Some Aspects of Statistical Volatility Analysis

MIN XU

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN
MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO
AUGUST 2015

Abstract

Volatility is the key of the option price in the stock market. Changes in volatility will dramatically lead to changes of the option price.

One of the most important volatilities is historical volatility (HV). The HV is essentially the annualized standard deviation of the first order difference of logarithm of the asset price. Therefore, changes in HV in finance may be detected by the variance change detection methods in statistics.

We propose a weighted sum of powers of variances method to detect single change in HV. It is noted that this method only examines if there is one single change-point in the data sequence. In the second part of the dissertation, we propose the empirical Bayesian information criterion (emBIC) method to detect multiple change-points simultaneously. The empirical BIC method can not only detect change-points in HV, but also in mean, and mean-and-variance. Simulation study shows that both of the above methods perform very well. We also apply these methods to detect changes in HV by using real stock data.

Another important volatility is the implied volatility (IV). IV is the volatility of asset implied by the market option price based on Black-Sholes model [Black and Scholes, 1973]. The long term IV and HV have totally different behaviours. We find the optimal time range by using the emBIC method aforementioned above. We explain the long term IV behaviour by interest rate risk and capital charge in the last part of the dissertation.

Acknowledgment

First of all, I would like to express my deepest appreciation and sincere gratitude to Dr. Yuehua Wu who has been my supervisor since the beginning of my PhD study. She not only taught me knowledge, but also gave me many useful suggestions and advices. Without her persistent help and patient instruction, this dissertation would not be finished.

I would express my sincere gratitude to Dr. Hong Xie at Manulife Financial as a member of my supervisory committee. He has spent numerous his own time in guiding and instructing me. He gave me many ideas and suggestions in the area of financial statistics. Without his superior knowledge and experience in financial area, the dissertation would not be complete.

In addition, I wish to express my appreciation to Dr. Huaiping Zhu as a member of my supervisory committee and Dr. Guoqi Qian at the University of Melbourne for his great help and suggestion in my research.

I would thank my whole family, my parents, wife, daughter and son. Without their selfless support, I couldn't complete the dissertation.

Finally, I wish to express my appreciation to all who help me finish the dissertation.

TABLE OF CONTENTS

1 Introduction and Outline

1.1 Volatility, Historical Volatility and Implied Volatility

Volatility is a measure of uncertainty about the returns provided by the stock [Hull, 2011]. There are two most important volatilities in the market: historical volatility (HV) and implied volatility (IV).

HV is the volatility of the underlying asset over a period observed in the past or called realized volatility:

$$\sigma_{HV} = \sqrt{252} \cdot \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(r_t - \bar{r})^2}, \tag{1.1}$$

where $r_t = \log(S_t/S_{t-1})$, S_t is the asset price at day t, $\bar{r} = r_t/n$, and 252 is the approximate trading days per year. It is easy to see that HV is essentially an annualized standard deviation of the first order difference of logarithm of the asset price.

[Figlewski, 1997] pointed out that better forecasts are normally obtained if we ignore term $\bar{r}$ in (1.1). Also setting n instead of $n-1$ hardly affects results if n is large, while it makes calculation more convenient. Therefore we can redefine HV as:

$$\sigma^*_{HV} = \sqrt{\frac{252}{n}\sum_{t=1}^{n} r_t^2}. \tag{1.2}$$

In the dissertation, we use (1.2) to calculate the HV.

IV is the volatility of underlying asset implied by the market price of the option based on the Black-Scholes (BS) model [Black and Scholes, 1973].

The BS model assumes that there is a riskless asset with expected return μ and constant volatility σ. The dynamics of the price S of the underlying asset are

$$dS = \mu S dt + \sigma S dB_t,$$

where B_t is a standard Brownian motion and satisfies

$$dB_t = \epsilon\sqrt{dt}$$

with ϵ being the normally distributed with mean 0 and variance 1.

Let $G = \log S$. By Ito lemma [Ito, 1951], we can obtain

$$d \log S = (\mu - \frac{1}{2}\sigma^2)dt + \sigma dB_t.$$

Therefore, the dynamics of price S can be expressed as

$$S_T = S_0 \exp\{(\mu - \frac{1}{2}\sigma^2)T + \sigma B_T\}.$$

Thus, $(\log S_T/S_0)$ has a normal distribution with mean $(\mu - \sigma^2/2)T$ and variance $\sigma^2 T$.

In the dissertation, the call option refers to the European call option. A call option allows the holder to buy the asset at a certain price in a prefixed day. The certain price is the strike price. The prefixed day is the expiration day. The call option price C can be expressed as:

$$C = E[e^{-rT}(S_T - K, 0)^+],$$

where r is risk free interest rate, S_T is the price of asset at time T, and K is the strike price. If we know the distribution of S_T, we can calculate the option price directly. Considering the BS model, S_T has a log-normal distribution. Therefore, the call option price can be calculated by

$$C = S_0 N(d_1) - K e^{-rT} N(d_2), \tag{1.3}$$

where S_0 is the asset price at time 0, r is risk free interest rate, T is expired time, $N(\cdot)$ is the cdf of standard normal distribution, and

$$d_1 = \frac{\log(S_0/K) + (r + \sigma^2/2)/T}{\sigma\sqrt{T}},$$

$$d_2 = d_1 - \sigma\sqrt{T}$$

with σ being the volatility of the asset.

By the BS model, the call option price is calculated by (1.3), in which all parameters except volatility (σ) can be observed in the market. If we treat call option price C as a function of σ, i.e. $C = f(\sigma)$, then σ can be calculated by the inverse function of C, i.e. $\sigma = f^{-1}(C)$. Therefore, we call the volatility σ implied by the market price of the option based on the BS model as implied volatility.

1.2 Changes in Volatility

The volatility is the key of the option price in the stock market. Changes in volatility will dramatically lead to changes in the option price.

(1.1) shows that the HV is essentially the annualized standard deviation of the first order difference of logarithm of the asset price. Therefore, changes in HV may be detected by the variance change detection methods in statistics.

Time series models have been widely used to test change-points in variance in earlier years, see [Wichern *et al.*, 1976]'s first-order autoregressive time series model, and [Tsay, 1988]'s ARIMA model. Detecting changes by Bayesian framework is also popular, see [Inclan, 1993], [Barry and Hartigan, 1993], [Lavielle, 2005], etc.. A well-known method in this field is called CUSUM of squares introduced by [Inclan and Tiao, 1994]. They used the cumulative sum of squares of a series of uncorrelated random variables to find change-points in variance. Furthermore, penalized likelihood approach introduced by [Yao, 1988] is also widely used. [Chen and Gupta, 2012] illustrated the likelihood approach and BIC Informational approach in their book (See Appendix [A-1] in details).

Let X_i, $i = 1, 2, \ldots, n$, be a sequence of independent random variables with mean 0 and σ_i. For the single change-point detection in variance, the corresponding hypotheses are:

$$\text{Null hypothesis: } H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_n$$

and

$$\text{Alternative hypothesis: } H_1 : \sigma_1 = \cdots = \sigma_k \neq \sigma_{k+1} = \cdots = \sigma_n,$$

, where $1 < k < n$ is unknown.

Assume X_i, $i = 1, 2, \ldots, n$, is a sequence of independently and identically distributed (I.I.D.) random variables with mean 0 and variance σ_i^2, $i = 1, 2, \ldots, , n$. If there exists a k^* such that

$$\sigma_i^2 = \begin{cases} \sigma_1^2 & \text{if } 1 \leq i \leq k^*, \\ \sigma_n^2 & \text{if } k^* < i \leq n \end{cases}$$

with $\sigma_1^2 \neq \sigma_n^2$, then we say k^* is the change-point in variance.

For multiple change-point detection, the corresponding null and alternative hypotheses are respectively:

$$H_0 : \sigma_1 = \sigma_2 = \cdots = \sigma_n$$

and

$$H_1 : \sigma_1 = \cdots = \sigma_{j_1} \neq \sigma_{j_1+1} = \cdots = \sigma_{j_2} \neq \cdots \neq \sigma_{j_K+1} = \cdots = \sigma_n,$$

where $1 < j_1 < \cdots < j_K < n$ are unknown positive integers and K is the number of change-points. If $K \neq 0$, we call each j_i $(i = 1, \cdots, K)$ a *change-point location*, and $(j_1, \cdots, j_K; K)$ a *configuration of change-points* in the data sample. If $K = 0$, we say that there is no change-point. In case of multiple changes in variance, the change-points $j_1, \cdots, j_K$ and the number of change-points K are all unknown, and need to be estimated.

Allowing for occurrence of multiple change-points in the data is a more realistic approach. [Vostrikova, 1981] proposed a binary segmentation procedure (BSP) in which a single change-point detection method is first applied to find the most significant change-point in the data sequence, which accordingly divides the sequence into two sub-ones; this method is then repeatedly applied to each sub-sequence; and each of further sub-sequences until no more significant change-point can be detected. [Chen and Gupta, 1997] applied the BSP to test and locate multiple variance change-points, where the problem was formulated as testing a sequence of change-point hypotheses by using Schwarz Bayesian information criterion (SIC or BIC). [Inclan and Tiao, 1994] used a statistic of cumulative sums of squares to test and locate multiple variance change-points in a sequential way similar to BSP. It is not difficult to see that BSP can be implemented to detect multiple change-points of a parameter other than the variance one. However, an undesirable feature of BSP is no guarantee of satisfactory size and power in the associated sequential testing: Once it does not reject the hypothesis of no change-point in a segment it will be impossible to detect change-points in any subsequent sub-segment; also once a data point is wrongly detected as a change-point at certain stage there is no chance for the error to be revoked under BSP in the subsequent stages.

Instead of using the binary segmentation procedure one can perform sequential testing of multiple change-points by progressively testing "H_0: the data sequence has ℓ change-points" versus "H_a: the data sequence has $\ell+1$ change-points", $\ell = 0, 1, 2, \ldots$. [Bai and Perron, 2003]

developed such a procedure for analyzing models with multiple structural changes, where a dynamic programming algorithm was used to reduce the computation load.

Optimally estimating the number and locations of changes is unequivocally an important task in analysis of multiple change-points. This leads to the development of several model selection approaches, including e.g. the Schwarz Bayesian information criterion (SIC or BIC) by [Yao, 1988], unbiased SIC by [Chen and Gupta, 1997], the minimum description length (MDL) criterion by [Davis *et al.*, 2006] and the Akaike's information criterion (AIC) by citepKurozumiTuvaandorj2011. While these criteria have been shown to possess some good asymptotic properties related to consistency and asymptotic unbiasedness under regularity conditions, they tend to overestimate the number of change-points in finite sample situations. Implementing these criteria for estimating multiple change-points is a difficult optimization problem because the number of possible change-points configurations is up to 2^{n-1} for a sequence of n data points. [Davis *et al.*, 2006] developed a genetic algorithm (GA) for this optimization resulting in excellent empirical properties in a number of situations; but as said in that paper, the expectation of GA working well depends on one's belief in Darwin's *Theory of Natural Selection*. A different procedure was proposed in [Lavielle, 2005] and [Lavielle and Teyssiere, 2006] to estimate the number of change-points. This procedure computes a criterion function $J(K)$ and choose the best K as the maximum number at which the second order difference of $J(K)$ is greater than a pre-specified threshold. Since the behaviour of $J(K)$ is mostly data-dependent, it can be a difficult task to determine this pre-specified threshold. A brief review on methods for estimating the number of change-points is provided in section 4 of [Aue and Horváthe, 2013] and references therein.

1.3 Long Term Implied Volatility Behaviour

It is well known that IV and HV have different behaviours. IV is the volatility of underlying asset implied by the market price of the option based on BS model while HV is the volatility of underlying asset over a period observed in the past or called realized volatility. The long term option market IV observed e.g. for 5-year, 10-year, 15-year or longer, tends to increase after 5-year term, while HV tends to decrease over the terms until it converges to a relatively stable level. For example, on the day of Dec 13, 2011, IVs of 5-year, 10-year and 15-year are

28.87%, 30.38%, 31.37%, respectively, while corresponding HVs are 26.62%, 22.01%, 21.41%. It is in general that long term IV is higher than corresponding HV and long term IV increases as the length of the term increases, but HV behaves oppositely.

[Eraker, 2008] studied a general equilibrium model based on long-run risk to explain the volatility premium of 1-year difference between IV and HV. [Bollerslev *et al.*, 2008] and [Zhou, 2010] explained the difference as time-varying economic uncertainty. Also it has been interpreted as the risk aversion by several researchers like [Bakshi and Madan, 2006] and [Bollerslev *et al.*, 2011].

All researchers mentioned above explain why the difference exists, but why does the difference increase as the length of the term increases? In fact, it is the same question as why IV tends to increase after a 5-year term since HV is relatively stable after a 5-year term. [Tehranchi, 2010] gave some theoretical explanation. He proved that a long term implied volatility cannot fall when the expired time goes to infinity by using Dybvig-Ingersoll-Ross theorem [Dybvig *et al.*, 1996] that says a long zero-coupon rate never fall. [Heston, 1993] proposed a stochastic volatility model, which has very similar form as the CIR [Cox, Ingersolla and Ross, 1985] interest rate model. It is essentially the BS model with all assumptions except the stochastic volatility. All of these studies fail to point out into real drivers of the pricing, hedging costs from interest rate risk, cost of capital required for writing derivatives, and some possible other factors such as credit valuation adjustment.

Both [Tehranchi, 2010]'s theory and [Heston, 1993]'s model mentioned above are based on the constant risk free interest rate. [Bakshi *et al.*, 2000] pointed out: It is a common understanding in the literature that stochastic interest rates may not be important for the pricing and hedging of short-term options, but should be so for long-term options. They indicated interest rate risk may affect the long term option pricing but they did not go further to explain the behaviour of the implied volatility term structure.

After examining long term interest rate and IV in the market, we find that the long term IV may not be related to the long term interest rate at the same day. It can be seen from Table 1 that for both June 13 and November 8, 2011, index values, long term IVs and HVs are almost the same, however the long term interest on June 13, 2011 is much higher than the one on November 8, 2011. Also by examining the IV market, it becomes clear that long

term IVs are much more stable than the shorter term IVs. This phenomenon tells us that the long term IV should be related to the long run average interest rates.

Table 1: Long Term IV, HV and Interest Rate

Date	Index	$IV5$	$IV10$	$IV15$	$HV5$	$HV10$	$HV15$	IR5Y	IR10	IR20	IR30
13/06/2011	1271.83	23.02%	26.84%	30.15%	25.04%	21.42%	20.77%	1.59%	3.00%	3.89%	4.20%
08/11/2011	1275.92	27.01%	28.38%	30.02%	26.37%	21.90%	21.31%	0.92%	2.10%	2.84%	3.13%

To capture interest rate risk, we use a widely adopted short interest rate model that describes a process of the interest rate r_t. A stochastic differential equation for r_t has the form

$$dr_t = \mu_r(t)dt + v_r(t)dB_t,$$

where $\mu_r(t)$ is the drift term, $v_r(t)$ is the diffusion term, and B_t is the Brownian motion term.

