variance of the tensile strength. This conclusion matches the one drawn in Shewhart (1931).

### 2.3.3 Application to Biomedical Data

We show an application of the mean and variance change point model to the analysis of aCGH data introduced in Section 2.1.3. In Linn et al. (2003) and Olshen et al. (2004), DNA copy number changes were viewed as a mean change point model (MCM) with a fixed variance in the distributions of the sequence  $\{X_i\}$ . As pointed out by Hodgson et al. (2001), the aCGH technology may not guarantee the aCGH data to have a constant variance; it is more reasonable to analyze the DNA copy number changes using the mean and variance change model (MVCM) proposed in Chen and Wang (2009) for the distributions of the sequence  $\{X_i\}$ . Observing the following normalized log-ratio intensities obtained through aCGH experiments of Lucito et al. (2003) on breast cancer cell line SK-BR-3 (see Figure 2.13), it is evident that both mean and variance of the sequence have changed.



**Fig. 2.13** Genome of the fibroblast cell line GM07408 Snijders et al. (2001)

The multiple DNA copy number changes in the sequence of log ratio intensities can be defined as the hypothesis testing problem stated earlier in this section (see (2.32)). Specifically, using BSP, we just need to focus on how to detect the single change (the most significant one), specified by testing (2.32) versus (2.33), each time and repeat the searching scheme of BSP to get all the significant changes. Here,  $\mu$  and  $\sigma^2$  are the unknown common mean and variance under the null hypothesis, and  $k$ ,  $1 < k < n$ , is the unknown position of the single change at each single stage. For a given significance level  $\alpha$ , when  $H_0$  is not rejected, there is no change in the DNA copy number sequence and the search scheme stops at this stage. If  $H_0$  is rejected at a given significance level  $\alpha$ , there is a significant change in the DNA copy number sequence and the search scheme of the BSP continues until no more significant changes are found.

As pointed out in Chen and Wang (2009), the advantage of using the MVCMM model is that MVCMM leads to fewer change points than that of the mean change point model (MCM) as MCM tends to divide large segments into smaller pieces so that the homogeneous variance assumption for all segments can be met (Picard et al., 2005). Therefore, the MVCMM model has the potential to give fewer false positives than MCM. Adding the variance component in the change point analysis will improve the estimation of the change point location even if just the mean shifts greatly. This is because in the MVCMM model, the variances under the alternative hypothesis are estimated for each subsequence without pooling all subsequences (with possible different means) together, whereas in MCM the homogeneous variance under the alternative hypothesis is estimated by pooling all subsequences with different means together. Using either MVCMM or MCM also depends on the biological experiment in which the scientists may have prior knowledge on whether there are potential variance changes. In that case, the MVCMM model

is proposed as an alternative to MCM when possible variance changes exist in the sequence.

To carry out the hypothesis testing of the null hypothesis (2.32), which claims no DNA copy number changes, versus the alternative hypothesis (2.33), the research hypothesis that there is a change in the mean and variance and hence a change in the DNA copy number, Chen and Wang introduced the SIC-based procedure along with an approximate  $p$ -value given by

$$p - \text{value} = 1 - \exp\{-2 \exp[b(\log n) - a(\log n)\lambda_n^{1/2}]\}, \quad (2.68)$$

where  $\lambda_n = 2 \log n - \Delta_n$ ,  $\Delta_n = \min_{2 \leq k \leq n-2} [\text{SIC}(k) - \text{SIC}(n)]$ , and  $\text{SIC}(k)$  and  $\text{SIC}(n)$  are given by (2.63) and (2.64), respectively.

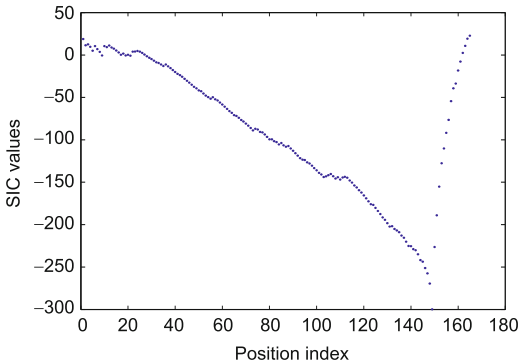
The applications of the SIC method to the detection of change point loci in the 15 fibroblast cell lines (Snijders et al., 2001) and other known aCGH data are given in Chen and Wang (2009). There are also comparisons of using the mean and variance change point model with the CBS method which is based on a mean change point model in Chen and Wang (2009).

