INFORMATION CRITERION AND CHANGE POINT PROBLEM FOR REGULAR MODELS[1]

Jiahua Chen
Department of Statistics & Actuarial Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

A. K. Gupta
Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, OH 43403-0221 U.S.A.

Jianmin Pan
Department of Statistics & Actuarial Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

Information criteria are commonly used for selecting competing statistical models. They do not favor the model which gives the best fit to the data and little interpretive value, but simpler models with good fit. Thus, model complexity is an important factor in information criteria for model selection. Existing results often equate the model complexity to the dimension of the parameter space. Although this notion is well founded in regular parametric models, it lacks some desirable properties when applied to irregular statistical models. We refine the notion of model complexity in the context of change point problems, and modify the existing information criteria. The modified criterion is found consistent in selecting the correct model and has simple limiting behavior. The resulting estimator $\hat{\tau}$ of the location of the change point achieves the best convergence rate $O_p(1)$, and its limiting distribution is obtained. Simulation results indicate that the modified criterion has better power in detecting changes compared to other methods.

Key words: consistency; irregular parametric model; limiting distribution; location; model complexity; regular parametric model; convergence rate.

1991 AMS Subject Classification: Primary 62H15; Secondary 62H10.

[1]Running title: Change point problem

1 Introduction

Out of several competing statistical models, we do not always use the one with the best fit to the data. Such models may simply interpolate the data and have little interpretive value. Information criteria, such as the Akaike information criterion and the Schwarz information criterion, are designed to select models with simple structure and good interpretive value, see Akaike (1973) and Schwarz (1978). The model complexity is often measured in terms of the dimensionality of the parameter space.

Consider the problem of making inference on whether a process has undergone some changes. In the context of model selection, we want to choose between a model with a single set of parameters, or a model with two sets of parameters plus the location of change. The Akaike and the Schwarz information criteria can be readily adopted to this kind of change point problems. There have been many fruitful research done in this respect such as Hirotsu, Kuriki and Hayter (1992) and Chen and Gupta (1997), to name a few.

Compared to usual model selection problems, the change point problem contains a special parameter: the location of the change. When it approaches the beginning or the end of the process, one of the two sets of the parameter becomes completely redundant. Hence, the model is un-necessarily complex. This observation motivates the notion that the model complexity also depends on the location of the change point. Consequently, we propose to generalize the Akaike and Schwarz information criteria by making the model complexity also a function of the location of the change point. The new method is shown to have a simple limiting behavior, and favourable power properties in many situations via simulation.

The change point problem has been extensively discussed in the literature in recent years. The study of the change point problem dates back to Page (1954, 1955 and 1957) which tested the existence of a change point. Parametric approaches to this problem have been studied by a number of researchers, see Chernoff and Zacks (1964), Hinkley (1970), Hinkley et. al. (1980), Siegmund (1986) and Worsley (1979, 1986). Nonparametric tests and estimations have also been proposed (Brodsky and Darkhovsky, 1993; Lombard, 1987; Gombay and Hušková, 1998). Extensive discussions on the large sample behavior of likelihood ratio test statistics can be found in Gombay and Horváth (1996) and Csörgö and Horváth (1997).

The detail can be found in some survey literatures such as Bhattacharya (1994), Basseville and Nikiforov (1993), Zacks (1983), and Lai (1985). The present study deviates from other studies by refining the traditional measure of the model complexity, and by determining the limiting distribution of the resulting test statistic under very general parametric model settings.

In Section 2, we define and motivate the new information criterion in detail. In Section 3, we give the conditions under which the resulting test statistic has chi-square limiting distribution and the estimator $\hat{\tau}$ of change point attains the best convergence rate. An application example and some simulation results are given in Section 4. The new method is compared to three existing methods and found to have good finite sample properties. The proofs are given in the Appendix.

2 Main Results

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables. It is suspected that X_i has density function $f(x, \theta_1)$ when $i \leq k$ and density $f(x, \theta_2)$ for $i > k$. We assume that $f(x, \theta_1)$ and $f(x, \theta_2)$ belong to the same parametric distribution family $\{f(x, \theta) : \theta \in R^d\}$. The problem is to test whether this change has indeed occurred and if so, find the location of the change k. The null hypothesis is $H_0 : \theta_1 = \theta_2$ and the alternative is $H_1 : \theta_1 \neq \theta_2$ and $1 \leq k < n$.

Equivalently, we are asked to choose a model from H_0 or a model from H_1 for the data.

For regular parametric (not change point) models with log likelihood function $\ell_n(\theta)$, Akaike and Schwarz information criteria are defined as

$$AIC = -2\ell_n(\hat{\theta}) + 2\dim(\hat{\theta});$$

$$SIC = -2\ell_n(\hat{\theta}) + \dim(\hat{\theta})\log(n)$$

where $\hat{\theta}$ is the maximum point of $\ell_n(\theta)$. The best model according to these criteria is the one which minimizes AIC or SIC. The Schwarz information criterion is asymptotically optimal according to certain Bayes formulation.

The log likelihood function for the change point problem has the form

$$\ell_n(\theta_1, \theta_2, k) = \sum_{i=1}^{k} \log f(X_i, \theta_1) + \sum_{i=k+1}^{n} \log f(X_i, \theta_2). \tag{1}$$

The Schwarz information criterion for the change point problem becomes

$$SIC(k) = -2\ell_n(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) + [2\text{dim}(\hat{\theta}_{1k}) + 1] \log(n)$$

and similarly for Akaike information criterion, where $\hat{\theta}_{1k}$, $\hat{\theta}_{2k}$ maximize $\ell_n(\theta_1, \theta_2, k)$ for given k. See, for example, Chen and Gupta (1997). When the model complexity is the focus, we may also write it as

$$SIC(k) = -2\ell_n(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) + \text{complexity}(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) \log(n).$$

We suggest that the notion of $\text{complexity}(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) = 2\text{dim}(\hat{\theta}_{1k}) + 1$ needs re-examination in the context of change point problem. When k takes values in the middle of 1 and n, both θ_1 and θ_2 are effective parameters. When k is near 1 or n, either θ_1 or θ_2 becomes redundant. Hence, k is an increasingly undesirable parameter as k getting close to 1 or n. We hence propose a modified information criterion with

$$\text{complexity}(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) = 2\text{dim}(\hat{\theta}_{1k}) + (\frac{2k}{n} - 1)^2 + \text{constant}. \tag{2}$$

For $1 \leq k < n$, let

$$MIC(k) = -2\ell_n(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) + [2\text{dim}(\hat{\theta}_{1k}) + (\frac{2k}{n} - 1)^2] \log(n). \tag{3}$$

Under the null model, we define

$$MIC(n) = -2\ell_n(\hat{\theta}, \hat{\theta}, n) + \text{dim}(\hat{\theta}) \log(n) \tag{4}$$

where $\hat{\theta}$ maximizes $\ell_n(\theta, \theta, n)$. If $MIC(n) > \min_{1 \leq k < n} MIC(k)$, we select the model with a change point and estimate the change point by $\hat{\tau}$ such that

$$MIC(\hat{\tau}) = \min_{1 \leq k < n} MIC(k).$$

Clearly, this procedure can be repeated when a second change point is suspected.

The size of model complexity can be motivated as follows. If the change point is at k, the variance of $\hat{\theta}_{1k}$ would be proportional to k^{-1} and the variance of $\hat{\theta}_{2k}$ would be proportional to $(n-k)^{-1}$. Thus, the total variance is

$$\frac{1}{k}+\frac{1}{n-k}=n^{-1}\left[\frac{1}{4}-(\frac{k}{n}-\frac{1}{2})^2\right]^{-1}.$$

The specific form in (2) reflects this important fact. Thus, if a change at an early stage is suspected, relatively stronger evidence is needed to justify the change. Hence, we should place larger penalty when k is near 1 or n. This notion is shared by many researchers. The method in Inclán and Tiao (1994) scales down the statistics heavier when the suspected change point is near 1 or n. The U-statistic method in Gombay and Horváth (1995) is scaled down by multiplying the factor $k(n-k)$.

To assess the error rates of the method, we can simulate the finite sample distribution, or find the asymptotic distribution of the related statistics. For Schwarz information criterion, the related statistic is found to have type I extreme value distribution asymptotically (Chen and Gupta, 1997; Csörgö and Horváth 1997). We show that the MIC statistic has chi-square limiting distribution for any regular distribution family, the estimator $\hat{\tau}$ achieves the best convergence rate $O_p(1)$ and has a limiting distribution expressed via a random walk.

Our asymptotic results under alternative model is obtained under the assumption that the location of the change point, k, evolves with n as $n\to\infty$. Thus, $\{X_{in}: 1\le i\le n, n\ge 2\}$ form a triangle array. The classical results on almost sure convergence for independent and identically distributed (iid) random variables cannot be directly applied. However, the conclusions on weak convergence will not be affected as the related probability statements are not affected by how one sequence is related to the other. Precautions will be taken on this issue but details will be omitted.

Let

$$S_n=MIC(n)-\min_{1\le k<n}MIC(k)+\dim(\theta)\log n \tag{5}$$

where $MIC(k)$ and $MIC(n)$ are defined in (3) and (4). Note that this standardization removes the constant term $\dim(\theta)\log n$ in the difference of $MIC(k)$ and $MIC(n)$.