To capture interest rate risk in options, we price long term options by a system of two stochastic variables: underlying equity and short interest rate. We assume that the underlying equity follows Black-Scholes model [Black and Scholes, 1973] with short interest rate, i.e.

$$\begin{aligned} dS_t &= r_t S_t dt + \sigma_s S_t dB_1, \\ dr_t &= \mu_r(t)dt + v_r(t)dB_2, \\ dB_1 dB_2 &= \rho dt, \end{aligned}$$

where S_t is the asset price, r_t is the interest rate, σ_s is the asset volatility, dB_1 and dB_2 are both Brownian motions with correlation ρ.

Furthermore, we find the capital charge for options has some impact on IV. Without loss of generality, we assume all capital requirements for the call option and corresponding hedging strategy exactly meet the Basel III regulatory (see **Chapter 4.1.5**).

Finally we find there are three most important factors which affect the long term IV behaviour: market expectation for long run average interest rate, equity volatility and capital charge. We can conclude that market expectation for long run average interest rate dominates the behaviour of long term IV.

1.4 Purpose and Outline of the Dissertation

The main purpose of this dissertation is to solve some volatility problems in finance by using statistical tools. We want to find change-points in volatility by proposing some statistical methods. We intend to explain long term implied volatility behaviour.

In Chapter 2, we propose a statistical method for detecting a single change-point in volatility. In Chapter 3, we present a statistical method for locating multiple change-points simultaneously. In Chapter 4, we first find an optimal time range in terms of change-points in volatility by the method introduced in Chapter 3. We then explain the long term implied volatility behaviour. In Chapter 5, we summarize this dissertation and discuss future research.

2 Historical Volatility Change Detection by Weighted Power of Variance

In this chapter, we propose a procedure to detect and estimate single change-point in historical volatility. HV is essentially an annualized standard deviation of the first order difference of logarithm of the asset price. Therefore, a change in HV can be detected by using a variance change detection method.

The basic idea is inspired by the well-known Jensen's inequality [Jensen, 1906]:

$$f(\alpha t_1 + \beta t_2) \leq \alpha f(t_1) + \beta f(t_2),$$

where f is a convex function and $\alpha > 0$, $\beta > 0$, and $\alpha + \beta = 1$.

Let the function $f(t) = t^\lambda$, where $t > 0$ and $\lambda > 1$. Put $\alpha = k/n$, $\beta = (n-k)/n$, $t_1 = \sigma_k^2$, and $t_2 = \sigma_{n-k}^2$, where $1 < k < n$. According to Jensen's inequality, we have

$$\left(\frac{k}{n}\sigma_k^2 + \frac{n-k}{n}\sigma_{n-k}^2\right)^\lambda \leq \frac{k}{n}\left(\sigma_k^2\right)^\lambda + \frac{n-k}{n}\left(\sigma_{n-k}^2\right)^\lambda.$$

Let $\sigma^2 = \frac{k}{n}\sigma_k^2 + \frac{n-k}{n}\sigma_{n-k}^2$, then for $\lambda > 1$,

$$\frac{k}{n}(\sigma_k^2)^\lambda + \frac{n-k}{n}(\sigma_{n-k}^2)^\lambda \geq (\sigma^2)^\lambda.$$

Thus we can propose a method for detecting single change-point in variance, which is named as the weighted sum of powers of variances (WSPV).

This chapter is organized as follows. Section 2.1 introduces weighted sum of powers of variances (WSPV) method. Section 2.2 lists the assumptions, and gives preliminary lemmas needed in the rest of the chapter. Section 2.3 states main results, proof and asymptotic property. Modified weighted sum of powers of variances (MWSPV) is introduced in Section 2.4. Simulation study and a real data analysis are provided in Section 2.5. Section 2.6 extends this method to a general form, i.e., the generalized weighted sum of functions of variances.

2.1 Weighted Sum of Powers of Variances (WSPV)

Let X_i be a sequence of I.I.D. random variables with constant mean 0 and variance σ_i^2, $i = 1, 2, \ldots, n$, respectively. We assume that

$$0 < \sigma_i^2 < \infty \qquad \text{and} \qquad 0 < E[X_i^4] < \infty. \tag{2.1}$$

For the single change-point detection problem, the corresponding hypothesis tests are the null hypothesis $H_0 : \sigma_1 = \cdots = \sigma_n = \sigma$ versus the alternative hypothesis $H_1 : \sigma_1 = \cdots = \sigma_k \neq \sigma_{k+1} = \cdots = \sigma_n$, where $1 < k < n$ is unknown.

Assume there exists one change-point k^* such that

$$\sigma_i^2 = \begin{cases} \sigma_1^2 & \text{if } 1 \leq i \leq k^* \\ \sigma_n^2 & \text{if } k^* < i \leq n, \end{cases}$$

where $\sigma_1^2 \neq \sigma_n^2$.

We define $v_{\lambda,k}$ as weighted sum of powers of variances (WSPV):

$$v_{\lambda,k} = k(\hat{\sigma}_{1,k}^2)^\lambda + (n-k)(\hat{\sigma}_{n,k}^2)^\lambda - n(\hat{\sigma}^2)^\lambda, \tag{2.2}$$

where $1 < k < n, \lambda$ is the power, $\hat{\sigma}_{1,k}^2$ is maximum likelihood estimator (MLE) of σ_1^2, $\hat{\sigma}_{n,k}^2$ is the MLE of σ_n^2 under H_1, and $\hat{\sigma}^2$ is the MLE of σ^2 under H_0, which are

$$\hat{\sigma}_{1,k}^2 = \frac{\sum_{i=1}^k x_i^2}{k}, \hat{\sigma}_{n,k}^2 = \frac{\sum_{i=k+1}^n x_i^2}{n-k}, \hat{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2}{n}. \tag{2.3}$$

We define $\hat{k}$ as

$$\hat{k} = \arg\max_k \{v_{\lambda,k}\}, \quad \lambda > 1 \text{ or } \lambda < 0$$

and

$$\hat{k} = \arg\min_k \{v_{\lambda,k}\}, \quad 0 < \lambda < 1.$$

Here is an example. We generate a sequence of 500 normal random variables. The first 300 variables have mean 0 and variance 1. The remaining 200 have mean 0 and variance 3. Figure 1 shows the behaviour of $v_{\lambda,k}$ when $\lambda = 2$. The maximum value of $v_{\lambda,k}$ comes up at $\hat{k}$.

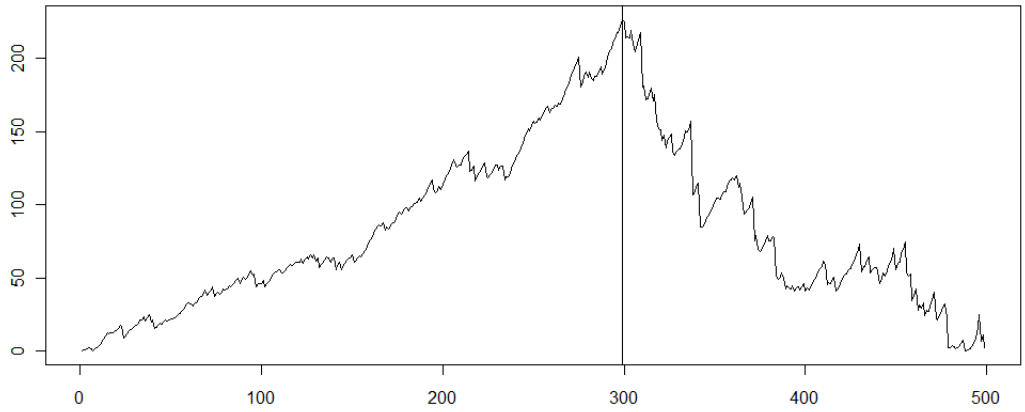

Figure 1: Behaviour of WSPV When $\lambda > 1$

2.2 Preliminaries

We apply a three-term Taylor expansion for $(\hat{\sigma}^2_{1,k})^\lambda$,$(\hat{\sigma}^2_{n,k})^\lambda$, and $(\hat{\sigma}^2)^\lambda$ in (2.2), respectively:

$$\begin{aligned}
(\hat{\sigma}^2_{1,k})^\lambda =&(\sigma_1^2)^\lambda + \lambda(\sigma_1^2)^{\lambda-1}(\hat{\sigma}^2_{1,k} - \sigma_1^2) + \frac{1}{2!}\lambda(\lambda-1)(\sigma_1^2)^{\lambda-2}(\hat{\sigma}^2_{1,k} - \sigma_1^2)^2 \\
&+\frac{1}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta_{1,k}})^{\lambda-3}(\hat{\sigma}^2_{1,k} - \sigma_1^2)^3, \\
(\hat{\sigma}^2_{n,k})^\lambda =&(\sigma_n^2)^\lambda + \lambda(\sigma_n^2)^{\lambda-1}(\hat{\sigma}^2_{n,k} - \sigma_n^2) + \frac{1}{2!}\lambda(\lambda-1)(\sigma_n^2)^{\lambda-2}(\hat{\sigma}^2_{n,k} - \sigma_n^2)^2 \\
&+\frac{1}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta_{n,k}})^{\lambda-3}(\hat{\sigma}^2_{n,k} - \sigma_n^2)^3, \\
(\hat{\sigma}^2)^\lambda =&(\sigma^2)^\lambda + \lambda(\sigma^2)^{\lambda-1}(\hat{\sigma}^2 - \sigma^2) + \frac{1}{2!}\lambda(\lambda-1)(\sigma^2)^{\lambda-2}(\hat{\sigma}^2 - \sigma^2)^2 \\
&+\frac{1}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta})^{\lambda-3}(\hat{\sigma}^2 - \sigma^2)^3,
\end{aligned}$$

where $\sigma^2_{\eta_{1,k}} \in (\min(\hat{\sigma}^2_{1,k}, \sigma_1^2), \max(\hat{\sigma}^2_{1,k}, \sigma_1^2))$, $\sigma^2_{\eta_{n,k}} \in (\min(\hat{\sigma}^2_{n,k}, \sigma_n^2), \max(\hat{\sigma}^2_{n,k}, \sigma_n^2))$, and $\sigma^2_{\eta} \in (\min(\hat{\sigma}^2, \sigma^2), \max(\hat{\sigma}^2, \sigma^2))$.

Denote

$$\begin{aligned}\xi_{1,k} =& \frac{k}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta_{1,k}})^{\lambda-3}(\hat{\sigma}^2_{1,k}-\sigma^2_1)^3,\\ \xi_{n,k} =& \frac{n-k}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta_{n,k}})^{\lambda-3}(\hat{\sigma}^2_{n,k}-\sigma^2_n)^3,\\ \xi \ =& \frac{n}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta})^{\lambda-3}(\hat{\sigma}^2-\sigma^2)^3.\end{aligned}$$

Lemma 2.1 *Let $X_i, i = 1, 2, \ldots, n$ be a sequence of I.I.D. random variables with $E[X_i] = 0$ and $E[X_i^2] = 1$. Let $S_k = \sum_{i=1}^{k} X_i$, then*

$$\lim_{n\to\infty}(2\log\log n)^{-1/2}\max_{1\le k\le n}\frac{|S_k|}{\sqrt{k}} = 1 \quad a.s..$$

See [Chen, 2013] (1.6) for proof.

Lemma 2.2 *Let $X_i, i = 1, 2, \ldots, n$ be a sequence of I.I.D. random variables with $E[X_i] = 0$ and $E[X_i^2] = \sigma_i^2$. Assume (2.1) holds, $\xi, \xi_{1,k}, \xi_{n,k}$ are denoted as above, then we have*

$$\begin{aligned}&(i)\ \lim_{n\to\infty} n^{1/2}(\log\log n)^{-3/2}\xi = O_p(1),\\ &(ii)\ \max_{1<k<n} k^{1/2}(\log\log n)^{-3/2}\xi_{1,k} = O_p(1),\\ &(iii)\ \max_{1<k<n}(n-k)^{1/2}(\log\log n)^{-3/2}\xi_{n,k} = O_p(1).\end{aligned}$$

Proof.

(i) Under H_0, let $\delta^2 = Var(X_1^2)$. Since $E[X_1^4] < \infty$ and $E[X_1^2] = \sigma^2 < \infty$,

$$\delta^2 = Var(X_1^2) = E[X_i^4] - E[X_i^2]^2 = E[X_i^4] - \sigma^4 < \infty. \tag{2.4}$$

By the law of the iterated logarithm, we have

$$\limsup_{n\to\infty}\frac{\sum_{i=1}^n X_i^2 - n\sigma^2}{\delta\sqrt{2n\log\log n}} = 1 \quad a.s..$$

It is

$$\limsup_{n\to\infty}\frac{(\hat{\sigma}^2-\sigma^2)}{\delta\sqrt{2\log\log n/n}} = 1 \quad a.s..$$

Therefore

$$n^{1/2}(\log\log n)^{-1/2}|\hat{\sigma}^2-\sigma^2| = O_p(1). \tag{2.5}$$

Then

$$n^{3/2}(\log\log n)^{-3/2}|\hat{\sigma}^2-\sigma^2|^3=O_p(1).$$

Thus

$$n^{1/2}(\log\log n)^{-3/2}\frac{n}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma_\eta^2)^{\lambda-3}|(\hat{\sigma}^2-\sigma^2)^3|=O_p(1),$$

which is $n^{1/2}(\log\log n)^{-3/2}|\xi|=O_p(1)$. (i) is proved.

(ii) Under $H1$, let $\delta_1^2=Var(X_1^2)$. By (2.4), we have $\delta_1^2<+\infty$. Let $Z_i=(X_i^2-\sigma_i^2)/\sqrt{Var(X_i^2)}$, then we have $E[Z_i]=0$ and $E[Z_i^2]=1$. By Lemma 2.1, we have

$$\lim_{n\to\infty}(2\log\log n)^{-1/2}\max_{1<k<n}\frac{|\sum_{i=1}^{k}Z_i|}{\sqrt{k}}=1\quad a.s.. \tag{2.6}$$

We know

$$\hat{\sigma}_{1,k}^2-\sigma_1^2=\frac{\sum_{i=1}^{k}(X_i-\sigma_1^2)}{k}=\frac{\sum_{i=1}^{k}(X_i-\sigma_1^2)/\delta_1}{k/\delta_1}=\frac{\sum_{i=1}^{k}Z_i}{k/\delta_1},$$

i.e., $\sum_{i=1}^{k}Z_i=k(\hat{\sigma}_{1,k}^2-\sigma_1^2)/\delta_1$. Put it into (2.6), we obtain

$$\max_{1<k<n}k^{1/2}(\log\log n)^{-1/2}|\hat{\sigma}_{1,k}^2-\sigma_1^2|=O_p(1). \tag{2.7}$$

Therefore

$$\max_{1<k<n}k^{3/2}(\log\log n)^{-3/2}|\hat{\sigma}_{1,k}^2-\sigma_1^2|^3=O_p(1).$$

Thus

$$\max_{1<k<n}k^{1/2}(\log\log n)^{-3/2}\frac{k}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma_\eta^2)^{\lambda-3}|\hat{\sigma}_{1,k}^2-\sigma_1^2|^3=O_p(1),$$

which is $\max_{1<k<n}k^{1/2}(\log\log n)^{-3/2}|\xi_{1,k}|=O_p(1)$. (ii) is proved.