There are important aCGH copy number experiments conducted by Snijders et al. (2001) on 15 fibroblast cell lines, namely GM03563, GM00143, GM05296, GM07408, GM01750, GM03134, GM13330, GM03576, GM01535, GM07081, GM02948, GM04435, GM10315, GM13031, and GM01524, and the obtained aCGH data on the genome of all such cell lines are regarded as benchmark aCGH datasets. There are many different computational and statistical methodology research articles published on how to analyze such aCGH datasets. The change point methods, CBS and MVCN, which were used for the analysis of the fibroblast aCGH data, were compared in Chen and Wang (2009) in terms of the change loci identified, the sensitivity, and specificity of the two methods. Two applications of the SIC approach are presented below and a predetermined significant level of  $\alpha = .001$  is used.

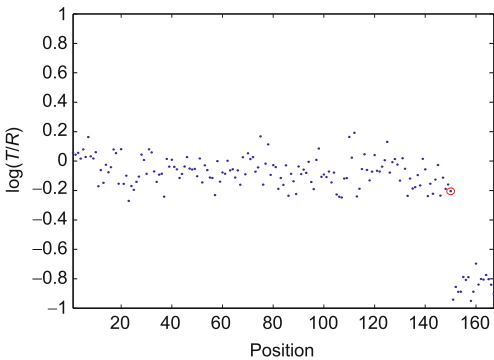
The first one is a chromosomewide copy number change search using SIC in MVCN on chromosome 4 of the fibroblast cell line GM13330. There are 167 genomic positions on which log base 2 ratio of intensities were recorded. The SIC values at all of the genomic locations were calculated according to expressions (2.63) and (2.64). The minimum SIC occurred at location index 150 with  $\min \text{SIC} = -299.8695$  and corresponding  $p$ -value (according to (2.68)) of  $6.465314 \times 10^{-9}$ . The graph of SIC values for this chromosome is given in Figure 2.14. Transferring back to the log ratio intensities, a scatterplot of the log ratio intensities of chromosome 4 of the fibroblast cell line GM13330 is provided as Figure 2.15 with the red circle indicating the change point identified.

The second application is a genomewide CNV search using SIC in MVCN on the cell line GM07408. It is found that the minimum SIC value of the whole sequence of 2027 log ratio intensity values occurs at locus 1841 with the  $p$ -value of 0.00000. The BSP is applied to the searching process. For the subsequence containing the first through the 1841st observations, the search

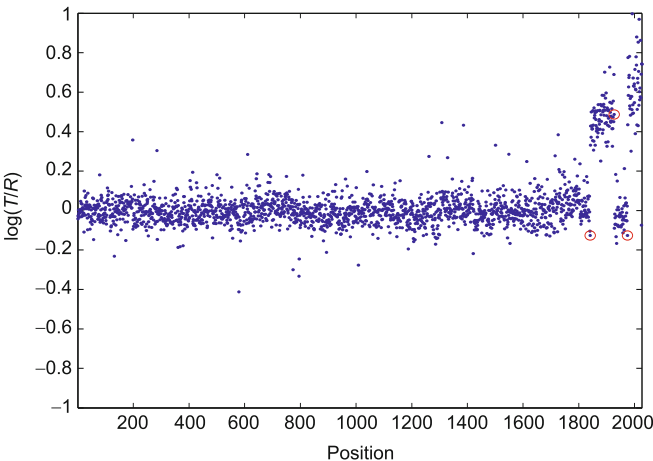




**Fig. 2.14** SIC values for every locus on chromosome 4 of the fibroblast cell line GM13330



**Fig. 2.15** Chromosome 4 of the fibroblast cell line GM13330 (Snijders et al., 2001)



**Fig. 2.16** Genome of the fibroblast cell line GM07408 with changes identified by red circles

locates no significant CNV, and for the subsequence consisting of the 1842nd through the 2027th log ratio intensity value, the minimum SIC occurs at the 1927th locus of the original sequence with the  $p$ -value of  $5.246360 \times 10^{-5}$ . After the identification of the 1927th change location, the subsequence is further broken into two subsubsequences and a third change is found at locus 1975 with the  $p$ -value of  $2.75359 \times 10^{-6}$ . These three loci are circled as red in the scatterplot, [Figure 2.16](#), of the genome of the fibroblast cell line GM07408.

Parametric Statistical Change Point Analysis

With Applications to Genetics, Medicine, and Finance

Chen, J.; Gupta, A.K.

2012, XIII, 273 p. 24 illus., 23 illus. in color., Hardcover

ISBN: 978-0-8176-4800-8

A product of Birkhäuser Basel