THEOREM 1 *Under Wald conditions W1-W7 and the regularity conditions R1-R3, to be*

specified in Section 3, as $n \to \infty$,

$$S_n \to \chi^2_d \tag{6}$$

in distribution under H_0, *where* d *is the dimension of* θ *and* S_n *is defined in (5).*

In addition, if there has been a change at τ *such that as* $n \to \infty$, τ/n *has a limit in (0, 1), then*

$$S_n \to \infty \tag{7}$$

in probability.

Theorem 1 implies that the new information criterion is consistent in the sense that when there is indeed a fixed amount of change in θ at τ such that τ/n has a limit in (0, 1), the model with a change point will be selected with probability approaching 1.

Our next result claims that the estimator $\hat{\tau}$ achieves the best convergence rate and obtains its limiting distribution.

THEOREM 2 *Assume that the Wald conditions W1-W7 and the regularity conditions R1-R3 are satisfied by the parametric distribution family* $\{f(x;\theta), \theta \in \Theta\}$. *As* n *goes to infinity, the change point satisfies* $\tau = [n\lambda]$ *with* $0 < \lambda < 1$. *Then, the change point estimator*

$$\hat{\tau} - \tau = O_p(1).$$

Theorem 2 shows that the estimator $\hat{\tau}$ of the change point attains the best convergence rate.

We further show that the asymptotic distribution of the estimator $\hat{\tau}$ can be characterized by the minimizer of a random walk. Let $\{Y_i, i = -1, -2, \ldots\}$ be a sequence of iid random variables with common density function $f(x, \theta_{10})$. Similarly, let $\{Y_i, i = 1, 2, \ldots\}$ be iid random variables with the common density function $f(x, \theta_{20})$, and the two sequences are independent. Let Y_0 be a number such that $f(Y_0, \theta_{10}) = f(Y_0, \theta_{20})$.

Define $W_k = \sum_{j=0}^{k} \text{sgn}(k)[\log f(Y_j, \theta_{20}) - \log f(Y_j, \theta_{10})]$ for $k = 0, \pm 1, \pm 2, \ldots$.

THEOREM 3 *Assume the Wald conditions W1-W7, and the regularity conditions R1-R3. Assume that there exists a change point at* $\tau = [n\lambda]$ *where* $0 < \lambda < 1$. *As* $n \to \infty$, *we have*

$$\hat{\tau} - \tau \to \xi$$

in distribution, where $\xi = \arg\min_{-\infty<k<\infty}\{W_k\}$, *and* $(\theta_{10}, \theta_{20})$ *are the true value of* (θ_1, θ_2) *under the alternative.*

We spell out the conditions in the next Section and give a few examples when these conditions are satisfied. The proof of Theorems 1-3 will be given in the Appendix.

3 Conditions and Examples

The first crucial step in proving Theorem 1 is to establish the consistency of $\hat{\theta}_{1k}$, $\hat{\theta}_{2k}$ and $\hat{\theta}$ for estimating the single true value θ_0 under the null hypothesis H_0. Our approach is similar to that of Wald (1949). Consequently, the following conditions look similar to the conditions there.

W1. The distribution of X_1 is either discrete for all θ or is absolutely continuous for all θ.

W2. For sufficiently small ρ and for sufficiently large r, the expected values $E[\log f(X;\theta,\rho)]^2 < \infty$ and $E[\log \varphi(X,r)]^2 < \infty$ for all θ, where

$$f(x,\theta,\rho) = \sup_{\|\theta'-\theta\|\le\rho} f(x,\theta'); \quad \varphi(x,r) = \sup_{\|\theta'-\theta_0\|>r} f(x,\theta').$$

W3. The density function $f(x,\theta)$ is continuous in θ for every x.

W4. If $\theta_1 \neq \theta_2$, then $F(x,\theta_1) \neq F(x,\theta_2)$ for at least one x, where $F(x,\theta)$ is the cumulative distribution function corresponding to the density function $f(x,\theta)$.

W5. $\lim_{\|\theta\|\to\infty} f(x,\theta) = 0$ for all x.

W6. The parameter space Θ is a closed subset of the d-dimensional Cartesian space.

W7. $f(x,\theta,\rho)$ is a measurable function of x for any fixed θ and ρ.

The notation E will be understood as expectation under the true distribution which has parameter value θ_0 unless otherwise specified.

LEMMA 1 *Under Wald conditions W1-W7 and the null model,*

$$\hat{\theta}_{1k}, \hat{\theta}_{2k} \to \theta_0 \tag{8}$$

in probability uniformly for all k such that $\min(k, n-k)/\sqrt{n} \to \infty$ *as* $n \to \infty$.

Under Wald conditions W1-W7 and the alternative model with the change point at $\tau = [n\lambda]$ *for some* $0 < \lambda < 1$,

$$(\hat{\theta}_{1k}, \hat{\theta}_{2k}) \to (\theta_{10}, \theta_{20}) \tag{9}$$

in probability uniformly for all k such that $|k - \tau| < n(\log n)^{-1}$ *as* $n \to \infty$.

Lemma 1 allows us to focus on a small neighborhood of θ_0 or a small neighborhood of $(\theta_{10}, \theta_{20})$ under appropriate model assumptions. The precise asymptotic behavior of S_n is determined by the properties of the likelihood when θ is close to θ_0. This is where the regularity conditions are needed. The following conditions can be compared to those given in Serfling (1980).

R1. For each $\theta \in \Theta$, the derivatives

$$\frac{\partial \log f(x,\theta)}{\partial \theta}, \quad \frac{\partial^2 \log f(x,\theta)}{\partial \theta^2}, \quad \frac{\partial^3 \log f(x,\theta)}{\partial \theta^3}$$

exist for all x.

R2. For each $\theta_0 \in \Theta$, there exist functions $g(x)$ and $H(x)$ (possibly depending on θ_0) such that for θ in a neighborhood $N(\theta_0)$ the relations

$$\left|\frac{\partial f(x,\theta)}{\partial \theta}\right| \le g(x), \quad \left|\frac{\partial^2 f(x,\theta)}{\partial \theta^2}\right| \le g(x), \quad \left|\frac{\partial^2 \log f(x,\theta)}{\partial \theta^2}\right|^2 \le H(x), \quad \left|\frac{\partial^3 \log f(x,\theta)}{\partial \theta^3}\right| \le H(x)$$

hold for all x, and

$$\int g(x)dx < \infty, \quad E_\theta[H(X)] < \infty \quad \text{for } \theta \in N(\theta_0).$$

R3. For each $\theta \in \Theta$,

$$0 < E_\theta\left\{\left(\frac{\partial \log f(X,\theta)}{\partial \theta}\right)^2\right\} < \infty, \quad E_\theta\left\{\left|\frac{\partial \log f(X,\theta)}{\partial \theta}\right|^3\right\} < \infty.$$

Some of the Wald conditions are implied by regularity conditions. For clarity, we do not combine the two sets of conditions into a concise set of conditions. When θ is a vector, the above conditions are true for all components

Let a_n be a sequence of positive numbers and A_n be a sequence of random variables, $n = 1, 2, \ldots$. If $P(A_n \leq \epsilon a_n) \to 1$, for each $\epsilon > 0$, we say that $A_n \leq o_p(a_n)$. This convention will be used throughout the paper.

LEMMA 2 *Under the null hypothesis and assuming the Wald conditions W1-W7 and the regularity conditions R1-R3, we have*

$$\max_{1 \leq k < n} [\sup_{\theta_1, \theta_2} \ell_n(\theta_1, \theta_2, k) - \ell_n(\theta_0, \theta_0, n)] \leq o_p(\log \log n).$$

Lemma 2 indicates that random fluctuation in $MIC(k)$ is less than the penalty $(2k/n - 1)^2 \log n$ under the null model. Hence, the minimum of $MIC(k)$ is attained in the middle of 1 and n in probability under the null model. Lemmas 1 and 2 together imply that the MIC value is mainly determined by the likelihood function when θ and k are close to θ_0 and $n/2$.

LEMMA 3 *Under the null hypothesis and assuming Wald conditions W1-W7 and the regularity conditions R1-R3,* $\hat{t}_\tau = \frac{\hat{\tau}}{n} \to \frac{1}{2}$ *in probability as* $n \to \infty$.

The conditions required for these Theorems are not restrictive. In the following we use two examples to illustrate.

Example 1: Assume

$$f(x, \theta) = \frac{\theta^x}{x!} e^{-\theta}, \quad \theta \in (0, \infty)$$

and $x = 0, 1, 2, \ldots$. To prove that the Poisson model satisfies conditions W2, it is sufficient to show that $E[X \log(X) I(X > 1)]^2 < \infty$. This is obvious by using the Sterling's Formula.

Example 2: Assume that for all $x \geq 0$, $f(x, \theta) = \theta e^{-\theta x}$ with $\Theta = (0, \infty)$. Note that $E[\log f(X, \theta, \rho)]^2 < \infty$ for small ρ. On the other hand, for all $r > 0$, $\varphi(x, r) = \sup_{\|\theta - \theta_0\| > r} f(x, \theta)$ $\leq x^{-1}$. Hence, $E[\log \varphi(X, r)]^2 \leq E[\log X]^2 < \infty$.