(iii) Similarly as proof of (ii), we obtain (iii).

Lemma 2.3 *Assume (2.1) holds, $\xi,\xi_{1,k},\xi_{n,k}$ are denoted as above, then we have*

$$(i)\ \lim_{n\to\infty}[a(\log n)]^2\xi-(b(\log n)+x)^2\xrightarrow{P}-\infty,$$
$$(ii)[a(\log n)]^2\max_{1<k<n}(\xi_{1,k}+\xi_{n,k})-(b(\log n)+x)^2\xrightarrow{P}-\infty.$$

where $a(\log n) = (2\log\log n)^{1/2}$, $b(\log n) = 2\log\log n + \frac{1}{2}\log\log\log n - \log\Gamma(\frac{1}{2})$, and x is for any real number, $x \in R$.

Proof.

(i)

As $n \to \infty$,

$$\frac{[a(\log n)]^2\xi}{[b(\log n)]^2} \leq \frac{(2\log\log n)n^{-1/2}(\log\log n)^{3/2}(n^{1/2}(\log\log n)^{-3/2}\xi)}{(2\log\log n)^2}$$
$$= \frac{(\log\log n)^{1/2}(n^{1/2}(\log\log n)^{-3/2}\xi)}{2n^{1/2}}.$$

By (2.5), we know $n^{1/2}(\log\log n)^{-3/2}|\xi| = O_p(1)$, then

$$\lim_{n\to\infty} \frac{[a(\log n)]^2\xi}{[b(\log n)]^2} \xrightarrow{P} 0.$$

Hence for any $x \in R$,

$$[a(\log n)]^2\xi - (b(\log n) + x)^2 \xrightarrow{P} -\infty$$

It is (i).

(ii) As $\log n \leq k < n$,

$$\max_{\log n\leq k<n} \frac{[a(\log n)]^2\xi_{1,k}}{[b(\log n)]^2} \leq \max_{\log n\leq k<n} \frac{(2\log\log n)k^{1/2}(\log\log n)^{-3/2}\xi_{1,k}}{(2\log\log n)^2k^{1/2}(\log\log n)^{-3/2}}$$
$$= \max_{\log n\leq k<n} \frac{(\log\log n)^{1/2}}{2k^{1/2}}k^{1/2}(\log\log n)^{-3/2}\xi_{1,k}$$
$$\leq \frac{(\log\log n)^{1/2}}{2(\log n)^{1/2}} \max_{\log n\leq k<n} k^{1/2}(\log\log n)^{-3/2}\xi_{1,k}.$$

By (2.7), $\max_{1<k<n} k^{1/2}(\log\log n)^{-3/2}|\xi_{1,k}| = O_p(1)$, we can obtain

$$\max_{\log n\leq k<n} \frac{[a(\log n)]^2\xi_{1,k}}{[b(\log n)]^2} \xrightarrow{P} 0. \tag{2.8}$$

As $1 < k < \log n$,

$$\max_{1<k<\log n} \frac{[a(\log n)]^2\xi_{1,k}}{[b(\log n)]^2} \leq \max_{1<k<\log n} \frac{(2\log\log n)k^{1/2}(\log\log\log n)^{-3/2}\xi_{1,k}}{(2\log\log n)^2k^{1/2}(\log\log\log n)^{-3/2}}$$
$$= \max_{1<k<\log n} \frac{(\log\log\log n)^{3/2}}{(2\log\log n)k^{1/2}}k^{1/2}(\log\log\log n)^{-3/2}\xi_{1,k}.$$

By (2.7), we can obtain

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

Also we have

$$\lim_{n\to\infty} \frac{(\log\log\log n)^{3/2}}{\log\log n} \to 0.$$

Hence

$$\max_{1<k<\log n} \frac{[a(\log n)]^2\xi_{1,k}}{[b(\log n)]^2} \xrightarrow{P} 0. \tag{2.9}$$

Combine (2.8) and (2.9), we obtain

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

Similarly we can obtain

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

Hence

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

(ii) is proved.

2.3 Asymptotic Properties

Theorem 2.1 *Let $X_i, i = 1, 2, \ldots, n$ be a sequence of I.I.D. random variables with mean 0 and variance σ_i^2. Assume (2.1) holds, and $\hat{\sigma}_{1,k}^2, \hat{\sigma}_{n,k}^2$ and $\hat{\sigma}^2$ are defined in (2.3). Then*

$$[\frac{1}{\hat{\sigma}^2}(\frac{k(n-k)}{2n^2})^{1/2}](\frac{k(n-k)}{n})^{1/2}(\hat{\sigma}_{1,k}^2 - \hat{\sigma}_{n,k}^2) \xrightarrow{D} B_0,$$

where B_0 is a Brownian bridge.

Proof.

$$\begin{aligned}
&[\frac{1}{\hat{\sigma}^2}(\frac{k(n-k)}{2n^2})^{1/2}](\frac{k(n-k)}{n})^{1/2}(\hat{\sigma}_{1,k}^2 - \hat{\sigma}_{n,k}^2) \\
=&\frac{\sqrt{n/2}}{n^2\hat{\sigma}^2}k(n-k)(\hat{\sigma}_{1,k}^2 - \hat{\sigma}_{n,k}^2) = \frac{\sqrt{n/2}}{n\sum_{i=1}^n x_i^2}((n-k)\sum_{i=1}^k x_i^2 - k\sum_{i=k+1}^n x_i^2) \\
=&\frac{\sqrt{n/2}}{n\sum_{i=1}^n x_i^2}(n\sum_{i=1}^k x_i^2 - k\sum_{i=1}^n x_i^2) = \sqrt{n/2}(\frac{\sum_{i=1}^k x_i^2}{\sum_{i=1}^n x_i^2} - \frac{k}{n})
\end{aligned}$$

Let $B = \sqrt{n/2}(\frac{\sum_{i=1}^{k} x_i^2}{\sum_{i=1}^{n} x_i^2} - \frac{k}{n})$, then B is exactly CUSUM introduced by [Inclan and Tiao, 1994] and they already proved that $B \xrightarrow{D} B_0$ in their paper.

Theorem 2.1 is proved.

Theorem 2.2 *Let $X_i, i = 1, 2, \ldots, n$ be a sequence of I.I.D. random variables with mean 0 and variance σ_i^2. Assume (2.1) holds. When $\lambda = 2$, we have*

$$\lim_{n\to\infty} P\{a(\log n) \cdot \frac{\max(v_{2,k}^{1/2})}{2^{1/2}\hat{\sigma}^2} - b(\log n) \le x\} = \exp(-2e^{-x}), \tag{2.10}$$

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

Proof.

When $\lambda = 2$, (2.2) goes to

$$v_{2,k} = k(\hat{\sigma}_{1,k}^2)^2 + (n-k)(\hat{\sigma}_{n,k}^2)^2 - n(\hat{\sigma}^2)^2.$$

Denote $\hat{\sigma}_{1,k}^2 = \frac{A}{k}$ and $\hat{\sigma}_{n,k}^2 = \frac{B}{n-k}$. We can easily obtain

$$\begin{aligned} v_{2,k} &= k(\frac{A}{k})^2 + (n-k)(\frac{B}{n-k})^2 - n(\frac{A+B}{n})^2 = \frac{A^2}{k} + \frac{B^2}{n-k} - \frac{(A+B)^2}{n} \\ &= \frac{(n-k)nA^2 + knB^2 - k(n-k)(A^2+2AB+B^2)}{kn(n-k)} \\ &= \frac{(nA)^2 + (kA)^2 + (kB)^2 - 2knA^2 - 2knAB + 2k^2AB}{kn(n-k)} \\ &= \frac{(nA-kA-kB)^2}{kn(n-k)} = \frac{((n-k)A-kB)^2}{kn(n-k)} = \frac{k(n-k)}{n}(\frac{A}{k} - \frac{B}{n-k})^2 \\ &= \frac{k(n-k)}{n}(\hat{\sigma}_{1,k}^2 - \hat{\sigma}_{n,k}^2)^2. \end{aligned}$$

Therefore, we have

$$v_{2,k}^{1/2} = (\frac{k(n-k)}{n})^{1/2} \left|\hat{\sigma}_{1,k}^2 - \hat{\sigma}_{n,k}^2\right|.$$

Let $B_n = [\frac{1}{\hat{\sigma}^2}(\frac{k(n-k)}{2n^2})^{1/2}](\frac{k(n-k)}{n})^{1/2}(\hat{\sigma}_{1,k}^2 - \hat{\sigma}_{n,k}^2)$, then $B_n \xrightarrow{D} B_0$ by Theorem 2.1, where B_0 is a Brownian bridge. Let $\delta = \sqrt{Var(X_i^2)} < +\infty$ under H_0, and define $U(0) = U(1) = U(n-1) = U(n) = 0$ and

$$U(k) = \frac{k(n-k)}{2\delta n^{3/2}}(\hat{\sigma}_{1,k} - \hat{\sigma}_{2,k}),\ 1 < k < n.$$

By Theorem 2.4.7 and Eg. 2.4.3 of [Csörgo and Horváth, 1997], we have

$$\max_{1<k<n} |U(k)| \xrightarrow{D} \sup_{0<t<1} |B(t)|,$$

where $B(t)$ is a Brownian bridge. Let $0 < t = k/n < 1$ and we define $U^*(t)$ as

$$U^*(t) = \frac{n^{1/2}t(1-t)}{2\delta}(\hat{\sigma}_{1,k} - \hat{\sigma}_{2,k}) = U(k).$$

By Theorem 2.4.9 of [Csörgo and Horváth, 1997], we have

$$\lim_{n\to\infty} P\{a(\log n) \cdot \sup_{0<t<1} |U^*(t)|/(t(1-t))^{1/2} - b(\log n) \le x\} = \exp(-2e^{-x}).$$

for all x, where $a(\log n) = (2\log\log n)^{1/2}$ and $b(\log n) = 2\log\log n + \frac{1}{2}\log\log\log n - \log\Gamma(\frac{1}{2})$.

After comparing $U(k)$ and B_n, we find the only difference is δ in $U(k)$ and $\hat{\sigma}^2$ in B_n. However neither δ nor $\hat{\sigma}^2$ can affect the distribution of $U(k)$ or B_n, since both of them converge to a Brownian bridge in distribution. Therefore, we have

$$\lim_{n\to\infty} P\{a(\log n) \cdot \sup_{0<t<1} |B_n|/(t(1-t))^{1/2} - b(\log n) \le x\} = \exp(-2e^{-x}).$$

After substituting t by k/n, the probability limit above goes to (2.10).

Theorem 2.2 is proved.

Theorem 2.3 *Let $X_i, i = 1, 2, \ldots, n$ be a sequence of I.I.D. random variables with mean 0 and variance σ_i^2. Assume (2.1) holds. Under $H_0, \sigma_1^2 = \sigma_n^2 = \sigma^2$, as $\lambda > 1$ or $\lambda < 0$, we have*

$$\lim_{n\to\infty} P\{a^2(\log n) \cdot \frac{\max(v_{\lambda,k})}{2\hat{\sigma}^4} \le \frac{1}{2}\lambda(\lambda-1)(\hat{\sigma_n}^2)^{\lambda-2}(b(\log n) + x)^2\} = \exp(-2e^{-x}). \quad (2.11)$$

Proof.

After applying a three-term Taylor expansion for $(\hat{\sigma}_{1,k}^2)^\lambda$,$(\hat{\sigma}_{n,k}^2)^\lambda$, and $(\hat{\sigma}^2)^\lambda$ in $v_{\lambda,k}$, we obtain

$$\begin{aligned} v_{\lambda,k} &= k(\hat{\sigma}_{1,k}^2)^\lambda + (n-k)(\hat{\sigma}_{n,k}^2)^\lambda - n(\hat{\sigma}^2)^\lambda \\ &= \frac{1}{2}\lambda(\lambda-1)(\sigma^2)^{\lambda-2}[k(\hat{\sigma}_{1,k}^2)^2 + (n-k)(\hat{\sigma}_{n,k}^2)^2 - n(\hat{\sigma}^2)^2] + (\xi_{1,k} + \xi_{n,k} - \xi) \\ &= \frac{1}{2}\lambda(\lambda-1)(\sigma^2)^{\lambda-2}v_{2,k} + (\xi_{1,k} + \xi_{n,k} - \xi). \end{aligned}$$

where

$$\begin{aligned}\xi_{1,k} =& \frac{k}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta_{1,k}})^{\lambda-3}(\hat{\sigma}^2_{1,k}-\sigma^2_1)^3,\\ \xi_{n,k} =& \frac{n-k}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta_{n,k}})^{\lambda-3}(\hat{\sigma}^2_{n,k}-\sigma^2_n)^3,\\ \xi \ =& \frac{n}{3!}\lambda(\lambda-1)(\lambda-2)(\sigma^2_{\eta})^{\lambda-3}(\hat{\sigma}^2-\sigma^2)^3.\end{aligned}$$

Therefore

$$\begin{aligned}\max\{v_{\lambda,k}\} &= \max\{k(\hat{\sigma}^2_{1,k})^\lambda + (n-k)(\hat{\sigma}^2_{n,k})^\lambda - n(\hat{\sigma}^2)^\lambda\}\\ &= \max\{\frac{1}{2}\lambda(\lambda-1)(\sigma^2)^{\lambda-2}v_{2,k} + (\xi_{1,k}+\xi_{n,k}-\xi)\}. \qquad (2.12)\end{aligned}$$

By Lemma 2.3, we know $[a(\log n)]^2 \max_{1<k<n}(\xi_{1,k}+\xi_{n,k}-\xi) - (b(\log n)+x)^2 \xrightarrow{P} -\infty$, then combine (2.12) and (2.10) in Theorem 2.2, we obtain:

$$\lim_{n\to\infty} P\{a^2(\log n)\cdot\frac{\max(v_{\lambda,k})}{2\hat{\sigma}^4} \le \frac{1}{2}\lambda(\lambda-1)(\hat{\sigma}^2)^{\lambda-2}(b(\log n)+x)^2\} = \exp(-2e^{-x}).$$

Theorem 2.3 is proved.

The significance level α can be applied here based on (2.11), i.e.:

$$1-\alpha = \lim_{n\to\infty} P\{a^2(\log n)\cdot\frac{\max(v_{\lambda,k})}{2\hat{\sigma}^4} \le \frac{1}{2}\lambda(\lambda-1)(\hat{\sigma}^2)^{\lambda-2}(b(\log n)+x)^2\} = \exp(-2e^{-x}). \quad (2.13)$$

The asymptotic critical value $C_{\lambda,\alpha}$ can be derived from (2.13):

$$C_{\lambda,\alpha} = \hat{\sigma}^{2\lambda}\lambda(\lambda-1)\frac{(2\log\log n + \frac{1}{2}\log\log\log n - \log\Gamma(\frac{1}{2}) - \log\log(1-\alpha)^{-1/2})^2}{2\log\log n} \qquad (2.14)$$

which satisfies

$$P(\max\{v_{\lambda,k}\} > C_{\lambda,\alpha}) = \alpha.$$

As $0<\lambda<1$, by the same way above, we can obtain

$$\min\{v_{\lambda,k}\} = \min\{\frac{1}{2}\lambda(\lambda-1)(\sigma^2)^{\lambda-2}(v_{2,k}) + \xi_{1,k} + \xi - \xi\} < 0$$

and

$$P(\min\{v_{\lambda,k}\} < C_{\lambda,\alpha}) = \alpha.$$

where $C_{\lambda,\alpha}$ is calculated as same as (2.14), but note $C_{\lambda,\alpha} < 0$ here.

2.4 Modified Weighted Sum of Powers of Variances (MWSPV)

The simulation study in Section 2.5.1 shows if sample size is small, the results are not good enough since the critical value is based on the large sample size. It also shows when the sample size is small, absolute value of the critical value is slightly larger than it should be, and when sample size is large, the critical value works as well as it should be. It tells us that an extra term may add here to improve WSPV's performance. The extra term should satisfy when sample size is small, absolute value of WSPV is a little bit larger, and when sample size is large, it won't affect value of WSPV.

According to Lemma 2.3, the impact of term $(\xi + \xi_{1,k} - \xi_{n,k})$ on the critical value can be ignored as $n \to \infty$. However, if the sample size is small, it may have a big impact on critical value. Recall Lemma 2.2, $n^{1/2}(\log\log n)^{-3/2}\xi = O_p(1)$. We see that the extra term ξ could be approximately proportional to $(\log\log n)^{3/2}n^{-1/2}$. Based on this idea, we set up the modified weighted sum of powers of variances MWSPV $v'_{\lambda,k}$:

$$v'_{\lambda,k} = k(\hat{\sigma}^2_{1,k})^\lambda + (n-k)(\hat{\sigma}^2_{n,k})^\lambda - n(\hat{\sigma}^2)^\lambda + \mathrm{sgn}(\lambda)\gamma(\log\log n)^{3/2}n^{-1/2}(\hat{\sigma}^2)^\lambda, \tag{2.15}$$

where $\mathrm{sign}(\lambda) \equiv \mathbf{I}(\lambda > 1 \text{ or } \lambda < 0) - \mathbf{I}(0 < \lambda < 1)$ and γ is a parameter to adjust the value of MWSPV.

Remark

- After comparing $v'_{\lambda,k}$ in (2.15) with $v_{\lambda,k}$ in (2.2), we can easily see that the MWSPV finds exactly the same change-point position as the WSPV does regardless of significance level since the penalty term is not related to k.
- When sample size n is smaller or say less than 200, the penalty term in the MWSPV has some big impact on $v'_{\lambda,k}$. Therefore, the MWSPV can catch more change-points than the WSPV does in case of small sample size.
- When sample size n is larger, the MWSPV and the WSPV should have almost the same performance since the penalty term goes to 0 when n goes to infinity. Simulation results show when n is larger than 200, the MWSPV and the WSPV have almost same performance.

The empirical value of λ in WSPV can be found by simulations. The hyperparameter γ adjusts the penalty term in MWSPV which becomes WSPV when $\gamma = 0$. A proper data-adaptive specification of the γ value is therefore very important. Our empirical study suggests this can be effectively achieved by setting $\gamma = q_\nu$ where q_ν is the level ν sample percentile of certain standardized data sequence deemed not containing any change-point. Such standardized sequence can be constructed in the following way: first, execute WSPV, i.e. MWSPV is run with $\gamma = 0$. Second, each segment in the partition is standardized by subtracting its sample mean and being divided by its sample standard deviation, i.e. the z-scores of each segment are calculated. The absolute values of the resultant z-scores give the referred standardized sequence. Common values of level ν are 0.9, 0.95 and 0.99.

2.5 Simulation Study and Real Data Analysis

- In order to see how accurate our method works, we measure the distance (D) between the detected change-point and the real change-point. We think $D < 10$ is successful detections.
- Numbers in tables of simulation parts are percentage of the successful detections based on 10,000 simulations.
- The optimal λ are developed by the simulation study in Chapter 2.5.1. The standardized sample percentile $\nu = 0.95$ and $\nu = 0.99$ are tested in MWSPV, respectively.

2.5.1 WSPV Simulation

A sequence of I.I.D. normal variables with mean 0 and variance $\sigma_i^2, i = 1, \ldots, n$ are set here. Different sample size n, 50, 100 and 200 are tested.

For no change-points (No CP) cases, samples are I.I.D. normal variables with mean 0 and variance 1. For one change-point (One CP) test, the position (τ) of the change-point is set to 0.5. Variance before and after change-point are 1 and 4. Table 2 shows percentage of successfully detecting.

In case of No CP, $\lambda = 0.1$ and $\lambda = 0.5$ are the best, while in case of One CP, $\lambda = -0.1$

and $\lambda = 0.1$ are the best. Obviously, $\lambda = 0.1$ is the best choice we can make.

Table 2: Change-points Test for WSPV with Different λ

CP	Size	λ									
		-2	-1	-0.5	-0.1	0.1	0.5	0.9	1.1	2	3
	50	62.0	80.9	91.9	98.7	**99.6**	99.7	99.5	99.3	96.7	91.5
No CP	100	61.1	81.1	91.6	98.4	**99.5**	99.6	99.2	99.0	95.7	89.6
	200	60.8	81.0	91.6	98.2	**99.3**	99.3	99.0	98.7	95.2	88.2
	50	39.6	46.2	47.7	45.5	**42.4**	35.4	29.4	27.1	20.5	20.5
One CP	100	48.5	69.7	79.4	82.5	**82.2**	80.0	75.8	73.3	61.3	48.7
$\tau = 0.5$	200	57.0	81.7	89.3	90.9	**90.6**	89.6	87.2	85.4	73.8	58.3

Note: unit in table is percentage(%)

2.5.2 MWSPV Simulation

We choose $\lambda = 0.1$ for both WSPV and MWSPV. The standardized sample percentile $\nu = 0.95$ and $\nu = 0.99$ are tested in MWSPV, respectively. We estimate γ by the method introduced in Chapter 2.4.

We compare WSPV and MWSPV with other methods, like CUSUM by Inclan and Tian [Inclan and Tiao, 1994] and BIC-type method by Chen and Gupta [Chen and Gupta, 2012], both of which are briefly described in [A-1], at the same significance level 0.05. Inclan and Tian [Inclan and Tiao, 1994] calculated the empirically critical values in their CUSUM paper and we use such empirical values at significance level 0.05 to do simulations. All results obtained below are based on 10,000 simulations.

Table 3 shows percentage of successful detections for No CP cases. 95% standardized sample percentile value makes the Type I error under 5 % no matter how big the sample size is. Type I error for the 99% standardized sample percentile value depends on the sample size. Type I error around 5% when sample size is as large as 200, and when sample size is as small as 50, the type I error is controlled under 10%.

Table 4 shows test results based on different sample size and variance change. The position (τ) of the change-point is set to be 0.3, 0.5 or 0.7. Variance change 1 vs 3 and 1 vs 4 are used

Table 3: No Change-point Test for MWSPV

Size	MWSPV		BIC	CUSUM
	$\nu = 0.95$	$\nu = 0.99$		
50	95.8	**90.5**	99.3	97.6
100	96.6	**92.7**	99.1	94.7
200	96.9	**94.3**	98.9	95.0

Note: unit in table is percentage(%)

to test method's applicability.

Table 4: Power Test under $\alpha = 0.05$ for One Change-point

Var Change		$\sigma_1 : \sigma_n = 1 : \sqrt{3}$				$\sigma_1 : \sigma_n = 1 : 2$			
		MWSPV				MWSPV			
Size	τ	$\nu = 0.95$	$\nu = 0.99$	BIC	CUSUM	$\nu = 0.95$	$\nu = 0.99$	BIC	CUSUM
50	0.3	29.9	**40.3**	11.5	7.2	50.1	**60.7**	25.3	13.3
	0.5	43.4	**51.4**	21.7	47.0	67.1	**73.9**	44.2	69.9
	0.7	42.2	**50.1**	21.8	52.4	65.3	**72.0**	43.4	75.0
100	0.3	57.6	**63.9**	39.2	39.7	81.5	**84.4**	70.2	57.5
	0.5	69.2	**72.8**	56.4	76.4	87.0	**88.0**	82.5	85.5
	0.7	67.2	**71.3**	53.1	81.9	86.8	**87.9**	80.8	93.8
200	0.3	78.8	**79.5**	75.6	55.3	90.1	**90.1**	90.1	60.9
	0.5	81.0	**81.2**	80.3	78.0	90.6	**90.6**	90.9	83.5
	0.7	81.0	**81.4**	79.3	86.8	91.2	**91.2**	91.3	93.8

Note: unit in table is percentage(%)

After thousands of simulations, we can draw following conclusions:

- When sample size is large, WSPV is as good as MWSPV.
- When sample size is smaller than 200, MWSPV is much better than WSPV, especially, when both changes in variance and sample size are small.
- When sample size is small, MWSPV works better than BIC method. When sample size is as large as or larger than 200, MWSPV works as well as BIC does.

- MWSPV works consistently no matter where the change-point is, however detection of CUSUM highly depends on the location of change-point. When the former variance is less than the latter variance and location of change-point is within the first half, MWSPV works much better than CUSUM, and when change-point is around the middle, MWSPV works slightly better than CUSUM, and when change-point is within the second half, CUSUM is slightly better than MWSPV.

2.5.3 Multiple Change-Points Detection

In the real world, we often face multiple change-points cases. To extend MWSPV to multiple change-points case, we apply a binary segmentation procedure:

- Step 1: Find the most significant change-point in data by MWSPV.
- Step 2: Split the data into 2 parts from the change-point found at Step 1.
- Step 3: Reapply MWSPV to 2 subsamples to check the change-point in each segment.
- Step 4: Repeat Step 1-3 until no change-point is found in any segment.

2.5.4 Real Data Analysis: IBM Stock Prices

In 1976, [Box and Jenkins, 1976] presented a time series data, IBM common stock daily closing prices from 17th May, 1961 to 2nd November, 1962. We transform it by the first order difference of logarithm and use MWSPV to analyze it with $\lambda = 0.1$.

The Figure 2 shows the first order difference of logarithm IBM stock closing prices. It looks like there is a change-point around 230 while others are not very clear. By MWSPV, we easily find the first one at point 235. Once we find the first one, we split the data into two parts from the first change-point 235 and then reapply MWSPV to that two parts. Between points 235 and 368, we find point 279. Between points 1 and 235, we find something interesting. We find the point 8 by setting sample percentile level $\nu = 0.99$ while no change-point is found by setting $\nu = 0.95$. Table 5 shows difference in HV between different parts. HV before point 8 is 1.8 times higher than HV after point 8, i.e. around 3.3 times higher in variance. Also we

know by simulation when sample size is around 200, Type I error of MWSPV is around 5%. Therefore we think point 8 is a change-point. After that, we continue to check if there is any change-point between existing change-points, but we cannot find any further change-points. Finally we obtain three change-points at 8, 235 and 279 in the data.

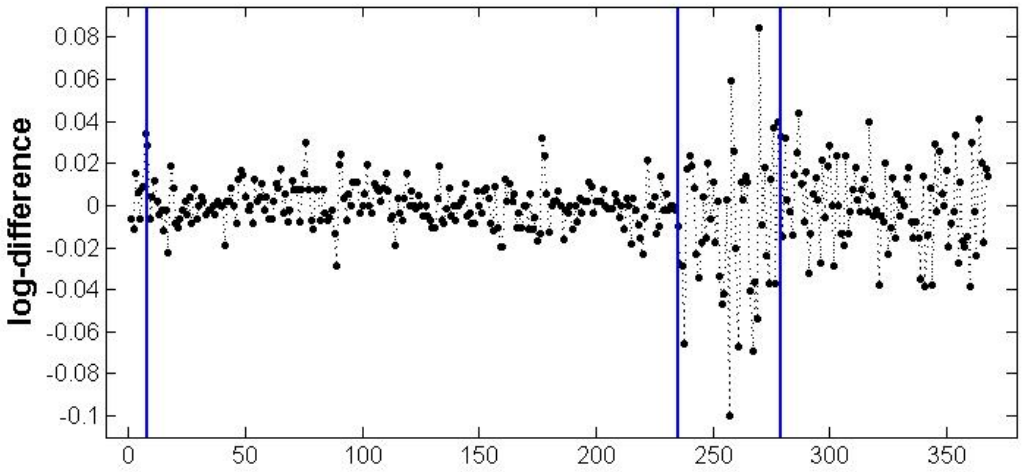

Figure 2: Historical Volatility Change Detection of IBM Stock Prices by MWSPV Algorithm

Table 5: Historical Volatility Change in Different Time Range

Position	1-8	9-235	236-279	280-368
HV	1.67	0.92	3.65	1.94

Note: unit in table is 10^{-2}

This famous series data have been analyzed by some researchers. [Wichern *et al.*, 1976] used ARIMA (1,1,0) to find two change-points, 180 and 235. [Tsay, 1988] used ARIMA(0,1,1) to detect the change-point at 237. CUSUM finds change-points at 235 and 279 while BIC-type method find change-points at 235 and 281.

2.6 Extensions: Generalized Weighted Variance

Since WSPV is essentially a function of $\hat{\sigma}^2_{1,k}$ and $\hat{\sigma}^2_{n,k}$, we are very interested if we can extend this concept to general case.

We redefine $\hat{k}$ as

$$\hat{k} = \begin{cases} \arg\max\limits_{1<k<n} \{v_k\} & \text{if } g(\cdot) \text{ is convex,} \\ \arg\min\limits_{1<k<n} \{v_k\} & \text{if } g(\cdot) \text{ is concave.} \end{cases}$$

Also we redefine v_k as

$$v_k = k \cdot g(\hat{\sigma}^2_{1,k}) + (n-k) \cdot g(\hat{\sigma}^2_{n,k}) - n \cdot g(\hat{\sigma}^2),$$

where $\hat{\sigma}^2_{1,k}$ and $\hat{\sigma}^2_{n,k}$ are defined as before, $g(\cdot)$ is either convex or concave function globally.

Finally we redefine $v_{\hat{k}}$ as

$$v_{\hat{k}} = \begin{cases} \max\limits_{1<k<n} (v_k) & \text{if } g(\cdot) \text{ is convex} \\ \min\limits_{1<k<n} (v_k) & \text{if } g(\cdot) \text{ is concave.} \end{cases}$$

Let $g(\cdot) = \log(\cdot)$. Since logarithm function is concave function, $\hat{k}$ can be calculated by

$$\hat{k} = \arg\min_{1<k<n} \{v_k\}.$$

where

$$v_k = k \cdot \log(\hat{\sigma}^2_{1,k}) + (n-k) \cdot \log(\hat{\sigma}^2_{n,k}) - n \cdot \log(\hat{\sigma}^2). \tag{2.16}$$

Denote $v_1 = (-v_{\hat{k}})^{1/2}$, then v_1 is likelihood procedure approach. [Chen and Gupta, 2012] showed that

$$\lim_{n\to\infty} P[a(\log n)v_1 - b(\log n) \le x] = \exp\{-2e^{-x}\}$$

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

[Schwarz, 1978] introduced Schwarz Information Criterion(SIC or BIC) for model selection. Based on the principle of BIC, we can define

$$v_2 = v_{\hat{k}} + \log n$$

where $v_{\hat{k}}$ is defined by (2.16). [Chen and Gupta, 2012] introduced it as BIC-typed statistic.