Some models do not satisfy these conditions directly. The most interesting case is the normal model with unknown mean and variance. Its likelihood function is unbounded at

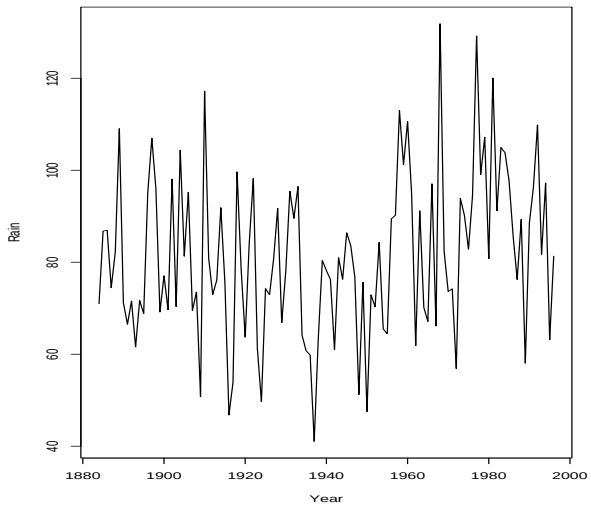

Figure 1: Yearly Data for Argentina Rainfall

$k = 1$, $\sigma^2 = 0$ and $\mu = X_1$. This obscure phenomenon quickly disappears by requiring $2 \leq k \leq n-2$. Other methods such as Schwarz information criterion also require $2 \leq k \leq n-2$.

It can be shown that all other conditions are satisfied by models in Examples 1 and 2. Most commonly used models satisfy the conditions of Theorems.

4 Application and Simulation Study

4.1 Example: Argentina Rainfall Data

In this section, we apply both our MIC and SIC criteria to a real data set, the Argentina rainfall data, which has been investigated by many researchers.

Argentina rainfall data set is a classical example where the change point analysis is used. The data set was provided by Eng Cesar Lamelas, a meteorologist in the Agricultural Experimental Station Obispo Colombres, Tucumán and contained yearly rainfall records collected from 1884 to 1996, with 113 observations in all. Lamelas believed that there was a change in the mean of the annual rain, caused by the construction of a dam in Tucumán between 1952 and 1962. Wu, Woodroofe and Mentz (2001) studied the data using the isotonic regression method and pointed out that there is an apparent increase in the mean.

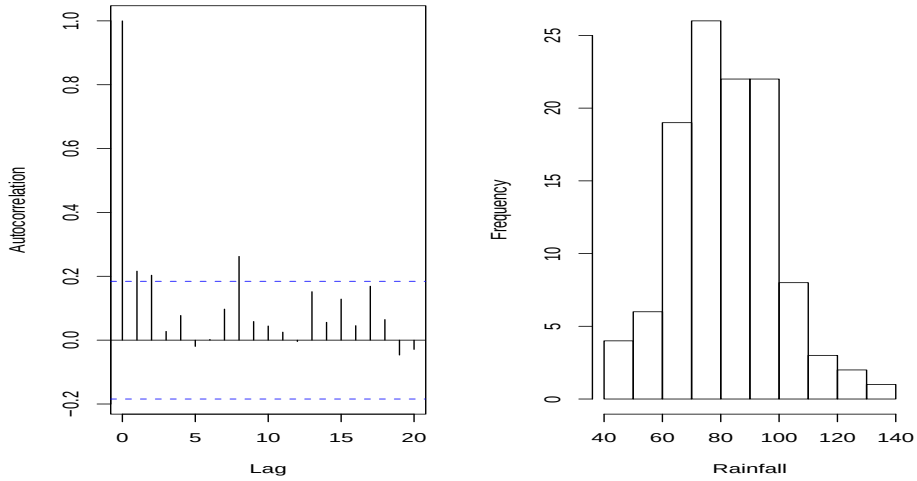

Figure 2: Autocorrelation Plot and Histogram for Argentina Rainfall Data

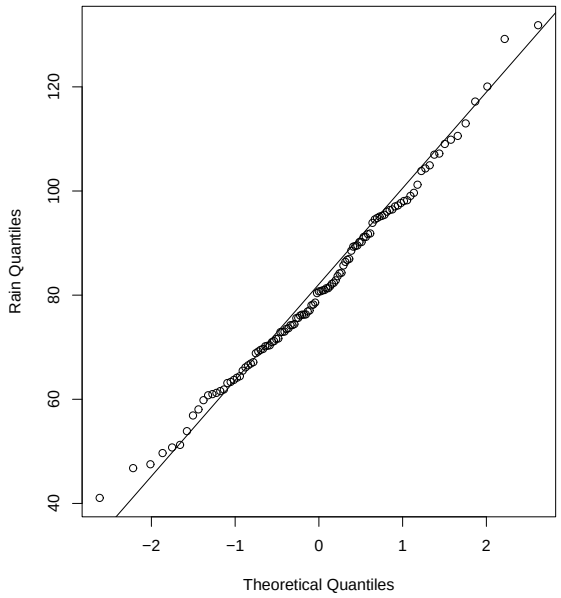

Figure 3: Normal Q-Q Plot for Argentina Rainfall Data

The rainfall data plot of $\{Y_t\}$, $t = 1, \ldots, 113$, is shown in Figure 1.

According to the autocorrelation and histogram plots in time series (see Figure 2) and the Normal Q-Q plot (see Figure 3), although there appears to be some autocorrelation in the sequence, it is not large enough to be of big concern. At this stage, we ignore the possibility of the autocorrelation. The normality seems to be reasonable. We use both the MIC and SIC based on the normal model to analyze the rainfall data, and examine the possibility of changes in mean and variance in the sequence. That is, we test the following hypothesis based on $\{Y_t\}$ series:

$$H_0 : \mu_1 = \mu_2 = \mu \quad \text{and} \quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \quad (\mu, \sigma^2 \quad \text{unknown})$$

against the alternative

$$H_1 : \mu_1 \neq \mu_2 \quad \text{or} \quad \sigma_1^2 \neq \sigma_2^2 \quad \text{and} \quad 1 < \tau < n.$$

Our analysis shows that $\min_{2<k<112} MIC(k) = MIC(71)$, $S_n = 17.97$ and the p-value is 1.3×10^{-4}. Hence $\tau = 71$ is suggested as a significant change point in mean and variance for the Y_t series, which corresponds to 1954. Not surprisingly, our conclusion fits very well with the fact that there was a dam constructed in Tucumán between 1952 and 1962.

If the Schwarz criterion is used, then we have $\min_{2<k<112} SIC(k) = SIC(71)$, $T_n = SIC(n) - \min_k SIC(k) + (d+1)\log n = 18.32$ and the p-value is 3.6×10^{-2}. Hence, two methods give similar conclusions.

For reference, we computed the means and standard deviations for the whole data set, the first subset (before the suggested change point 1954) and the second subset (after 1954) as follows: $\hat{\mu} = 81.33$, $\hat{\mu}_1 = 76.50$ and $\hat{\mu}_2 = 89.50$; $\hat{\sigma} = 17.59$, $\hat{\sigma}_1 = 15.46$ and $\hat{\sigma}_2 = 17.96$. From these values, we find that there is a substantial difference between $\hat{\mu}_1$ and $\hat{\mu}_2$, but not so apparent change in standard deviation. Hence the change point 1954 suggested by both the MIC and SIC methods is mainly due to the change in the mean precipitation.

4.2 Simulation Study

4.2.1 The Power Comparison between MIC and Other Methods

In this section, we use simulation to investigate finite sample properties of several methods. In addition to the modified information criterion, we also consider the Schwartz information criterion, the U-statistic method by Gombay and Horváth (1995) and the method given by Inclán and Tiao (1994). The U-statistics method computes the values of the U-statistics based on the first k observations, and based on the rest $n-k$ observations. If there is a change at k, the difference between these values will be stochastically larger than the difference when there is no change. They choose to standardize the difference by $k(n-k)$. The method in Inclán and Tiao (1994) computes the ratio of the sum of squares of the first k observations to the sum of squares of all observations. If the variance of the first k observations is different from the variance of the rest observations, this ratio likely deviates from k/n significantly. The test statistic is defined as the maximum absolute difference between the ratio and k/n.

We considered three models in our simulation: normal distribution with a change in the mean, normal distribution with a change in the variance, and exponential distribution with a change in the scale. For the U-statistics, we choose kernel function $h(x) = x$ for the first model, and $h(x, y) = (x - y)^2$ for the second and third models. Since the sample mean and variance are complete and sufficient statistics in the first two models respectively, the choices of the kernel function are very sensible.

The sample sizes are chosen to be $n = 100$ and 200. Under alternative model, we placed the change at 25%, 50% and 75% points. The amount of change in normal mean was a difference of 0.5, in normal variance was a factor of 2, and in exponential mean is a factor of $\sqrt{2}$. The simulation was repeated 5000 times for all combinations of sample size, location of change and the model. In Table 1, we report the null rejection rates (in percentage) of the Schwarz and Modified information criteria in the rows of SIC and MIC. Because two criteria have rather different null rejection rates, direct power comparison is not possible. In the row called SIC*, we calculated the powers of the Schwarz information criterion after its null rejection rates were made the same as the corresponding MIC (by increasing/decreasing its critical values). The row called U*, T* are obtained from two other methods similarly.