In terms of principle of BIC, we accept H_0 if $v_2 < 0$. Furthermore, [Chen and Gupta, 2012] set up a critical value $c_\alpha(\ge 0)$. H_0 is accepted when $v_2 < c_\alpha$ rather than $v_2 < 0$. They also estimated the critical value c_α based on both significance level α and sample size.

Conclusion

The WSPV has very good performance to find change-point in variance. The MWSPV highly improves the performance of the WSPV and has better performance than BIC and CUSUM in the most of cases when sample size is small (less than 200). When sample size is larger, MWSPV is comparable to other single change-point detection methods.

3 Multiple Historical Volatility Change Detection by Empirical Bayesian Information Criteria

In Chapter 2, we introduce the method which just can find single change-point in HV. However, allowing for occurrence of multiple change-points in HV is a more realistic approach. A binary segmentation procedure has two major problems: Once it does not reject the hypothesis of no change-point in a segment it will be impossible to detect change-points in any subsequent sub-segment; also once a data point is wrongly detected as a change-point at certain stage there is no chance for the error to be revoked. Instead of using the binary segmentation procedure, in this chapter, we propose an emBIC method to find multiple change-points simultaneously. Furthermore, this method can not only detect changes in HV, but also other changes like mean changes or mean and variance changes.

Many developments of Markov chain Monte Carlo (MCMC) since 1990's have enabled Bayesian methods to be effective for analysis of multiple change-points. Under a Bayesian setup, number and locations of change-point are estimated simultaneously by the posterior distributions via MCMC computing. [Barry and Hartigan, 1993] used a product partition prior distribution to construct a Bayesian change-points model, and used Gibbs sampler, which they called Markov sampling, to estimate the posterior distribution of change-points locations and sizes. Instead of using the product partition prior distribution one can use a multivariate discrete prior defined on the change-points indicator process $V_1, \cdots, V_n$ corresponding to the data $Y_1, \cdots, Y_n$, where $V_i = 1$ if i is a change-point location and $V_i = 0$ otherwise. Independent Bernoulli prior distributions were used for the V_i's in [Lavielle and Lebarbier, 2001] for instance, while a truncated Poisson on $\sum_{i=1}^{n} V_i$ was used in [Kim and Cheon, 2010]. In addition to multiple change-points, other parameters and/or hyper-parameters in a Bayesian model also need to be estimated. Due to the parameterization complication in many Bayesian models, estimating these parameters often entails dedicated Monte Carlo methods, such as hybrid MCMC, reversible jump MCMC, stochastic approximation expectation-maximization (SAEM) ([Delyon *et al.*, 1999]), stochastic approximation Monte Carlo (SAMC) ([Liang *et al.* (2007)]), and annealing stochastic approximation Monte Carlo (ASAMC) ([Liang, 2007]) etc.. See [Kim and Cheon, 2010] for more discussions on this. Computations involved in these Monte Carlo methods are usually very intensive. The parameterization complication needs to

be carefully treated as it can cause further difficulties on interpreting the simulation results of some posterior distributions ([Lavielle and Lebarbier, 2001]).

Motivated by the aforementioned research we propose a new statistical procedure to simultaneously estimate the change-points configuration and the associated parameters. In the procedure, we use empirical Bayesian and maximum likelihood principles to first derive a model selection or change-points selection criterion which we call emBIC (standing for empirical Bayesian information criterion). In developing the emBIC we choose a special product partition distribution as the non-informative prior for the change-points indicator process, and use the maximum likelihood estimates for the parameters in the model. The resultant emBIC tends to achieve the minimum when the model correctly specifies all change-points. The optimal change-points configuration is thus naturally defined as the one minimizing the emBIC. We estimate the optimal change-points configuration by inducing a multivariate discrete distribution on the change-points indicator process and generating Markov chains from the induced distribution using the Gibbs sampler. The optimal change-points configuration is the mode of the induced distribution by definition. Therefore, the optimal change-points configuration has the highest probability to appear in a generated Markov chain; and accordingly can be accurately and efficiently estimated by a stochastic optimization method. Our proposed procedure combines the advantages of both Bayesian and model selection approaches for multiple change-points, yet the MCMC computing it requires is relatively simple and less intensive. The procedure is expected to work in general multiple change-points problems with different statistical models involved. For the sake of clear and concrete presentation, we will use only examples of mean and variance changes to illustrate.

This chapter is organized as follows. In section 3.1 we derive a general form of emBIC for multiple change-points estimation. We provide specific forms of emBIC in cases of mean and variance change-points, and discuss their connections with related existent methods. In section 3.2 we show how a Gibbs sampler engineered stochastic sampling and search method is used to operate the change-points estimation based on emBIC and other related methods in a computationally feasible way. We also derive some asymptotic properties on convergence and efficiency of the stochastic sampling and search method. In section 3.3 we address various computing and diagnosis issues associated with the use of the proposed and other methods in

practice. We provide a simulation study and two applications involving finance and genetics real data in section 3.4. The chapter ends with concluding remarks in section 3.5.

3.1 An Empirical Bayesian Information Criterion

3.1.1 General Case of Multiple Change-points

Consider a sequence of independent random variables $Y_1, \cdots, Y_n$ with parameters $\theta_1, \cdots, \theta_n$. The parameter θ_i can be multi-dimensional but its dimension $\dim(\theta_i) \equiv d$ is assumed fixed. Let

$$\theta_1 = \cdots = \theta_{j_1} = \phi_1 \neq \theta_{j_1+1} = \cdots = \theta_{j_2} = \phi_2 \neq \cdots \neq \theta_{j_K+1} = \cdots = \theta_n = \phi_{K+1}.$$

where $\boldsymbol{\phi}_K = (\phi_1, \cdots, \phi_{K+1})$ specifies the distinct values of $\boldsymbol{\theta}_n = (\theta_1, \cdots, \theta_n)$. Both $\boldsymbol{\phi}_K$ and the change-points configuration $(j_1, \cdots, j_K; K)$ are unknown and need to be estimated, but both have different conceptual meanings: $\boldsymbol{\phi}_K$ is treated as an unknown vector parameter with unknown length while $(j_1, \cdots, j_K; K)$ is regarded as a latent random vector or latent data.

Denote $\mathbf{Y}_{ns} = (Y_{j_{(s-1)}+1}, \cdots, Y_{j_s})$ and $\mathbf{y}_{ns} = (y_{j_{(s-1)}+1}, \cdots, y_{j_s})$ its realization, where $s = 1, \cdots, K+1$. According to the change-points configuration $(j_1, \cdots, j_K; K)$ the data sequence $\mathbf{Y}_n = (Y_1, \cdots, Y_n)$ can be partitioned into $K+1$ segments $\mathbf{Y}_{n1}, \cdots, \mathbf{Y}_{n(K+1)}$. Let $f_{ns}(y_{ns}|\phi_s)$ be the joint pdf of Y_{ns} which involves the sth parameter ϕ_s. The joint pdf of $Y_1, \cdots, Y_n$ given the change-points configuration is then

$$f_n(y_n|\boldsymbol{\theta}_n; j_1, \cdots, j_K; K) = \prod_{s=1}^{K+1} f_{ns}(\mathbf{y}_{ns}|\phi_s), \quad \text{where } \mathbf{y}_n = (y_1, \cdots, y_n). \tag{3.1}$$

Note that in obtaining (3.1) there is no need of mutual independence in $Y_1, \cdots, Y_n$. Rather, only the mutual independence among the $K+1$ segments is assumed.

The latent data $(j_1, \cdots, j_K; K)$ can be equivalently represented by an indicator random process $\mathbf{V}_n = (V_1, \cdots, V_n)$ for simplicity of presentation, where $V_i = 1$ if $i = j_s$ for some $s \in \{1, \cdots, K\}$ and $V_i = 0$ otherwise; $i = 1, \cdots, n$. Note $V_n \equiv 0$ and $\sum_{i=1}^{n-1} V_i = K$ according to our notations on change-points. Both $(j_1, \cdots, j_K; K)$ and $(V_1, \cdots, V_n)$ determine the same partition of $Y_1, \cdots, Y_n$ giving $Y_{n1}, \cdots, Y_{n(K+1)}$. Thus we can use the idea of product partition

probability distribution in [Barry and Hartigan, 1993] to model the latent data. Namely

$$\Pr(V_1, \cdots, V_n) = \Pr(j_1, \cdots, j_K; K) = \Pr(j_1, \cdots, j_K | K) \cdot \Pr(K).$$

Note that when $\dim(\theta_i) \equiv d = 1$ the number of possible configurations of $(j_1, \cdots, j_K; K)$ given K is $\binom{n-1}{K} = \frac{(n-1)!}{K!(n-1-K)!}$; and conditional on K each such configuration may be equally likely *a priori*. Assigning a prior distribution to K is not so obvious. We initially used a hierarchical Bayesian approach where a binomial$(n-1, \xi)$ distribution is assumed for K given ξ, and ξ is assumed a beta(a, b) distribution. By this approach it can be shown that

$$\Pr(K) = \frac{(n-1)^{(n-1)}}{K^{(K)}(n-1-K)^{(n-1-K)}} \cdot \frac{(K+a-1)^{(K)}(n-1-K+b-1)^{(n-1-K)}}{(n-1+a+b-1)^{(n-1)}}$$

where $x^{(m)} = x(x-1)\cdots(x-m+1)$ is an order m factorial power function with $m \geq 1$ being a natural number. The hyper-parameters a and b need to be specified for this $\Pr(K)$ to be usable, which is difficult in practice. However it implies $\Pr(K) = \frac{1}{n}$ when $a = b = 1$ and $\Pr(K)$ is related to $\binom{n-1}{K}$ in certain manner in general.

This discussion suggests a simplified approach which introduces a tuning parameter γ and assumes

$$\Pr(j_1, \cdots, j_K | K) \propto \binom{n-1}{K}^{-1} \quad \text{and} \quad \Pr(K) \propto \binom{n-1}{K}^{1-\gamma}, \ \ \gamma \geq 0 \tag{3.2}$$

where $\Pr(K)$ is a constant when $\gamma = 1$; positively related to $\binom{n-1}{K}$ when $0 < \gamma < 1$; and negatively related to $\binom{n-1}{K}$ when $\gamma > 1$. A $\gamma \in [0, 2\log n]$ seems sufficient in practice (Section 3.3 for more detail). Now only γ needs to be specified in order to use (3.2).

It is not difficult to see that (3.2) can be extended by replacing 1 and γ there with d and $d\gamma$ respectively when $d > 1$. It follows that

$$\Pr(V_1, \cdots, V_n) = \Pr(j_1, \cdots, j_K; K) \propto \left[\sum_{k=0}^{n-1} \binom{n-1}{k}^{d-d\gamma}\right]^{-1} \binom{n-1}{K}^{-d\gamma}. \tag{3.3}$$

From (3.1) and (3.3) the joint probability density function of $(Y_1, \cdots, Y_n; j_1, \cdots, j_K; K)$ is

$$f_n(y_n; j_1, \cdots, j_K; K | \boldsymbol{\theta}_n) \propto \prod_{s=1}^{K+1} f_{ns}(y_{ns} | \phi_s) \cdot \left[\sum_{k=0}^{n-1} \binom{n-1}{k}^{d-d\gamma}\right]^{-1} \binom{n-1}{K}^{-d\gamma}, \tag{3.4}$$

which is also the complete data likelihood function of $\boldsymbol{\phi}_K = (\phi_1, \cdots, \phi_{K+1})$.

[Schwarz, 1978] derived SIC by using a conditional uniform prior distribution on the parameter space and a Laplace approximation for the likelihood function. [Yao, 1988] applied SIC to multiple change-points problem which has resulted in the following criterion function called BIC using our notations

$$\mathrm{BIC}(\mathbf{V}_n) \equiv \mathrm{BIC}(j_1, \cdots, j_K; K) = -\sum_{s=1}^{K+1} \log f_{ns}(y_{ns}|\hat{\phi}_{ns}) + \frac{d}{2}(K+1)\log n \tag{3.5}$$

where $\hat{\phi}_{ns}$ is the maximum likelihood estimator of ϕ_s. The BIC estimator of the change-points is defined as $\breve{\mathbf{V}}_n \equiv (\breve{j}_1, \cdots, \breve{j}_{\breve{K}}; \breve{K}_n) = \operatorname{argmin} \mathrm{BIC}(\mathbf{V}_n) \equiv \operatorname{argmin} \mathrm{BIC}(j_1, \cdots, j_K; K)$.

Applying the spirit of BIC aforementioned to the complete data likelihood function (3.4), we obtain the following empirical Bayesian information criterion (emBIC) function

$$\mathrm{emBIC}_\gamma(\mathbf{V}_n) \equiv \mathrm{emBIC}_\gamma(j_1, \cdots, j_K; K) = -\sum_{s=1}^{K+1} \log f_{ns}(y_{ns}|\hat{\phi}_{ns}) + d\gamma \log \binom{n-1}{K} + \frac{d}{2}(K+1)\log n \tag{3.6}$$

where we do not include $\log \sum_{k=0}^{n-1} \binom{n-1}{k}^{d-d\gamma}$, a constant not affecting change-points selection. The best estimator of the change-points configuration is the one minimizing emBIC_γ, i.e.

$$\hat{\mathbf{V}}_n \equiv (\hat{j}_1, \cdots, \hat{j}_{\hat{K}}; \hat{K}_n) = \operatorname{argmin} \mathrm{emBIC}_\gamma(\mathbf{V}_n) \equiv \operatorname{argmin} \mathrm{emBIC}_\gamma(j_1, \cdots, j_K; K) \tag{3.7}$$

BIC (3.5) in the context of multiple change-points selection has been proved, under some standard conditions, to be asymptotically consistent in selecting the true change-points by [Yao, 1988]. It is well known that the key condition for a consistent model selection criterion is that its penalty term, e.g. $\frac{d}{2}(K+1)\log n$ in (3.5), is of an order between $o(n)$ and $O(\log \log n)$ and is an increasing function of model dimension, see [Hannan and Quinn, 1979]. The two penalty terms in emBIC_γ (3.6) satisfies this condition. Therefore emBIC is also consistent; its proof can be done by following the same line as in [Yao, 1988] but will not be detailed here as it is not the focus of this chapter.

Even though BIC is asymptotically consistent, it has been found to have tendency to over-estimate the true number of change-points in finite sample situations. This can be seen in the simulations in [Bai and Perron, 2003], [Lavielle, 2005] and [Lavielle and Teyssiere, 2006], for example. Our simulation study in section 3.4 will also confirm this. We expect

the extra penalty term $d\gamma \log \binom{n-1}{K}$ in emBIC_γ (3.6) will effectively correct this finite-sample over-estimation.