Table 1: Comparison between SIC and MIC

	n=100				n=200			
	normal model: change in the mean(c = 0.5)							
k	0	25	50	75	0	50	100	150
MIC	14.7	58.3	78.8	58.4	10.2	79.1	94.4	78.0
SIC	4.94	37.2	49.1	36.4	3.06	61.0	75.7	59.7
SIC*		56.3	67.1	56.4		76.8	87.8	76.2
U*		59.7	78.3	60.0		80.5	94.3	79.5
	normal model: change in the variance (c = 2)							
k	0	25	50	75	0	50	100	150
MIC	13.6	53.3	75.3	54.3	9.2	74.9	94.3	77.0
SIC	5.70	31.8	45.7	37.4	4.58	51.5	72.9	60.1
SIC*		46.9	61.0	50.9		67.9	84.3	73.5
U*		21.1	50.5	49.1		28.1	70.9	66.6
T*		24.6	65.5	54.1		33.9	86.2	77.2
	Exponential model, change in the mean (c=$\sqrt{2}$)							
k	0	25	50	75	0	50	100	150
MIC	14.1	36.9	53.3	35.7	10.2	48.7	71.4	49.5
SIC	6.46	18.7	24.8	18.9	3.72	26.5	37.8	28.9
SIC*		33.9	40.8	34.6		43.4	56.3	45.2
U*		20.8	36.0	32.1		21.8	46.2	38.3
T*		22.7	44.2	38.8		29.9	62.5	51.1

We can make several conclusions from the simulation results. First, both information criteria are consistent. When the sample size increases from $n = 100$ to $n = 200$, the probabilities of type I errors decrease and the powers increase. Second, after the probabilities of type I errors have been lined up, the powers of the modified information criterion are higher than the corresponding Schwarz information criterion in all cases, higher than U^* and T^* in most cases. There is only one case when the power of the MIC is more than 2% lower than another method. Third, the MIC is most powerful when the change point is in the middle of the sequence. The differences of 2% or more are considered significant with 5000 repetitions.

4.2.2 The Comparison of $\hat{\tau}$ and its Limiting Distribution

The limiting distributions of the estimators of the location of the change point based on the MIC and SIC methods are the same. To investigate their finite sample properties, we simulated the limiting distribution given in Theorem 3, and the distributions of $\hat{\tau} - \tau$ under the MIC and SIC criteria. We considered four models. The first three models are the same as the models in the last subsection. The additional model is a normal model with changes in both mean and variance with mean difference 0.5 and variance ratio 2. To save space, we only report results when sample sizes are 100 and 1000, and placed the change at $\tau = n/4$ and $n/2$ in the simulation.

We conducted simulation with 5000 repetitions for each combination and obtained the rates of ξ, $\hat{\tau}_{MIC} - \tau$ and $\hat{\tau}_{SIC} - \tau$ belong to specified intervals. The results are presented in Tables 2 and 3. It is seen that the limiting distribution (that of ξ) is a good approximation for both $\hat{\tau}$ when n is in the range of 1000.

When n is as small as 100, $P\{|\hat{\tau}_{MIC} - \tau| \leq \delta\} \geq P\{|\hat{\tau}_{SIC} - \tau| \leq \delta\}$ in all cases; The difference narrows when n increases. At $n = 1000$, two estimators are almost the same.

5 Appendix: Proofs

We split the proof of Lemmas 1 and 2 into several steps.

Lemma 1A. Let $f(x, \theta, \rho)$ and $\varphi(x, r)$ be the functions defined in W2 such that $0 < \rho < \|\theta - \theta_0\|$ and $E \log[f^*(X, \theta, \rho)] < \infty$. Then, for some $\rho > 0$ small enough (may depend on

Table 2: The Comparison of the Distributions $\hat{\tau} - \tau$ and ξ for $\tau = n/2$

$P_\xi = P\{\lvert\xi\rvert \le \delta\}$, $P_M = P\{\lvert\hat{\tau}_{MIC} - \tau\rvert \le \delta\}$ and $P_S = P\{\lvert\hat{\tau}_{SIC} - \tau\rvert \le \delta\}$						
Models	Prob	$\delta = 5$	$\delta = 10$	$\delta = 20$	$\delta = 30$	$\delta = 40$
	n=100					
Model 1	P_ξ	0.4698	0.6488	0.8288	0.9048	0.9426
	P_M	0.5156	0.7016	0.8692	0.9416	0.9796
	P_S	0.4100	0.5632	0.7250	0.8154	0.8952
Model 2	P_ξ	0.4562	0.6248	0.7922	0.8820	0.9344
	P_M	0.4846	0.6670	0.8502	0.9308	0.9712
	P_S	0.3644	0.5036	0.6688	0.7780	0.8672
Model 3	P_ξ	0.3066	0.4590	0.6432	0.7538	0.8298
	P_M	0.3852	0.5682	0.7856	0.8888	0.9502
	P_S	0.2486	0.3704	0.5360	0.6582	0.7864
Model 4	P_ξ	0.5820	0.7632	0.9096	0.9636	0.9856
	P_M	0.5406	0.7040	0.8356	0.8924	0.9218
	P_S	0.4470	0.5806	0.7006	0.7676	0.8196
	n=1000					
Model 1	P_ξ	0.4804	0.6492	0.8152	0.8966	0.9384
	P_M	0.4580	0.6498	0.8150	0.8896	0.9322
	P_S	0.4562	0.6472	0.8122	0.8876	0.9304
Model 2	P_ξ	0.4456	0.6216	0.8002	0.8878	0.9332
	P_M	0.4524	0.6190	0.7926	0.8750	0.9228
	P_S	0.4510	0.6172	0.7910	0.8732	0.9212
Model 3	P_ξ	0.3062	0.4584	0.6288	0.7324	0.8000
	P_M	0.2972	0.4478	0.6180	0.7190	0.7920
	P_S	0.2912	0.4376	0.6034	0.7024	0.7738
Model 4	P_ξ	0.6000	0.7706	0.9058	0.9578	0.9782
	P_M	0.5882	0.7660	0.8978	0.9528	0.9744
	P_S	0.5878	0.7652	0.8970	0.9518	0.9736

Table 3: The Comparison of the Distributions $\hat{\tau} - \tau$ and ξ for $\tau = n/4$

$P_\xi = P\{\lvert\xi\rvert \le \delta\}$, $P_M = P\{\lvert\hat{\tau}_{MIC} - \tau\rvert \le \delta\}$ and $P_S = P\{\lvert\hat{\tau}_{SIC} - \tau\rvert \le \delta\}$						
Models	Prob	$\delta = 5$	$\delta = 10$	$\delta = 20$	$\delta = 30$	$\delta = 40$
	n=100					
Model 1	P_ξ	0.4698	0.6488	0.8288	0.9048	0.9426
	P_M	0.4048	0.5614	0.7422	0.8516	0.9114
	P_S	0.3812	0.5284	0.7068	0.8072	0.8402
Model 2	P_ξ	0.4562	0.6248	0.7922	0.8820	0.9344
	P_M	0.3724	0.5334	0.7234	0.8358	0.9028
	P_S	0.3552	0.5050	0.6884	0.7940	0.8304
Model 3	P_ξ	0.3066	0.4590	0.6432	0.7538	0.8298
	P_M	0.2642	0.3992	0.5980	0.7570	0.8564
	P_S	0.2336	0.3566	0.5428	0.6902	0.7406
Model 4	P_ξ	0.5820	0.7632	0.9096	0.9636	0.9856
	P_M	0.4332	0.5674	0.7262	0.8454	0.8842
	P_S	0.3754	0.5002	0.6488	0.7988	0.8204
	n=1000					
Model 1	P_ξ	0.4804	0.6492	0.8152	0.8966	0.9384
	P_M	0.4644	0.6456	0.8082	0.8826	0.9226
	P_S	0.4652	0.6498	0.8124	0.8870	0.9274
Model 2	P_ξ	0.4456	0.6216	0.8002	0.8878	0.9332
	P_M	0.4394	0.5978	0.7742	0.8596	0.9052
	P_S	0.4420	0.6048	0.7832	0.8686	0.9136
Model 3	P_ξ	0.3062	0.4584	0.6288	0.7324	0.8000
	P_M	0.2760	0.4142	0.5778	0.6804	0.7516
	P_S	0.2796	0.4194	0.5886	0.6918	0.7596
Model 4	P_ξ	0.6000	0.7706	0.9058	0.9578	0.9782
	P_M	0.5982	0.7602	0.9020	0.9546	0.9768
	P_S	0.5966	0.7584	0.9028	0.9548	0.9764

$\theta \neq \theta_0$),

$$\max_k \left[\sum_{i=1}^{k} \log f(X_i, \theta, \rho) - \sum_{i=1}^{k} \log f(X_i, \theta_0) \right] \leq o_p(\log \log n).$$

Similarly, for some $r > 0$ large enough,

$$\max_k \left[\sum_{i=1}^{k} \log \varphi(X_i, r) - \sum_{i=1}^{k} \log f(X_i, \theta_0) \right] \leq o_p(\log \log n).$$

PROOF: From the regularity conditions,

$$\lim_{\rho \to 0^+} E[\log f(X, \theta, \rho) - \log f(X, \theta_0)] = E[\log f(X, \theta) - \log f(X, \theta_0)] = -K(\theta, \theta_0) < 0,$$

where K is the Kullback Leibler Information. Hence, we can find ρ small enough such that $EY_i < 0$ for $Y_i = \log f(X_i, \theta, \rho) - \log f(X_i, \theta_0)$.

By the Lévy-Skorohod inequality (Shokack and Wellner, 1986, pp. 844), we have for any $\varepsilon > 0$ and $0 < c < 1$,

$$\begin{aligned} & P\left\{ \max_k \sum_{i=1}^{k} [\log f(X_i, \theta, \rho) - \sum_{i=1}^{k} \log f(X_i, \theta_0)] \geq \varepsilon \log \log n \right\} \\ \leq & \frac{P\left[\sum_{i=1}^{n} Y_i \geq c\varepsilon \log \log n\right]}{\min_{1 \leq i \leq n} P\left[\sum_{j=i+1}^{n} Y_j \leq (1-c)\varepsilon \log \log n\right]}. \end{aligned} \tag{10}$$

By Chebyshev's inequality and noting $EY_1 < 0$,

$$\begin{aligned} P\left[\sum_{i=1}^{n} Y_i \geq c\varepsilon \log \log n\right] &= P\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Y_i - EY_i) \geq -\sqrt{n} EY_1 + \frac{c\varepsilon}{\sqrt{n}} \log \log n\right] \\ &\to 0, \end{aligned} \tag{11}$$

and, for all i, $P\left[\sum_{j=i+1}^{n} Y_j \leq (1-c)\varepsilon \log \log n\right] \to 1$. Hence, the probability in (10) goes to 0. The result for $\varphi(x, r)$ can be proved in the same way.

Lemma 1B. Let $N(\theta_0)$ be an open neighborhood of θ_0 and define

$$A = \{\theta : \|\theta - \theta_0\| \leq r,\ \theta \notin N(\theta_0)\}.$$

Then,

$$\max_k \sup_{\theta \in A} \left[\sum_{i=1}^{k} \log f(X_i, \theta) - \sum_{i=1}^{k} \log f(X_i, \theta_0) \right] \leq o_p(\log \log n).$$

PROOF: Note that A is a bounded, closed subset of R^d and hence is compact. By finite coverage theorem, we can find $\{\theta_j : j = 1, \ldots, N\}$ and their corresponding $\{\rho_j, j = 1, \ldots, N\}$

such that $\bigcup_{j=1}^{N}\{\theta : \|\theta - \theta_j\| < \rho_j\} \supset A$. Thus,

$$\begin{aligned}
&\max_{1\le k<n} \sup_{\theta\in A}\left[\sum_{i=1}^{k}\log f(X_i,\theta) - \sum_{i=1}^{k}\log f(X_i,\theta_0)\right] \\
&\le \max_{1\le k<n} \max_{1\le j\le N}\left[\sum_{i=1}^{k}\log f(X_i,\theta_j,\rho_j) - \sum_{i=1}^{k}\log f(X_i,\theta_0)\right] \\
&\le o_p(\log\log n) \qquad (12)
\end{aligned}$$

by Lemma 1A and the finiteness of N.

To finish the job, we further consider θ in $N(\theta_0)$.

Lemma 1C. For $N(\theta_0)$ small enough,

$$\max_{k} \sup_{\theta\in N(\theta_0)}\left[\sum_{i=1}^{k}\log f(X_i,\theta) - \sum_{i=1}^{k}\log f(X_i,\theta_0)\right] \le o_p(\log\log n).$$

PROOF: Without loss of generality, we will proceed with the proof as if θ is one-dimensional. Let $C_n = (\log\log n)^{1/2}$ and consider the maximum over the range of $k \le C_n$. The quantity is less than $\max_{k\le C_n}\sum_{i=1}^{k}[\log f(X_i,\theta_0,\rho_0) - \log f(X_i,\theta_0)]$ for some small $\rho_0 > 0$ when $N(\theta_0)$ is small enough. Further, for all k in this range,

$$\begin{aligned}
&\sum_{i=1}^{k}[\log f(X_i,\theta_0,\rho_0) - \log f(X_i,\theta_0)] \\
&= \sum_{i=1}^{k}\{[\log f(X_i,\theta_0,\rho_0) - \log f(X_i,\theta_0)] - E[\log f(X_i,\theta_0,\rho_0) - \log f(X_i,\theta_0)]\} \\
&\quad + kE\left|\log f(X_i,\theta_0,\rho_0) - \log f(X_i,\theta_0)\right|.
\end{aligned}$$

The maximum of the first term is $O_p(C_n)$ by the Kolmogorov Maximal Inequality for martingales. The second term is uniformly smaller than C_n times the expectation.

For $k > C_n$, we will use Taylor's expansion and the usual law of large numbers. Define

$$g(x,\theta_0) = \sup_{\theta\in N(\theta_0)} \frac{\partial^2 \log f(x,\theta)}{\partial\theta^2}.$$

Under regularity conditions, $E[g(X,\theta_0)]$ converges to $E[\partial^2\log f(X,\theta)/\partial\theta^2]|_{\theta=\theta_0}$ which is negative (of Fisher information) as $N(\theta_0)$ shrinking to θ_0. Hence, $E[g(X,\theta_0)] < 0$ for sufficiently small $N(\theta_0)$. Let $t = \sqrt{k}(\theta - \theta_0)$. Then we get, for some ζ between θ and θ_0,

$$\sum_{i=1}^{k}[\log f(X_i,\theta) - \log f(X_i,\theta_0)]$$

$$
\begin{aligned}
&= \sum_{i=1}^{k} \frac{\partial \log f(X_i,\theta_0)}{\partial\theta}(\theta-\theta_0) + \frac{1}{2}\sum_{i=1}^{k} \frac{\partial^2 \log f(X_i,\zeta)}{\partial\theta^2}(\theta-\theta_0)^2 \\
&\leq \frac{1}{\sqrt{k}}\sum_{i=1}^{k} \frac{\partial \log f(X_i,\theta_0)}{\partial\theta}\cdot t + \frac{1}{2k}\sum_{i=1}^{k} g(X_i,\theta_0)\cdot t^2. \qquad (13)
\end{aligned}
$$

From Kolmogorov Maximum Inequality for reversed martingales under R2 (Sen and Singer, 1993, pp. 81)

$$
\max_{k>C_n}\left|\frac{1}{k}\sum_{i=1}^{k} g(X_i,\theta_0) - Eg(X,\theta_0)\right| = o_p(1).
$$

Recall that $E[g(X,\theta_0)] < 0$. Thus, the quadratic function of t in (13) is bounded by a constant times

$$
\max_{k>C_n}\left\{\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k} \frac{\partial \log f(X_i,\theta_0)}{\partial\theta}\right]^2 \Big/ \left|\frac{1}{k}\sum_{i=1}^{k} g(X_i,\theta_0)\right|\right\}. \qquad (14)
$$

Note that $\frac{\partial \log f(X_i,\theta_0)}{\partial\theta}$ has mean zero, and condition R3 implies that it has finite third moment. By Theorem 1 in Darling and Erdös (1956) on the maximum of normalized sums of independent random variables under finite third moment, we can easily get

$$
\max_{k>C_n}\frac{1}{\sqrt{k}}\sum_{i=1}^{k} \frac{\partial \log f(X_i,\theta_0)}{\partial\theta} = o_p((\log\log n)^{1/2}).
$$

Hence, (14) has order $o_p(\log\log n)$. The proof is similar for $d \geq 2$. This completes the proof of Lemma 1C.

Proof of Lemmas 1 and 2: Note that

$$
\begin{aligned}
\max_k\left[\ell_n(\theta_1,\theta_2,k) - \ell_n(\theta_0,\theta_0,n)\right] &= \max_k\left\{\left[\sum_{i=1}^{k}\log f(X_i,\theta_1) - \sum_{i=1}^{k}\log f(X_i,\theta_0)\right]\right. \\
&\quad \left. + \left[\sum_{i=k+1}^{n}\log f(X_i,\theta_2) - \sum_{i=k+1}^{n}\log f(X_i,\theta_0)\right]\right\} \qquad (15)
\end{aligned}
$$

Lemmas 1A, 1B, 1C conclude that the first term in (15) is less than $o_p(\log\log n)$ when θ_1 is close to θ_0 i.e., in $N(\theta_0)$; far away from θ_0, i.e. $\|\theta-\theta_0\| > r$; or in between ($\theta \in A$). The second term is symmetric to the first term in k, hence the same conclusion holds. This proves Lemma 2.

Now we prove the first part of Lemma 1. Similar to the proof in Wald (1949), we need only show that when $(\tilde{\theta}_1,\tilde{\theta}_2)$ is not in an open neighborhood of (θ_0,θ_0),

$$
\sup[\ell_n(\theta_1,\theta_2,k) - \ell_n(\theta_0,\theta_0,n)] < 0 \qquad (16)
$$

in probability uniformly for all k such that $\min(k, n-k)/\sqrt{n} \to \infty$, where the sup is taken over a very small neighborhood of $(\tilde{\theta}_1, \tilde{\theta}_2)$. After this, we use the compactness to conclude that $(\hat{\theta}_{1k}, \hat{\theta}_{2k})$ has diminishing probability to be outside of any open neighborhood of (θ_0, θ_0).

Let the range of sup in (16) be defined by $(\theta_1 - \tilde{\theta}_1)^2 + (\theta_2 - \tilde{\theta}_2)^2 < \rho^2$. Let $Y_i = \log f(X_i, \tilde{\theta}_1, \rho) - \log f(X, \theta_0)$ and $Z_i = \log f(X_i, \tilde{\theta}_2, \rho) - \log f(X, \theta_0)$ and ρ be small enough such that $EY_1 + EZ_1 < 0,\ EY_1 \leq 0$ and $EZ_1 \leq 0$.

By Kolomgorov Maximal Inequality again,

$$\begin{aligned} \sup[\ell_n(\theta_1, \theta_2, k) - \ell_n(\theta_0, \theta_0, n)] &\leq 2\sum_{i=1}^{k}(Y_i - EY_i) + 2\sum_{i=k+1}^{n}(Z_i - EZ_i) \\ &\quad +2kEY_1 + 2(n-k)EZ_1 \\ &\leq 2\min(k, n-k)(EY_1 + EZ_1) + O_p(n^{1/2}). \end{aligned}$$

Since $EY_1 + EZ_1 < 0$ and $\min(k, n-k)/\sqrt{n} \to \infty$, we have shown (16). The proof of the second part of Lemma 1 is similar. Thus we complete the proof of Lemma 1.