A different extension of BIC for detecting multiple change-points is the penalized contrast (PC) of the form $\text{PC}(\mathbf{V}_n) \equiv \text{PC}(j_1, \cdots, j_K; K) = J(\mathbf{V}_n; y_n) + \beta \text{pen}(\mathbf{V}_n)$ given by [Lavielle, 2005]. Here $J(\mathbf{V}_n; y_n)$ measures the fitness of the change-points configuration $\mathbf{V}_n$ to the data $\mathbf{y}_n = (y_1, \cdots, y_n)$, with the minus maximum log-likelihood function being a special form of $J(\mathbf{V}_n; \mathbf{y}_n)$. The penalty term $\text{pen}(\mathbf{V}_n)$ depends K only. The penalization parameter β adjusts the trade-off between $J(\mathbf{V}_n; \mathbf{y}_n)$ and $\text{pen}(\mathbf{V}_n)$, and is estimated according to the maximum likelihood principle which may require an adaptive intensive-computing procedure called the stochastic approximation expectation-maximization (SAEM) algorithm introduced by [Delyon *et al.*, 1999]. In the case of Gaussian mean change-points problem, where $J(\mathbf{V}_n; \mathbf{y}_n)$ is chosen to be the minus twice maximum log-likelihood times the constant variance σ^2, [Birge and Massart, 2001] showed that a penalty term of the form $\text{pen}(\mathbf{V}_n) = (K+1)(1 + c \log \frac{n}{K+1})$ and $\beta = \frac{2\sigma^2}{n}$ is optimal for minimizing the mean sum of squares of residuals. An empirical estimate of $c = 2.5$ was suggested for practical use. It is easy to see that the penalty term $\beta \text{pen}(\mathbf{V}_n)$ here is asymptotically equivalent to the penalty $d\gamma \log \binom{n-1}{K} + \frac{d}{2}(K+1) \log n$ in (3.6). [Lavielle, 2005] used the above penalty function given by [Birge and Massart, 2001] for the variance change-points problem as well, and he called the value of the change-points configuration minimizing the penalized contrast function the MPC estimator, being denoted here as $\tilde{\mathbf{V}}_n \equiv (\tilde{j}_1, \cdots, \tilde{j}_{\tilde{K}}; \tilde{K}) = \text{argmin}\, \text{PC}(\mathbf{V}_n) \equiv \text{argmin}\, \text{PC}(j_1, \cdots, j_K; K)$.

3.1.2 Special Cases on Mean and Variance Change-points

It is interesting to see how emBIC, BIC and penalized contrast (PC) will perform in some common but important cases of multiple change-points. The cases that we focus on are

C1. mean change-points, with constant unknown variance σ^2; i.e. $\phi_s = \mu_s$, $s = 1, \cdots, K+1$;

C2. variance change-points, with constant unknown mean μ; i.e. $\phi_s = \sigma_s^2$, $s = 1, \cdots, K+1$;

C3. mean-variance change-points; i.e. $\phi_s = (\mu_s, \sigma_s^2)$, $s = 1, \cdots, K+1$.

For ease of presentation we now assume the data points $Y_1, \cdots, Y_n$ are i.i.d. normal random variables. The formulas derived based on this assumption can still be used in situations where this assumption does not hold, but their validities are subject to their robustness and large-sample properties.

Consider case C1 first. It is easy to see that the MLE of each μ_s given the change-points configuration is $\hat{\mu}_{ns} = \overline{Y}_{ns}$, which is the sample mean of the sth segment. The MLE of σ^2 given the change-points configuration is $\hat{\sigma}^2_{0nK} = n^{-1}S(j_1, \cdots, j_K)$ with $S(j_1 \cdots, j_K) = \sum_{s=1}^{K+1}\sum_{i=j_{(s-1)}+1}^{j_s}(Y_i - \overline{Y}_{ns})^2$ being the total sum of squared residuals. By (3.6) and a straightforward calculation emBIC for case C1 is, up to an additive constant,

$$\begin{aligned}\mathrm{emBIC}_{1\gamma}(\mathbf{V}_n) &\equiv \mathrm{emBIC}_{1\gamma}(j_1, \cdots, j_K; K)\\ &= \frac{1}{2}\log\hat{\sigma}^2_{0nK} + \gamma\log\binom{n-1}{K} + \frac{1}{2}(K+1)\log n.\end{aligned}$$

Also BIC for case C1 is, up to an additive constant,

$$\mathrm{BIC}_1(\mathbf{V}_n) \equiv \mathrm{BIC}_1(j_1, \cdots, j_K; K) = \frac{1}{2}\log\hat{\sigma}^2_{0nK} + \frac{1}{2}(K+1)\log n.$$

The penalized contrast function for case C1 can be similarly found to be

$$\mathrm{PC}_1(\mathbf{V}_n) \equiv \mathrm{PC}_1(j_1, \cdots, j_K; K) = \hat{\sigma}^2_{0nK} + \frac{2\hat{\sigma}^2_{0nK}}{n}(K+1)\left(1 + c\log\frac{n}{K+1}\right)$$

based on normal log-likelihood.

Now consider case C2. The log-likelihood function is

$$\log L_2(\mu; \sigma_1^2, \cdots, \sigma^2_{K+1}) = -\sum_{s=1}^{K+1}\left[\frac{j_s - j_{(s-1)}}{2}\log\sigma_s^2 + \frac{1}{2\sigma_s^2}\sum_{i=j_{(s-1)}+1}^{j_s}(Y_i - \mu)^2\right]. \tag{3.8}$$

The MLE of $(\mu, \sigma_1^2, \cdots, \sigma^2_{K+1})$ can be found to satisfy

$$\hat{\mu}_{0nK} = \left[\sum_{s=1}^{K+1}\frac{j_s - j_{(s-1)}}{\hat{\sigma}^2_{ns}}\right]^{-1}\sum_{s=1}^{K+1}\sum_{i=j_{(s-1)}+1}^{j_s}\frac{Y_i}{\hat{\sigma}^2_{ns}} \quad \text{and} \quad \hat{\sigma}^2_{ns} = \frac{\sum_{i=j_{(s-1)}+1}^{j_s}(Y_i - \hat{\mu}_{0nK})^2}{j_s - j_{(s-1)}}, \tag{3.9}$$

$s = 1, \cdots, K+1$, which can be computed using Newton's iteration method. By (3.6), (3.8) and (3.9), emBIC for case C2 is, up to an additive constant,

$$\begin{aligned}\mathrm{emBIC}_{2\gamma}(\mathbf{V}_n) &\equiv \mathrm{emBIC}_{2\gamma}(j_1, \cdots, j_K; K)\\ &= \frac{1}{2}\sum_{s=1}^{K+1}(j_s - j_{(s-1)})\log\hat{\sigma}^2_{ns} + \gamma\log\binom{n-1}{K} + \frac{1}{2}(K+1)\log n.\end{aligned}$$

Also BIC for case C2 is, up to an additive constant,

$$\mathrm{BIC}_2(\mathbf{V}_n) \equiv \mathrm{BIC}_2(j_1, \cdots, j_K; K) = \frac{1}{2}\sum_{s=1}^{K+1}(j_s - j_{(s-1)}) \log \hat{\sigma}^2_{ns} + \frac{1}{2}(K+1)\log n.$$

[Lavielle, 2005] gave a penalized contrast function for case C2 as

$$\mathrm{PC}_2(\mathbf{V}_n) \equiv \mathrm{PC}_2(j_1, \cdots, j_K; K) = \frac{1}{n}\sum_{s=1}^{K+1} \log\left(\frac{\sum_{i=j_{(s-1)}+1}^{j_s}(Y_i - \bar{Y}_n)^2}{j_s - j_{(s-1)}}\right)^{j_s - j_{(s-1)}} + \beta(K+1) \tag{3.10}$$

where $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Instead of assigning a single value for β and using (3.10) to estimate $(j_1, \cdots, j_K)$ and K simultaneously, [Lavielle, 2005] proposed to estimate $(j_1, \cdots, j_K)$, for each possible K given, by $(\tilde{j}_1, \cdots, \tilde{j}_K)$ that minimised (3.10) at $\beta = 0$. The estimate $(\tilde{j}_1, \cdots, \tilde{j}_K)$ can be computed by using a dynamic programming algorithm. He then found that $\mathrm{argmin}_K \mathrm{PC}_2(\tilde{j}_1, \cdots, \tilde{j}_K; K)$, as a function of β, varied with β according to a step-decreasing function. By exploiting the curvature property of this function, he was able to develop a heuristic approach to estimate the optimal K without actually using β in the computing involved.

Finally we consider case C3. It is easy to see that emBIC for case C3, up to an additive constant is

$$\begin{aligned}\mathrm{emBIC}_{3\gamma}(\mathbf{V}_n) &\equiv \mathrm{emBIC}_{3\gamma}(j_1, \cdots, j_K; K) \\ &= \frac{1}{2}\sum_{s=1}^{K+1} \log\left(\frac{\sum_{i=j_{(s-1)}+1}^{j_s}(Y_i - \overline{Y}_{ns})^2}{j_s - j_{(s-1)}}\right)^{j_s - j_{(s-1)}} + 2\gamma \log\binom{n-1}{K} + (K+1)\log n. \end{aligned} \tag{3.11}$$

Also BIC for case C3 is, up to an additive constant,

$$\begin{aligned}\mathrm{BIC}_3(\mathbf{V}_n) &\equiv \mathrm{BIC}_3(j_1, \cdots, j_K; K) \\ &= \frac{1}{2}\sum_{s=1}^{K+1} \log\left(\frac{\sum_{i=j_{(s-1)}+1}^{j_s}(Y_i - \overline{Y}_{ns})^2}{j_s - j_{(s-1)}}\right)^{j_s - j_{(s-1)}} + (K+1)\log n. \end{aligned} \tag{3.12}$$

And the normal log-likelihood based penalized contrast function for case C3 is

$$\mathrm{PC}_3(\mathbf{V}_n) \equiv \mathrm{PC}_3(j_1, \cdots, j_K; K) = \frac{1}{2}\sum_{s=1}^{K+1} \log\left(\frac{\sum_{i=j_{(s-1)}+1}^{j_s}(Y_i - \overline{Y}_{ns})^2}{j_s - j_{(s-1)}}\right)^{j_s - j_{(s-1)}} + \beta(K+1) \tag{3.13}$$

Again, (3.13) is not used to estimate $(j_1, \cdots, j_K)$ and K simultaneously, as indicated in [Lavielle, 2005].

3.2 Iterative Stochastic Search of Change-points

3.2.1 Computational Challenges and Existent Methods

Given a change-points selection criterion such as emBIC, BIC or PC, the underlying change-points configuration is estimated as the one that minimises the criterion. But it is not straightforward to actually use such a criterion to compute the configuration estimate. Note that in theory there are $\binom{n-1}{K}$ possible configurations for each given K for a sequence of n data points. Thus there are in total $\sum_{K=0}^{n-1}\binom{n-1}{K} = 2^{n-1}$ possible change-points configurations in the data. It is computationally infeasible, even when the sample size n is moderately large, to evaluate the criterion function for all 2^{n-1} configurations to find the best change-points estimation.

[Vostrikova, 1981] proposed a hypothesis testing based sequential binary segmentation procedure to compute the significant change-points in the data. The procedure first finds a significant change-point by a test with null against one change-point hypotheses. Once such a change-point is detected at a given significance level, use it to divide the data into two sub-sequences. The procedure then tests for significant change-point in each sub-sequence; and if a significant change-point is found, use it to divide the associated sub-sequence into two further sub-sequences. The procedure continues this way until no significant change-point can be found and no sub-sequence can be further divided. For such a sequential testing procedure, it is usually difficult to balance the overall power of the procedure against its overall significance levels, and also against the involved computing expenses. Another drawback of such a procedure is its inability to recover any underlying change-point once this point is failed to be detected at some stage. Further, such a procedure tends to over-estimate the number of change-points ([Lavielle and Teyssiere, 2006]). Nevertheless, many papers have used the binary segmentation or similar sequential procedures to search for multiple change-points. See e.g. [Bai and Perron, 2003], [Chen and Gupta, 1997], and [Inclan and Tiao, 1994].

In particular, [Chen and Gupta, 1997] used the binary segmentation procedure to implement a BIC induced sequential test on multiple variance change-points detection. Their BIC induced test can also deal with change-points of a general parameter of dimension d. We now briefly describe their test in the context of general parameter change-points. Let SIC(n) be the value of BIC calculated when assuming no change-point in the data $Y_1, \cdots, Y_n$; and

$\mathrm{SIC}(j)$, $1 < j < n$, be that when assuming one change-point at location j. Note we use $\mathrm{SIC}(\cdot)$ here to distinguish it from $\mathrm{BIC}(\mathbf{V}_n)$ in (3.5) where K change-points are assumed. The null hypothesis of no change-point will be rejected at significance level α if

$$\mathrm{SIC}(n) \geq \min_{1<j<n} \mathrm{SIC}(j) + c_{n\alpha}, \tag{3.14}$$

and accordingly $\breve{j} = \mathrm{argmin}_{1<j<n}\,\mathrm{SIC}(j)$ is used to estimate the location of the significant change-point detected. Here $c_{n\alpha}$ is the right-tail level α critical value for the asymptotic distribution of $\mathrm{SIC}(n) - \min_{1<j<n}\mathrm{SIC}(j)$ under the null hypothesis. By Theorem 3.1 of [Chen and Gupta, 1997] and Theorem 1.3.1 of [Csörgo and Horváth, 1997], it can be shown that

$$c_{n\alpha} = \frac{1}{2}\left\{-\frac{1}{a(\log n)}\log\log[1-\alpha+\exp(-2e^{b(\log n)})]^{-1/2} + \frac{b(\log n)}{a(\log n)}\right\}^2 - \frac{d}{2}\log n, \tag{3.15}$$

where

$$a(\log n) = (2\log\log n)^{1/2} \quad \text{and} \quad b(\log n) = 2\log\log n + \frac{d}{2}\log\log\log n - \log\Gamma(\frac{d}{2}) \tag{3.16}$$

with $\Gamma(\cdot)$ being the gamma function. Note that n in (3.14) refers to the size of the data segment being tested in the binary segmentation procedure.

As mentioned in section 3.1, [Lavielle, 2005] used an intensive computing procedure to calculate the MPC estimate of the change-points configuration. The procedure involved both a dynamic programming algorithm for computing the contrast function and a heuristic determination of the penalization parameter β and the number of change-points K.

[Lavielle, 2005] also proposed to estimate the change-points configuration $\mathbf{V}_n$ by maximizing its posterior distribution

$$P(\mathbf{V}_n|y_n;\alpha,\beta) \propto e^{-\alpha[J(\mathbf{V}_n;y_n)+\beta\cdot\mathrm{pen}(\mathbf{V}_n)]}, \quad \alpha > 0;$$

and called the resultant estimate the maximum *a posteriori* (MAP) estimate. This Bayesian approach involves using SAEM for estimating (α,β) and simulated annealing for estimating $\mathbf{V}_n$ and K, which has been found to be computationally slow and tends to produce worse results than the MPC approach.

3.2.2 Change-points Sampling and Search by Gibbs Sampler

Now we propose a different computing procedure for finding the change-points configuration that minimises the empirical Bayesian information criterion emBIC derived in section 3.1. We will show that the new procedure is computationally efficient, easy to implement and interpret, and possesses some optimality properties under regularity conditions.