Before we start proving Lemma 3, define

$$MIC(\theta_1, \theta_2; k) = -2l_n(\theta_1, \theta_2; k) + [2\dim(\theta_1) + (\frac{2k}{n} - 1)^2]\log n.$$

Obviously, $MIC(k) = MIC(\hat{\theta}_{1k}, \hat{\theta}_{2k}; k) \leq MIC(\theta_1, \theta_2, k)$, for any $\theta_1,\ \theta_2 \in \Theta$.

Proof of Lemma 3: For any $\epsilon > 0$, define $\Delta = \{k : |\frac{k}{n} - \frac{1}{2}| < \epsilon\}$. Since the penalty term on the location of change point in MIC disappears if $\tau = n/2$ and $l_n(\theta_0, \theta_0, n/2) = l_n(\theta_0, \theta_0, n)$, it is seen that

$$\begin{aligned} P\{\hat{\tau} \notin \Delta\} &\leq P\{\min_{k\notin\Delta} MIC(k) \leq MIC(\theta_0, \theta_0, n/2)\} \\ &= P\{2\max_{k\notin\Delta}[l_n(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) - (2k/n - 1)^2\log n] \geq 2l_n(\theta_0, \theta_0, n/2)\} \\ &\leq P\{\max_{k\notin\Delta}[l_n(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) - l_n(\theta_0, \theta_0, n)] \geq C\epsilon^2\log n\} \\ &\to 0 \end{aligned}$$

by noting that

$$\max_{k\notin\Delta}[l_n(\hat{\theta}_{1k}, \hat{\theta}_{2k}, k) - l_n(\theta_0, \theta_0, n)] \leq o_p(\log\log n).$$

This proves Lemma 3.

PROOF OF THEOREM 1: We first prove the result when $d = 1$.

Lemma 3 implies that the range of k/n can be restricted to an arbitrarily small neighborhood of value $\frac{1}{2}$. If k is in such a neighborhood, we have $\min(k, n-k)/\sqrt{n} \to \infty$. Lemma 1 then enables us to consider only θ_1, θ_2 and θ an arbitrarily small neighborhood of true value θ_0. Mathematically, it means that for any $\varepsilon > 0$, $\delta > 0$,

$$\begin{aligned} S_n &= 2 \max_{|\frac{k}{n}-\frac{1}{2}|<\varepsilon} \left\{ \left[\sup_{|\theta_1-\theta_0|<\delta} \sum_{i=1}^{k} \log f(X_i, \theta_1) + \sup_{|\theta_2-\theta_0|<\delta} \sum_{i=k+1}^{n} \log f(X_i, \theta_2) \right. \right. \\ &\quad \left. \left. - \sup_{|\theta-\theta_0|<\delta} \sum_{i=1}^{n} \log f(X_i, \theta) \right] - \frac{1}{2} \left(\frac{2k}{n} - 1 \right)^2 \log n \right\} + o_p(1) \\ &\leq 2 \max_{|\frac{k}{n}-\frac{1}{2}|<\varepsilon} \left\{ \left[\sup_{|\theta_1-\theta_0|<\delta} \sum_{i=1}^{k} \log f(X_i, \theta_1) + \sup_{|\theta_2-\theta_0|<\delta} \sum_{i=k+1}^{n} \log f(X_i, \theta_2) \right. \right. \\ &\quad \left. \left. - \sup_{|\theta-\theta_0|<\delta} \sum_{i=1}^{n} \log f(X_i, \theta) \right] \right\} + o_p(1). \end{aligned} \tag{17}$$

Let $\tilde{\theta}_{1k}, \tilde{\theta}_{2k}$ and $\tilde{\theta}$ be the maximum points of $\sum_{i=1}^{k} \log f(X_i, \theta)$, $\sum_{i=k+1}^{n} \log f(X, \theta)$ and $\sum_{i=1}^{n} \log f(X_i, \theta)$ in the range $(\theta_0 - \delta,\ \theta_0 + \delta)$ respectively. Whenever θ equals one of $\tilde{\theta}_{1k}$, $\tilde{\theta}_{2k}$ or $\tilde{\theta}$, there exists ζ such that $|\zeta - \theta_0| < \delta$ and

$$\begin{aligned} \sum \log f(X_i, \theta) &= \sum \log f(X_i, \theta_0) + \sum \frac{\partial \log f(X_i, \theta_0)}{\partial \theta} (\theta - \theta_0) \\ &\quad + \frac{1}{2} \sum \frac{\partial^2 \log f(X_i, \theta_0)}{\partial \theta^2} (\theta - \theta_0)^2 \\ &\quad + \frac{1}{6} \sum \frac{\partial^3 \log f(X_i, \zeta)}{\partial \theta^3} (\theta - \theta_0)^3 \end{aligned} \tag{18}$$

where the range of summation is from $i = 1$ to k when $\theta = \tilde{\theta}_{1k}$; from $i = k+1$ to n when $\theta = \tilde{\theta}_{2k}$; or from $i = 1$ to n when $\theta = \tilde{\theta}$. The regularity condition R2 on the third derivative implies that

$$\sum \frac{\partial^3 \log f(X_i, \zeta)}{\partial \theta^3} (\tilde{\theta} - \theta_0)^3 = O_p(n(\tilde{\theta} - \theta_0)^3).$$

This term is negligible compared to

$$\sum \frac{\partial^2 \log f(X_i, \theta_0)}{\partial \theta^2} (\tilde{\theta} - \theta_0)^2$$

which is of order $n(\tilde{\theta} - \theta_0)^2$ in each case, as δ is arbitrarily small.

Thus, by ignoring the high order terms in (18) and substituting it into (17), we obtain

$$\begin{aligned} S_n &\leq \max_{|\frac{k}{n}-\frac{1}{2}|<\varepsilon} \left[\sup_{|\theta_1-\theta_0|<\delta} P_n(1,k,\theta_1) + \sup_{|\theta_2-\theta_0|<\delta} P_n(k+1,n,\theta_2) \right. \\ &\quad \left. - \sup_{|\theta-\theta_0|<\delta} P_n(1,n,\theta) \right] + o_p(1) \end{aligned} \tag{19}$$

where

$$P_n(k_1,k_2,\theta) = 2\sum_{i=k_1}^{k_2} \frac{\partial \log f(X_i,\theta_0)}{\partial\theta}(\theta-\theta_0) + \sum_{i=k_1}^{k_2} \frac{\partial^2 \log f(X_i,\theta_0)}{\partial\theta^2}(\theta-\theta_0)^2.$$

By the Kolmogorov Maximum Inequality for reversed martingales,

$$\begin{aligned} \max_{|\frac{k}{n}-\frac{1}{2}|<\varepsilon} \frac{1}{k}\sum_{i=1}^{k} \frac{\partial^2 \log f(X_i,\theta_0)}{\partial\theta^2} &= E\left[\frac{\partial^2 \log f(X_i,\theta_0)}{\partial\theta^2}\right] + o_p(1) \\ &= -E\left[\frac{\partial \log f(X_i,\theta_0)}{\partial\theta}\right]^2 + o_p(1) \\ &= -I(\theta_0) + o_p(1) \end{aligned} \tag{20}$$

where $I(\theta_0)$ is the Fisher information. Using the property of the quadratic function, we have

$$\sup_{|\theta-\theta_0|<\delta} P_n(1,k,\theta) = I^{-1}(\theta_0)\left[\frac{1}{\sqrt{k}}\sum_{i=1}^{k}\frac{\partial \log f(X_i,\theta_0)}{\partial\theta}\right]^2 + o_p(1).$$

Similarly for the other two terms in (19). Let $Y_i = \frac{\partial \log f(X_i;\theta_0)}{\partial\theta}$ and $W_k = \sum_{i=1}^{k} Y_k$. Then

$$\begin{aligned} S_n &\leq \max_{|\frac{k}{n}-\frac{1}{2}|<\varepsilon} I^{-1}(\theta_0)[\{k^{-1/2}W_k\}^2 + \{(n-k)^{-1/2}(W_n-W_k)\}^2 - \{n^{-1/2}W_n\}^2] + o_p(1) \\ &= \max_{|\frac{k}{n}-\frac{1}{2}|<\varepsilon} I^{-1}(\theta_0)[nt_k(1-t_k)]^{-1}(W_k - t_kW_n)^2 + o_p(1) \\ &\leq \max_{|t-\frac{1}{2}|<\varepsilon} I^{-1}(\theta_0)T_n^2(t) + o_p(1) \end{aligned} \tag{21}$$

where $t_k = k/n$ and

$$T_n(t) = \{\frac{[nt]}{n}(1-\frac{[nt]}{n})\}^{-1/2}n^{-1/2}\{W_{[nt]} + (nt-[nt])Y_{[nt]+1} - \frac{[nt]}{n}W_n\}.$$

Without loss of generality, we also assume $I(\theta_0) = 1$. By Donsker's theorem (Csörgö and Révész, 1981, pp. 13), as $n \to \infty$, for $t \in \left[\frac{1}{2}-\epsilon, \frac{1}{2}+\epsilon\right]$, $T_n(t) \to [t(1-t)]^{-1/2}B_0(t)$ in distribution as a random continuous function, and $B_0(t)$ is a Brownian bridge. As a consequence, as $n \to \infty$, we have

$$\sup_{|t-\frac{1}{2}|\leq\epsilon} T_n^2(t) \to \sup_{|t-\frac{1}{2}|\leq\epsilon} [t(1-t)]^{-1}B_0^2(t)$$

in distribution.

Consequently, from (21) we have shown that

$$S_n \leq \sup_{|t-\frac{1}{2}|<\varepsilon} T_n^2(t) + o_p(1) \rightarrow \sup_{|t-\frac{1}{2}|<\varepsilon} [t(1-t)]^{-1} B_0^2(t). \tag{22}$$

As $\varepsilon \rightarrow 0$, the P. Lévy modulus of continuity of the Wiener process implies,

$$\sup_{|t-\frac{1}{2}|\leq\epsilon} |B_0(t) - B_0(\frac{1}{2})| \rightarrow 0$$

almost surely. Since $\varepsilon > 0$ can be chosen arbitrarily small, and

$$\left[\frac{1}{2}\left(1-\frac{1}{2}\right)\right]^{-1} B_0^2(\frac{1}{2}) \sim \chi_1^2,$$

(22) implies

$$\underline{\lim}_{n\rightarrow\infty} P\{S_n \leq s\} \geq P\{\chi_1^2 \leq s\}$$

for all $s > 0$. It is straightforward to show that $S_n \geq MIC(n) - MIC(n/2) + d\log(n) \rightarrow \chi_1^2$. Thus,

$$\overline{\lim}_{n\rightarrow\infty} P\{S_n \leq s\} \leq P\{\chi_1^2 \leq s\}.$$

Hence, $S_n \rightarrow \chi_1^2$ in distribution as $n \rightarrow \infty$.

Consider the situation when θ has dimension $d \geq 2$. The current proof is valid up to (18). All we need after this point is to introduce vector notation. The subsequent order comparison remains the same as the Fisher Information is positive definite by the regularity conditions. This strategy also works for (21). Note that Y_k is now a vector. Reparameterizing the model so that the Fisher information is an identity matrix under the null hypothesis, and consequently the components of Y_k are un-correlated. We remark here that the test statistic remains unchanged as the likelihood method is invariant to re-parameterization. The term $I^{-1}(\theta_0)T_n^2(t)$ in (21) becomes $T_{n1}^2(t) + T_{n2}^2(t) + \cdots + T_{nd}^2(t)$.

Note that that finite dimensional distribution of $T_{n1}(t), T_{n2}(t), \cdots, T_{nd}(t)$ converge to that of independent Brownian motions $B_1(t), B_2(t), \ldots, B_d(t)$. By Condition R3 and the inequality in Chow and Teicher (1978, pp 357), it is easy to verify

$$\sum_{j=1}^{d} E|T_{nj}(s) - T_{nj}(t)|^3 \leq C|s-t|^{3/2}$$

for some constant C. Thus, $T_{n1}(t), T_{n2}(t), \cdots, T_{nd}(t)$ is tight. In conclusion, the multidimensional process $T_{n1}(t), T_{n2}(t), \cdots, T_{nd}(t)$ with continuous sample path (after some harmless smoothing) converges to $B_1(t), B_2(t), \ldots, B_d(t)$ in distribution according to Revuz and Yor (1999, pp 448-449). Hence, we have $S_n \to \chi^2_d$ in distribution as $n \to \infty$. This proves the conclusion of Theorem 1 for the null hypothesis.

To prove the conclusion of Theorem 1 under the alternative hypothesis, note that

$$\begin{aligned} S_n &\geq MIC(n) - MIC(\tau) + \dim(\theta) \log n \\ &\geq 2[\ell_n(\theta_{10}, \theta_{20}, \tau) - \ell_n(\hat{\theta}, \hat{\theta}, n)] + O(\log n) \end{aligned}$$

where τ is the change point and θ_{10} and θ_{20} be true values of the parameters. Hence $|\theta_{10} - \theta_{20}| > 0$. It is well known (Serfling, 1980)

$$2 \sup_{\theta} \sum_{i=1}^{\tau} \log\{f(X_i, \theta)/f(X_i, \theta_{10})\} \to \chi^2_d \tag{23}$$

in distribution when $\tau \to \infty$. Also, by Wald (1949),

$$\inf_{\theta:|\theta-\theta_{10}|\geq\delta} \sum_{i=1}^{\tau} \log\{f(X_i, \theta_{10})/f(X_i, \theta)\} \geq C_1\tau + o_p(\tau) \tag{24}$$

for any $\delta > 0$ with some positive constant C_1. Similarly for the sum from $\tau + 1$ to n.

Let $\delta = |\theta_{10} - \theta_{20}|/3$. Divide the parameter space into three regions: $A_1 = \{\theta : |\theta - \theta_{10}| \leq \delta\}$, $A_2 = \{\theta : |\theta - \theta_{20}| \leq \delta\}$, and A_3 be the complement of $A_1 \cup A_2$. We have

$$\inf_{\theta \in A_j} [\ell_n(\theta_{10}, \theta_{20}, \tau) - \ell_n(\theta, \theta, n)] \geq C_2 \min\{\tau, n - \tau\}$$

for $j = 1, 2, 3$ with some positive constant C_2. For example, when $j = 1$, (24) implies part of ℓ_n is larger than $C_1\tau$, and (23) implies the other part is of order 1. Hence, when both τ and $n - \tau$ go to infinity at a rate faster than $\log n$, $MIC(n) - MIC(\tau) + \dim(\theta) \log n$ go to infinity. This completes the proof.

Remark: Even though $\{X_{in}, 1 \leq i \leq \tau, n \geq 2\}$ under the alternative model becomes a triangle array with two sets of iid random variables, classical weak convergence results for a single sequence of iid random variables such as in Serfling (1980) and Wald (1949) can be applied to each set for the reasons discussed earlier.

Proof of Theorem 2: The proof will be done in two steps.

STEP 1. We first show that $\hat{\tau} - \tau = O_p[n(\log n)^{-1}]$, or equivalently

$$Pr\{\hat{\tau} \notin A(n)\} \to 1, \text{ as } n \to \infty,$$

for $A(n) = \{k : 0 < k < n, |k - \tau| > n(\log n)^{-1}\}$.

For this purpose, it suffices to show that,

$$P\{MIC(k) > MIC(\tau), \text{ for all } k \in A(n)\} \to 1.$$

Since $MIC(\tau) \leq MIC(\theta_{10}, \theta_{20}; \tau)$, it will be an easy consequence of

$$MIC(\hat{\theta}_{1k}, \hat{\theta}_{2k}; k) - MIC(\theta_{10}, \theta_{20}; \tau) \geq Cn(\log n)^{-1} + o_p[n(\log n)^{-1}] \tag{25}$$

for some constant $C > 0$ uniformly for $k \in A(n)$. Hence, we prove (25) instead.

Consider the case $k < \tau - n(\log n)^{-1}$. We have

$$\begin{aligned} & MIC(\hat{\theta}_{1k}, \hat{\theta}_{2k}; k) - MIC(\theta_{10}, \theta_{20}; \tau) \\ = \; & 2\sum_{i=1}^{k}[\log f(X_i, \theta_{10}) - \log f(X_i, \hat{\theta}_{1k})] + 2\sum_{i=k+1}^{\tau}[\log f(X_i, \theta_{10}) - \log f(X_i, \hat{\theta}_{2k})] \\ & +2\sum_{i=\tau+1}^{n}[\log f(X_i, \theta_{20}) - \log f(X_i, \hat{\theta}_{2k})] + [(\frac{2k}{n} - 1)^2 - (\frac{2\tau}{n} - 1)^2]\log n. \end{aligned}$$

Let us call four terms as R_j, $j = 1, 2, 3$ and R_0. Thus, the difference in (25) is divided into four parts: the difference in penalty, R_0; the differences in the log-likelihoods for observations between 1 and k, R_1; between k and τ, R_2; and between τ and n, R_3. Obviously $R_0 = O(\log n)$. Since X_j, $j = 1, \ldots, k$ are iid observations, $R_1 = O_p(1)$ under regularity conditions as it is no larger than the ordinary likelihood ratio statistic.

Both R_2 and R_3 are sum of at least $n(\log n)^{-1}$ terms. They share the same parameter estimator $\hat{\theta}_{2k}$ which cannot be close to both θ_{10} and θ_{20} as $\theta_{10} \neq \theta_{20}$. Hence, there exists $\rho > 0$, either $|\theta_{10} - \hat{\theta}_{2k}| \geq \rho$ or $|\theta_{20} - \hat{\theta}_{2k}| \geq \rho$. If it is the former, then as $n \to \infty$, by Wald(1949),

$$R_2 \geq Cn(\log n)^{-1} + o_p[n(\log n)^{-1}]$$

while $R_3 \geq O_p(1)$. If it is the latter, we simply switch R_2 and R_3 in the above. That is, (25) is proved when $k < \tau - n(\log n)^{-1}$. The proof for the case of $k > \tau + n(\log n)^{-1}$ is the same. Hence, we have completed step 1.

STEP 2. We show $\hat{\tau} - \tau = O_p(1)$.

According to Step 1, the convergence rate of $\hat{\tau}$ is at least $O_p[n(\log n)^{-1}]$. We need only tighten this rate to obtain the result of the theorem. For this purpose, it suffices to show

$$MIC(\hat{\theta}_{1k}, \hat{\theta}_{2k}; k) > MIC(\theta_{10}, \theta_{20}; \tau)$$

with probability tending to 1 uniformly for $\tau - n(\log n)^{-1} \leq k \leq \tau - M$, and for $\tau + M \leq k \leq \tau + n(\log n)^{-1}$. We only give the proof for the first case.