We now Use the indicator process $\mathbf{V}_n$ defined in section 3.1 to represent the change-points configuration. It is easy to see that the estimate $\hat{\mathbf{V}}_n$ in (3.7), which minimizes the $\text{emBIC}_\gamma(\mathbf{V}_n)$ given by (3.6), is also the one that maximizes the following probability function

$$\text{PemBIC}_{\tau\gamma}(\mathbf{V}_n) = D(y_n; \tau, \gamma) \exp\{-\tau \cdot \text{emBIC}_\gamma(\mathbf{V}_n)\}, \tag{3.17}$$

where $D(y_n; \tau, \gamma) = \left[\sum_{\mathbf{V}_n} \exp\{-\tau \cdot \text{emBIC}_\gamma(\mathbf{V}_n)\}\right]^{-1}$ is the normalization constant. Note $\text{PemBIC}_{\tau\gamma}(\mathbf{V}_n)$ is deemed to be a joint probability mass function of the latent binary random vector $\mathbf{V}_n = (V_1, \cdots, V_n)$ given the observed data $Y_1 = y_1, \cdots, Y_n = y_n$. This implies that, if we can generate samples of $\mathbf{V}_n$ from (3.17), $\hat{\mathbf{V}}_n$ would have the highest probability among all possible values of $\mathbf{V}_n$ to appear in the samples, and tend to appear early in the samples. Hence the empirical distribution of the generated $\mathbf{V}_n$ samples can be used to consistently estimate $\hat{\mathbf{V}}_n$. In particular we can either use the empirical marginal distributions of $V_1, \cdots, V_{n-1}$ (note $V_n \equiv 0$) based on the samples, or do finite search over the samples, to find the optimal $\mathbf{V}_n$ in the samples. This optimal $\mathbf{V}_n$ is expected to converge to $\hat{\mathbf{V}}_n$ very quickly when sufficient samples are generated; and will eventually converge to the true change-points configuration in the data when $n \to \infty$. Comparing to an exhaustive search to find $\hat{\mathbf{V}}_n$, which involves enumerating all 2^{n-1} possible configurations of $\mathbf{V}_n$, the search based on generating random samples from (3.17) requires just a fractional amount of the computing for enumeration.

The normalization constant $D(y_n; \tau, \gamma)$, which is actually a function of y_n and (τ, γ), involves enumerating 2^{n-1} terms thus cannot be easily evaluated when n is large. Hence it is difficult to generate random samples from $\text{PemBIC}_{\tau\gamma}(\mathbf{V}_n)$ directly. However, a Markov chain Monte Carlo method can be used to generate random samples in such a situation. In particular, the Gibbs sampler ([Casella and George, 1992]) can be used here in a very simple manner, because the conditional distributions $\text{PemBIC}_{\tau\gamma}(V_i|\mathbf{V}_{-i})$, which are required in using the Gibbs sampler, are Bernoulli ones with easily computable probabilities of "success". Here

$\mathbf{V}_{-i} = (\mathbf{V}_{1:(i-1)}, \mathbf{V}_{(i+1):n})$, $i = 1, \cdots, n-1$; and $\mathbf{V}_{i:j} = (V_i, \cdots, V_j)$ if $i \le j$ and $= \emptyset$ otherwise. It is easy to see that the conditional probability mass function of V_i given $\mathbf{V}_{-i}$ is

$$\text{PemBIC}_{\tau\gamma}(V_i|\mathbf{V}_{-i}) = \frac{e^{-\tau \cdot \text{emBIC}_\gamma(\mathbf{V}_n)}}{e^{-\tau \cdot \text{emBIC}_\gamma(\mathbf{V}_{1:(i-1)}, V_i=1, \mathbf{V}_{(i+1):n})} + e^{-\tau \cdot \text{emBIC}_\gamma(\mathbf{V}_{1:(i-1)}, V_i=0, \mathbf{V}_{(i+1):n})}}, \quad (3.18)$$

$i = 1, \cdots, n-1$, which does not involve the intractable $D(y_n; \tau, \gamma)$.

We propose to compute the change-points configuration estimate based on the samples that are generated from $\text{PemBIC}_{\tau\gamma}(\mathbf{V}_n)$ by using the Gibbs sampler. Our computing procedure is described by the following algorithm.

Algorithm 1. *Gibbs sampler + emBIC for detecting change-points*

- Arbitrarily choose a starting indicator vector $\mathbf{V}_n^{(0)} = (V_1^{(0)}, \cdots, V_{n-1}^{(0)}, 0)$. For example, take $(V_1^{(0)}, \cdots, V_{n-1}^{(0)})$ to be generated from Bernoulli(0.2), or set each $V_i^{(0)}$ to 1 if i is a multiple of 10 and 0 otherwise, etc..

- Generate $V_i^{(\ell)}$ from the probability mass function $\text{PemBIC}_{\tau\gamma}(V_i|\mathbf{V}_{1:(i-1)}^{(\ell)}, \mathbf{V}_{(i+1):n}^{(\ell-1)})$, where $i = 1, \cdots, n-1$ and $\ell = 1, \cdots, L$, sequentially for given L, τ and γ.

- Return the sampled sequence $\{\mathbf{V}_n^{(1)}, \cdots, \mathbf{V}_n^{(L)}\}$ where $\mathbf{V}_n^{(l)} = (V_1^{(\ell)}, \cdots, V_{n-1}^{(\ell)}, 0)$.

- Compute the marginal empirical distribution for each component of $\mathbf{V}_{1:(n-1)}$. This is equivalent to computing the marginal sampling probabilities $p_i^{(L)} = \frac{1}{L}\sum_{\ell=1}^{L} V_i^{(\ell)}$, $i = 1, \cdots, n-1$. Then determine $\mathbf{V}_n^{*L} = (V_1^{*L}, \cdots, V_{n-1}^{*L}, 0)$, where $V_i^{*L} = 1$ if $p_i^{(L)} > p^*$ and $V_i^{*L} = 0$ otherwise, with p^* being a given threshold probability (e.g. $p^* = 0.5$). Note that the empirical distribution of any function of $\mathbf{V}_n$ can be similarly computed.

- The generated $V_i^{(\ell)}$ values result in $(n-1)L$ vectors of $\mathbf{V}_n$, each being of the form $(V_1^{(\ell)}, \cdots, V_i^{(\ell)}, V_{i+1}^{(\ell-1)}, \cdots, V_{n-1}^{(\ell-1)}, 0)$; $i = 1, \cdots, n-1$ and $\ell = 1, \cdots, L$. Compute the emBIC for these vectors; denote the results as $\text{emBIC}_\gamma^{(1,1)}, \cdots, \text{emBIC}_\gamma^{(n-1,L)}$. Also compute $\mathbf{V}_n^{+L} = \arg\min\{\text{emBIC}_\gamma^{(i,\ell)}; 1 \le i \le n-1, 1 \le \ell \le L\}$. Note $(\mathbf{V}_n^{+L}, \text{emBIC}_\gamma(\mathbf{V}_n^{+L}))$ should have little difference from $(\mathbf{V}_n^{*L}, \text{emBIC}_\gamma(\mathbf{V}_n^{*L}))$ when L is sufficiently large. We use $\mathbf{V}_n^{*L}$ as the change-points configuration estimate if $\text{emBIC}_\gamma(\mathbf{V}_n^{*L}) < \text{emBIC}_\gamma(\mathbf{V}_n^{+L})$; and use $\mathbf{V}_n^{+L}$ otherwise.

In section 3.2.3 we will reason that, under regularity conditions, both $\mathbf{V}_n^{*L}$ and $\mathbf{V}_n^{+L}$ converge to $\hat{\mathbf{V}}_n$ almost surely when $L \to \infty$, and further converge to the true population change-points configuration almost surely when $n \to \infty$. In section 3.3 we will provide detail on how to implement Algorithm 1; and will address practice issues, e.g. how to choose L, τ, γ and p^* etc.. At the moment we would like to just provide the following remark.

Remark 1. [Qian, 1999] introduced the idea of combining the Gibbs sampler with a variable selection criterion and performing variable selection by iterative sampling and stochastic search. This idea was later used in [Qian and Field, 2002] for logistic regression variable selection; in [Qian and Zhao, 2007] for autoregressive moving average (ARMA) time series variable selection; and in [Cui *et al.*, 2010] for general estimating equation variable selection. The context in which we apply this idea here is very different from the aforementioned ones. For example, the model space in which we are to find the change-points is that of the latent indicator process $(V_1, \cdots, V_{n-1})$, which depends on the sample size of the data. This is in contrast to the model space in variable selection which does not depend on the sample size.

Gibbs sampler can also be used with the BIC induced test of [Chen and Gupta, 1997] for change-points detection. The key is to replace (3.18) by the conditional probability function

$$\mathrm{PtBIC}_{\tau\alpha}(V_i = 1|\mathbf{V}_{-i}) = \frac{1}{1 + e^{\tau[\mathrm{BIC}(\mathbf{V}_{1:(i-1)}, V_i=1, \mathbf{V}_{(i+1):n}) + c_{n\alpha} - \mathrm{BIC}(\mathbf{V}_{1:(i-1)}, V_i=0, \mathbf{V}_{(i+1):n})]}} \tag{3.19}$$

and $\mathrm{PtBIC}_{\tau\alpha}(V_i = 0|\mathbf{V}_{-i}) = 1 - \mathrm{PtBIC}_{\tau\alpha}(V_i = 1|\mathbf{V}_{-i})$, $i = 1, \cdots, n-1$, where $\mathrm{BIC}(\mathbf{V}_n)$ is given by (3.5) and $c_{n\alpha}$ is given by (3.15). The joint probability mass function resulting in (3.19) is

$$\mathrm{PtBIC}_{\tau\alpha}(\mathbf{V}_n) = D'(y_n; \tau, \alpha) \exp\{-\tau \cdot \mathrm{tBIC}_\alpha(\mathbf{V}_n)\} \tag{3.20}$$

where

$$\mathrm{tBIC}_\alpha(\mathbf{V}_n) = -\sum_{s=1}^{K+1} \log f_s(\mathbf{y}_s|\hat{\phi}_s) + \frac{d}{2}(K+1)\log n + Kc_{n\alpha} = \mathrm{BIC}(\mathbf{V}_n) + c_{n\alpha}\sum_{i=1}^{n-1} V_i. \tag{3.21}$$

Now it amounts to finding the change-points configuration minimizing $\mathrm{tBIC}_\alpha(\mathbf{V}_n)$ rather than minimizing $\mathrm{BIC}(\mathbf{V}_n)$. The computing is processed through the following algorithm.

Algorithm 2. *Gibbs sampler + BIC induced test for detecting change-points*

- Arbitrarily choose a starting indicator vector $\mathbf{V}_n^{(0)} = (V_1^{(0)}, \cdots, V_{n-1}^{(0)}, 0)$. For example, take $(V_1^{(0)}, \cdots, V_{n-1}^{(0)})$ to be generated from Bernoulli(0.2), or set each $V_i^{(0)}$ to 1 if i is a multiple of 10 and 0 otherwise, etc..

- Generate $V_i^{(\ell)}$ from the Bernoulli probability function $\text{PtBIC}_{\tau\alpha}(V_i|\mathbf{V}_{1:(i-1)}^{(\ell)}, \mathbf{V}_{(i+1):n}^{(\ell-1)})$, where $i = 1, \cdots, n-1$ and $\ell = 1, \cdots, L$, sequentially for given L, τ and α.

- Return the sampled sequence $\{\mathbf{V}_n^{(1)}, \cdots, \mathbf{V}_n^{(L)}\}$ where $\mathbf{V}_n^{(l)} = (V_1^{(\ell)}, \cdots, V_{n-1}^{(\ell)}, 0)$.

- Compute the marginal empirical distribution for each component of $\mathbf{V}_{1:(n-1)}$. This is equivalent to computing the marginal sampling probabilities $\breve{p}_i^{(L)} = \frac{1}{L}\sum_{\ell=1}^{L} V_i^{(\ell)}$, $i = 1, \cdots, n-1$. Then determine $\breve{\mathbf{V}}_n^{*L} = (\breve{V}_1^{*L}, \cdots, \breve{V}_{n-1}^{*L}, 0)$, where $\breve{V}_i^{*L} = 1$ if $\breve{p}_i^{(L)} > p^*$ and $\breve{V}_i^{*L} = 0$ otherwise, with p^* being a given threshold probability (e.g. $p^* = 0.5$). The empirical distribution of any function of $\mathbf{V}_n$ can be similarly computed.

- The generated $V_i^{(\ell)}$ values result in $(n-1)L$ vectors of $\mathbf{V}_n$, each being of the form $(V_1^{(\ell)}, \cdots, V_i^{(\ell)}, V_{i+1}^{(\ell-1)}, \cdots, V_{n-1}^{(\ell-1)}, 0)$; $i = 1, \cdots, n-1$ and $\ell = 1, \cdots, L$. Compute the tBIC values for these vectors; denote the results as $\text{tBIC}_\alpha^{(1,1)}, \cdots, \text{tBIC}_\alpha^{(n-1,L)}$. Also compute $\breve{\mathbf{V}}_n^{+L} = \arg\min\{\text{tBIC}_\alpha^{(i,\ell)}; 1 \le i \le n-1, 1 \le \ell \le L\}$. Note $(\breve{\mathbf{V}}_n^{+L}, \text{tBIC}_\alpha(\breve{\mathbf{V}}_n^{+L}))$ should have little difference from $(\breve{\mathbf{V}}_n^{*L}, \text{tBIC}_\alpha(\breve{\mathbf{V}}_n^{*L}))$ when L is sufficiently large. We use $\breve{\mathbf{V}}_n^{*L}$ as the change-points configuration estimate if $\text{tBIC}_\alpha(\breve{\mathbf{V}}_n^{*L}) < \text{tBIC}_\alpha(\breve{\mathbf{V}}_n^{+L})$; and use $\breve{\mathbf{V}}_n^{+L}$ otherwise.

As for Algorithm 1, it can be similarly reasoned that, under regularity conditions, both $\breve{\mathbf{V}}_n^{*L}$ and $\breve{\mathbf{V}}_n^{+L}$ converge almost surely to the change-points configuration minimizing $\text{tBIC}_\alpha(\mathbf{V}_n)$ in (3.21) when $L \to \infty$, and further converge to the true population change-points configuration almost surely when $n \to \infty$. Both $\breve{\mathbf{V}}_n^{*L}$ and $\breve{\mathbf{V}}_n^{+L}$ are expected to be asymptotically equivalent to the solution of the binary segmentation procedure of [Chen and Gupta, 1997] as being summarized in section 3.2.1. But Algorithm 2 gives an iterative stochastic search procedure, which does not have those difficult finite-sample issues of the binary segmentation and sequential procedures as mentioned in section 3.2.1. In section 3.3 we will discuss on how to implement Algorithm 2; and will address practice issues, e.g. how to choose L, τ, α and p^* etc..

Clearly, an algorithm combining Gibbs sampler with BIC in a way similar to that in Algorithms 1 and 2 can be formulated to detect change-points. The resultant algorithm would generate random samples from $\text{PBIC}_\tau(\mathbf{V}_n) \propto \exp\{-\tau \cdot \text{BIC}(\mathbf{V}_n)\}$. But we will not proceed with it since BIC tends to over-detect change-points.