We use the same strategy as in Step 1. Due to the narrower range of k under consideration, we have $R_0 = O(1)$ in the current case. Since X_j, $j = 1, \ldots, k$ are iid observations, $R_1 = O_p(1)$ under regularity conditions as it is no larger than the ordinary likelihood ratio statistic. Similarly, $R_3 = O_p(1)$.

The focus is now on R_2. We intend to show that $R_2 > CM + M \cdot o_p(1)$ for some constant $C > 0$ uniformly in the range of consideration. If so, we could choose large enough M and claim that the probability of $MIC(\hat{\theta}_{1k}, \hat{\theta}_{2k}; k) > MIC(\theta_{10}, \theta_{20}; \tau)$ in the range of consideration is larger than $1 - \epsilon$ for any pre-given $\epsilon > 0$.

By Lemma 1 since $|k - \tau| < n(\log n)^{-1}$, $\hat{\theta}_{2k} \to \theta_{20} \neq \theta_{10}$. Thus, we may assume that it falls outside a small neighborhood of θ_{10}. According to Wald(1949), this implies, as $\tau - k \geq M$,

$$R_2 = \sum_{i=k+1}^{\tau} [\log f(X_i; \theta_{10}) - \log f(X_i; \hat{\theta}_{2k})] \geq CM$$

for some $C > 0$ in probability as $M \to \infty$. Consequently, the theorem is proved.

PROOF OF THEOREM 3: It suffices to show that for any given $M > 0$,

$$MIC(\tau + k) - MIC(\tau) \to W_k \tag{26}$$

in probability uniformly for all k such that $|k| \leq M$.

Keep in mind that M is finite, for $-M \leq k \leq 0$,

$$\begin{aligned} & MIC(\tau + k) - MIC(\tau) \\ = \; & 2\sum_{i=1}^{\tau} \log f(X_i, \hat{\theta}_{1\tau}) + 2\sum_{i=\tau+1}^{n} \log f(X_i, \hat{\theta}_{2\tau}) - 2\sum_{i=1}^{\tau+k} \log f(X_i, \hat{\theta}_{1(\tau+k)}) \\ & -2\sum_{i=\tau+k+1}^{n} \log f(X_i, \hat{\theta}_{2(\tau+k)}) + [(\frac{2(\tau+k)}{n} - 1)^2 - (\frac{2\tau}{n} - 1)^2] \log n \end{aligned}$$

$$
\begin{aligned}
&= 2\sum_{i=\tau+k+1}^{\tau}[\log f(X_i,\hat{\theta}_{1\tau})-\log f(X_i,\hat{\theta}_{2(\tau+k)})]+2\sum_{i=1}^{\tau+k}[\log f(X_i,\hat{\theta}_{1\tau})-\log f(X_i,\hat{\theta}_{1(\tau+k)})]\\
&\quad +2\sum_{i=\tau+1}^{n}[\log f(X_i,\hat{\theta}_{2\tau})-\log f(X_i,\hat{\theta}_{2(\tau+k)})]+o_p(1). \qquad (27)
\end{aligned}
$$

By Lemma 1, $\hat{\theta}_{1\tau}\to\theta_{10},\hat{\theta}_{2(\tau+k)}\to\theta_{20}$ uniformly for all k such that $|k|\le M$. Hence, the first term

$$
\sum_{i=\tau+k+1}^{\tau}[\log f(X_i,\hat{\theta}_{1\tau})-\log f(X_i,\hat{\theta}_{2(\tau+k)})]=W_k+o_p(1).
$$

For the second term, we have,

$$
\begin{aligned}
&2\sum_{i=1}^{\tau+k}[\log f(X_i,\hat{\theta}_{1\tau})-\log f(X_i,\hat{\theta}_{1(\tau+k)})]\\
&= 2\sum_{i=1}^{\tau}[\log f(X_i,\hat{\theta}_{1\tau})-\log f(X_i,\theta_{10})]-2\sum_{i=\tau+k+1}^{\tau}[\log f(X_i,\hat{\theta}_{1\tau})-\log f(X_i,\theta_{10})]\\
&\quad -2\sum_{i=1}^{\tau+k}[\log f(X_i,\hat{\theta}_{1(\tau+k)})-\log f(X_i,\theta_{10})]\\
&\hat{=} P_1-P_{k2}-P_{k3}.
\end{aligned}
$$

Since it has at most M terms and $\hat{\theta}_{1\tau}\to\theta_{10}$, $P_{k2}=o_p(1)$ uniformly for $-M\le k\le 0$. For other two terms, we have expansions

$$
P_1=\frac{1}{I(\theta_{10})}[\frac{1}{\sqrt{\tau}}\sum_{i=1}^{\tau}\frac{\partial\log f(X_i,\theta_{10})}{\partial\theta}]^2+o_p(1)
$$

and

$$
P_{k3}=\frac{1}{I(\theta_{10})}[\frac{1}{\sqrt{\tau+k}}\sum_{i=1}^{\tau+k}\frac{\partial\log f(X_i,\theta_{10})}{\partial\theta}]^2+o_p(1),
$$

for all $-M\le k\le 0$. The difference between P_1 and P_{k3} is seen to be $o_p(1)$ and so is the second term in (27).

The last term in (27) is $o_p(1)$ for the same reason. Thus, the limiting distribution of $\hat{\tau}$ is solely determined by the first term of (27) which gives the result of the theorem.

Acknowledgment

The research of Jiahua Chen is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

References

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle, *2nd Int. Symp. Inf. Theory* (B. N. Petrov and E. Csaki, Eds.). Budapest: Akademiai Kiado, 267-281.

Basseville, M. and Nikiforov, I. V. (1993). Detection of Abrupt Changes: Theory and Application, *PTR Prentice-Hall*.

Bhattacharya, P. K. (1994). Some Aspects of Change-Point Analysis, In *Change-Point Problems*, **23**, 28-56.

Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric Methods in Change-Point Problems*. Kluwer Academic Publishers, The Netherlands.

Chen, Jie and Gupta, A. K. (1997). Testing and locating variance change points with application to stock prices, *J. Amer. Statist. Assoc.*, **92**, 739-747.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to change in time, *Ann. Math. Statist.*, **35**, 999-1018.

Chow, Y. and Teicher, H. (1978). *Probability theory, independence, interchangeability, Martingales*. New York: Springer-Verily.

Csörgö, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. John Wiley & Sons, New York.

Csörgö, M. and Révész, P. (1981). *Strong approximations in probability and statistics*. Academic Press, New York.

Darling, D. A. and Erdös, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables, *Duke Math. J.* **23** 143-155.

Gombay, E. and Horváth, L. (1996). On the rate of approximations for maximum likelihood tests in change-point models, *J. Multi. Analy.* **56** 120-152.

Gombay, E. and Horváth, L. (1995). An application of U-statistics to change-point analysis *Acta. Sci. Math.* **60** 345-357.

Gombay, E. and Hušková, M. (1998). Rank based estimators of the change-point, *J. Statist. Plann. Inf.* **67** 137-154.

Hinkley, D. V. (1970). Inference about the change point in a sequence of random variables, *Biometrika*, **57**, 1-7.

Hinkley, D. V., Chapman, P., and Rungel, G. (1980). Changepoint problems, Technical Report 382. University of Minnesota, Minneapolis, MN.

Hirotsu, C., Kuriki, S., and Hayter, A. J. (1992). Multiple comparison procedures based on the maximal component of the cumulative chi-squared statistic, *Biometrika*, **79**, 381-392.

Inclán, C., and Tiao, G. C. (1994), Use of sums of squares for retrospective detection of changes of variance, *J. Amer. Statist. Assoc.,* **89**, 913-923.

Lai, T. L. (1995). Sequential Changepoint detection in Quality Control and Dynamical Systems, *J. R. Statist. Soc. B.*, **57**, 613-658.

Lombard, F. (1987). Rank tests for change point problems, *Biometrika*, **74**, 615-624.

Page, E. S. (1954). Continuous Inspection Schemes, *Biometrika*, bf 41, 100-115.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point, *Biometrika*, bf 42, 523-527.

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point, *Biometrika*, bf 44, 248-252.

Revuz, D. and Yor, M. (1999). *Continuous Martingales and Brownian Motion*. Second Edition. Springer-Verlag. New York.

Schwarz, G. (1978). Estimating the dimension of a model, *Ann. Statist.*, **6**, 461-464.

Sen, P. K. and , Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, New York.

Shorack, G. R. and Wellner (1986). *Empirical Processes with Applications to Statistics*. John Wiley & Sons, New York.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate, *Ann. Math. Statist.*, **20**, 595-601.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations, *J. Amer. Statist. Assoc.*, **74**, 365-367.

Worsley, K. J. (1986). Confidence regions and tests for a change point in a sequence of exponential family random variables, *Biometrika*, **73**, 91-104.

Wu, W. B,, Woodroofe, M., and Mentz, G. (2001), Isotonic regression: Another look at the changepoint problem, *Biometrika*, **88(3)**, 793-804.

Zacks, S. (1983). Survey of classical and bayesian approaches to the change-point problem: Fixed sample and sequential procedures in testing and estimation, In *Recent Advances in Statistics*, Academic Press, 245-269.