MCMC including the Gibbs sampler were combined with the penalty contrast function to estimate the change-points by [Lavielle and Lebarbier, 2001] in a maximum *a posteriori* (MAP) approach. The combination is done in a much more complicated framework than the one presented in Algorithms 1 and 2, yet the resultant MAP estimates are not as good as the MPC ones. The penalty contrast framework does not allow for a simple integration with the Gibbs sampler as well as MCMC.

3.2.3 Asymptotic Optimality of Gibbs Sampler Plus emBIC or tBIC

The sequence $\{\mathbf{V}_n^{(1)}, \cdots, \mathbf{V}_n^{(L)}\}$ generated by each of Algorithms 1 and 2 is a Markov chain rather than a sequence of i.i.d. samples. It can be verified the generated Markov chain is aperiodic, irreducible and reversible; thus it is uniformly ergodic and converges to its stationary distribution given by (3.17) or (3.20) as $L \to \infty$. Hence by the ergodicity theorem [Robert and Richardson, 1998, p.2] the change-points detected by Algorithm 1 or 2 converge to that defined by minimizing emBIC or tBIC almost surely as $L \to \infty$ with respect to the probability space defined by (3.17) or (3.20). The change-points minimizing emBIC or tBIC, although can be proved to be strongly consistent using e.g. [Yao, 1988], are not necessarily the same as the true population change-points associated with the data. So it is of interest to see how likely the true population change-points can be detected by Algorithms 1 and 2.

First we need to establish the concept of true population change-points. Recall that the data is $\mathbf{Y}_n = (Y_{n1}, \cdots, Y_{nn})$, which can be sampled from a random process $\{X(t), t \in (0, 1]\}$ such that $Y_{ni} = X(i/n)$, $i = 1, \cdots, n$. Suppose there exist K_0 constants $t_1, \cdots, t_{K_0}$ satisfying $0 = t_0 < t_1 < t_2 < \cdots < t_{K_0} < t_{K_0+1} = 1$, and $K_0 + 1$ distinct probability distributions $F_1, \cdots, F_{K_0+1}$ such that $X(t) \stackrel{d}{=} F_k$ for any $t \in (t_{k-1}, t_k]$, $k = 1, \cdots, K_0 + 1$. It is easy to see that $t_1, \cdots, t_{K_0}$ are the *true population change-points* of the random process $X(t)$, and we formally call them the *population fraction change-points* of the data $Y_{n1}, \cdots, Y_{nn}$. [Yao, 1988] has shown that the BIC estimator $\check{K}_n$ of K_0 is strongly consistent, i.e. $\check{K}_n \stackrel{\text{a.s.}}{\to} K_0$ as $n \to \infty$

under regularity conditions. This strong consistency can be extended to the emBIC and tBIC estimators of K_0 following a similar proof. Note that, although being strongly consistent, the BIC estimator $\check{K}_n$ tends to overestimate K_0 in finite sample situations.

Given a number k_n, $K_0 \le k_n \le n-1$, let $\hat{j}_1, \cdots, \hat{j}_{k_n}$ be the k_n most probable change-points of Y_n obtained from minimizing emBIC or tBIC. It is easy to see that $\hat{j}_1, \cdots, \hat{j}_{k_n}$ are just the maximum likelihood estimates. Following the asymptotic techniques in [Shi *et al.*, 2009] and [Qian *et al.*, 2014], it is not difficult to show that, when $k_n \ge K_0$, there exists a size K_0 subset $\{s_1, \cdots, s_{K_0}\}$ of $\{1, \cdots, k_n\}$ such that $|n^{-1}\hat{j}_{ns_i} - t_i| = o(n^{-1}q(n))$ a.s. for any $q(n) \uparrow \infty$, $i = 1, \cdots, K_0$. Namely, $(\hat{j}_1, \cdots, \hat{j}_{k_n}; k_n)$, with possible redundancy, almost surely gives a correct configuration of the true population change-points. Using these techniques one can also show that the difference between the maximum log-likelihood of the data over the configuration $(\hat{j}_1, \cdots, \hat{j}_{k_n}; k_n)$ and that over the true population change-points configuration is of order $O(\log\log n)$ a.s.. And this difference would be of order $O(n)$ a.s. if $(\hat{j}_1, \cdots, \hat{j}_{k_n}; k_n)$ does not give a correct configuration (e.g. if $k_n < K_0$).

For n data points $Y_{n1}, \cdots, Y_{nn}$, the set of all possible configurations of change-points can be denoted as $\bigcup_{k=0}^{n-1} \mathcal{S}_{nk} = \{\emptyset\} \bigcup \big(\bigcup_{k=1}^{n-1} \{(j_1, \cdots, j_k) : 1 \le j_1 < \cdots < j_k \le n-1\}\big)$, which contains $2^{n-1} - 1$ non-empty elements. With the discussions so far, we can divide $\bigcup_{k=0}^{n-1} \mathcal{S}_{nk}$ into two subsets $\mathcal{M}_n(\varepsilon)$ and its complement $\mathcal{M}_n^c(\varepsilon)$ for any $\varepsilon > 0$ sufficiently small and n sufficiently large. Here

$$\begin{aligned}\mathcal{M}_n(\varepsilon) \;=\; & \{(j_1, \cdots, j_{k_n}; k_n) :\; K_0 \le k_n \le n-1;\; 1 \le j_1 < \cdots < j_{k_n} \le n-1; \\ & \text{and for any } s = 1, \cdots, K_0, \text{ there exists exactly one } s_n \in \{1, \cdots, k_n\} \\ & \text{such that } |n^{-1}j_{s_n} - t_s| < \varepsilon.\}\end{aligned}$$

It is easy to see that the cardinality of $\mathcal{M}_n(\varepsilon)$ is $|\mathcal{M}_n(\varepsilon)| = 2^{n-1-K_0}$ if ε is sufficiently small and n is sufficiently large such that $0 < \varepsilon < (2n)^{-1}$; and each configuration in $\mathcal{M}_n(\varepsilon)$ covers all true population change-points of the data but may also contain some non-change-points. The cardinality of $\mathcal{M}_n^c(\varepsilon)$ can be found to be $2^{n-1} - 2^{n-1-K_0} = 2^{n-1}(1 - 2^{-K_0})$ if ε is sufficiently small and n is sufficiently large such that $0 < \varepsilon < (2n)^{-1}$. We single out a particular subset of $\mathcal{M}_n(\varepsilon)$, which is

$$\mathcal{M}_{0n}(\varepsilon) = \{(J_1, \cdots, J_{K_0}) :\; 1 \le J_1 < \cdots < J_{K_0} \le n-1;\; |n^{-1}J_k - t_k| < \varepsilon,\, k = 1, \cdots, K_0.\}$$

Note $\mathcal{M}_{0n}(\varepsilon) = \{\emptyset\}$ if $K_0 = 0$, otherwise $\mathcal{M}_{0n}(\varepsilon)$ contains only the true configuration of the change-points if ε is sufficiently small and n is sufficiently large such that $0 < \varepsilon < (2n)^{-1}$.

Following the above discussions on $\mathcal{M}_n(\varepsilon)$ and $\mathcal{M}_n^c(\varepsilon)$, we can reasonably assume the following two conditions hold for the maximum log-likelihood function for the data Y_n.

(A.1) For any $(j_1, \cdots, j_{k_n}; k_n) \in \mathcal{M}_n(\varepsilon)$ and $(j'_1, \cdots, j'_{k'_n}; k'_n) \in \mathcal{M}_n(\varepsilon)$ with $0 < \varepsilon < (2n)^{-1}$ and n sufficiently large

$$\left| \sum_{s=1}^{k_n+1} \log f_{ns}(Y_{ns} \mid \hat{\phi}_{ns}) - \sum_{s'=1}^{k'_n+1} \log f_{ns'}(Y_{ns'} \mid \hat{\phi}_{ns'}) \right| \asymp \log\log n \quad \text{a.s., where}$$

$Y_{ns} = (Y_{n(j_{(s-1)}+1)}, \cdots, Y_{nj_s})$ and $a_n \asymp b_n$ means $0 < \underline{\lim}_{n\to\infty} a_n b_n^{-1} \leq \overline{\lim}_{n\to\infty} a_n b_n^{-1} < \infty$.

(A.2) For any $(j_1, \cdots, j_{k_n}; k_n) \in \mathcal{M}_n(\varepsilon)$ and $(j''_1, \cdots, j''_{k''_n}; k''_n) \in \mathcal{M}_n^c(\varepsilon)$ with $0 < \varepsilon < (2n)^{-1}$ and n sufficiently large

$$0 < \sum_{s=1}^{k_n+1} \log f_{ns}(Y_{ns} \mid \hat{\phi}_{ns}) - \sum_{s''=1}^{k''_n+1} \log f_{ns''}(Y_{ns''} \mid \hat{\phi}_{ns''}) \asymp n \text{ a.s.}$$

We have the following results when use randomly generated samples from $\text{PemBIC}_{\tau\gamma}(\mathbf{V}_n)$ or $\text{PtBIC}_{\tau\alpha}(\mathbf{V}_n)$ for detecting change-points.

Proposition 1 *Consider the change-points detection criteria BIC in (3.5), emBIC$_\gamma$ in (3.6) and tBIC$_\alpha$ in (3.21); and their induced probability mass functions PBIC$_\tau(\mathbf{V}_n)$, PemBIC$_{\tau\gamma}(\mathbf{V}_n)$ in (3.17) and PtBIC$_{\tau\alpha}(\mathbf{V}_n)$ in (3.20). Suppose the number of the population fraction change-points is K_0 corresponding to data $Y_{n1}, \cdots, Y_{nn}$ which are observed from the random process $\{X(t), 0 < t \leq 1\}$. Also suppose both conditions (A.1) and (A.2) are satisfied. Let $Pr(\cdot)$ be a generic notation of a probability statement with respect to PBIC$_\tau(\mathbf{V}_n)$, PemBIC$_{\tau\gamma}(\mathbf{V}_n)$ or PtBIC$_{\tau\alpha}(\mathbf{V}_n)$; and a.s. be with respect to the probability space of $\{X(t), 0 < t \leq 1\}$. Also denote ξ, ξ_1, ξ_2 etc. as some generic positive constants. Then we have the following results.*

(R.1) $Pr(\mathcal{M}_n(\varepsilon)) \sim [1 + (2^{K_0} - 1)e^{-\xi n}]^{-1}$ *a.s.. Here* $a_n \sim b_n$ *means* $a_n b_n^{-1} \to 1$.

(R.2) $\dfrac{Pr(\mathcal{M}_n(\varepsilon))}{Pr(\mathcal{M}_n^c(\varepsilon))} \asymp (2^{K_0} - 1)^{-1} e^{\xi n}$ *a.s..*

(R.3) $Pr(\mathcal{M}_{0n}(\varepsilon)) \geq [1 + (2^{n-1-K_0} - 1)\xi_1 n^{-\xi_2}]^{-1} Pr(\mathcal{M}_n(\varepsilon))$ *a.s..*

Proposition 1 essentially says that the probability of a change-points configuration from $\mathcal{M}_n(\varepsilon)$ being selected is asymptotically 1 (R.1); the probability of selecting a correct configuration is exponentially larger than that selecting an incorrect configuration (R.2); but the lower bound in (R.3) is weak and not sufficient to guarantee a large probability of selecting the true population configuration of change-points.

Proof of Proposition 1 is tedious but straightforward by knowing that there are 2^{n-1-K_0} configurations in $\mathcal{M}_n(\varepsilon)$ and $2^{n-1-K_0}(2^{K_0} - 1)$ configurations in $\mathcal{M}_n^c(\varepsilon)$; and by knowing that $\Pr(\mathbf{V}_{n0})/\Pr(\mathbf{V}_{n1}) \asymp n^{\xi_2}$ a.s. and $\Pr(\mathbf{V}_{n2})/\Pr(\mathbf{V}_{n3}) \asymp e^{\xi n}$ a.s. under conditions (A.1) and (A.2) for any $\mathbf{V}_{n0} \in \mathcal{M}_{0n}(\varepsilon)$, $\mathbf{V}_{n1} \in \mathcal{M}_n(\varepsilon) \backslash \mathcal{M}_{0n}(\varepsilon)$, $\mathbf{V}_{n2} \in \mathcal{M}_n(\varepsilon)$ and $\mathbf{V}_{n2} \in \mathcal{M}_n^c(\varepsilon)$.

We know a change-points configuration $(j_1, \cdots, j_{k_n}; k_n)$ can be equivalently represented by an indicator vector $\mathbf{V}_n = (V_1, \cdots, V_{n-1}, 0)$ where those V_i's with $i \in \{j_1, \cdots, j_{k_n}\}$ equal 1 and the other ones equal 0. From the induced probability mass functions $\text{PBIC}_\tau(\mathbf{V}_n)$, $\text{PemBIC}_{\tau\gamma}(\mathbf{V}_n)$ in (3.17) and $\text{PtBIC}_{\tau\alpha}(\mathbf{V}_n)$ in (3.20), we can easily write down the marginal probability mass function of each V_i, $i = 1, \cdots, n-1$:

$$\begin{aligned} \Pr(V_i = 1) &= \Pr(i \in \{j_1, \cdots, j_{k_n}\}) = \sum_{V_i=1,\, V_{-i} \in \{0,1\}^{n-2}} \Pr(\mathbf{V}_n) \\ \Pr(V_i = 0) &= \Pr(i \in \{j_1, \cdots, j_{k_n}\}) = \sum_{V_i=0,\, V_{-i} \in \{0,1\}^{n-2}} \Pr(\mathbf{V}_n). \end{aligned}$$

We have the following results for the marginal distributions of those V_i components specified by the change-points configuration $(J_{n1}, \cdots, J_{nK_0}; K_0)$ in $\mathcal{M}_{0n}(\varepsilon)$.

Proposition 2 *Consider the same setting and conditions as in Proposition 1. Let* $\varepsilon > 0$ *be sufficiently small and* n *sufficiently large such that* $0 < \varepsilon < (2n)^{-1}$. *Then for any* $J \in \{J_{n1}, \cdots, J_{nK_0}\}$ *given in* $\mathcal{M}_{0n}(\varepsilon)$, *we have*

(R.4) $Pr(V_J = 1) \geq Pr(\mathcal{M}_n(\varepsilon)) \sim [1 + (2^{K_0} - 1)e^{-\xi n}]^{-1}$ *a.s..*

(R.5) $Pr(V_J = 0) \leq Pr(\mathcal{M}_n^c(\varepsilon)) \asymp (2^{K_0} - 1)e^{-\xi n}[1 + (2^{K_0} - 1)e^{-\xi n}]^{-1}$ *a.s..*

(R.6) $\dfrac{Pr(V_J = 1)}{Pr(V_J = 0)} \geq (2^{K_0} - 1)^{-1} e^{\xi n}$ *a.s..*

Proof of Proposition 2 is obvious from Proposition 1 and the definition of $\mathcal{M}_{0n}(\varepsilon)$. Proposition 2 essentially says that, under PBIC, PemBIC and PtBIC, the marginal probability of each true change-point J_{ni}, $i = 1, \cdots, K_0$, being detected is exponentially larger than that of it not being detected, provided that K_0 is fixed and n is sufficiently large. Therefore, although we cannot guarantee the true change-points $J_{n1}, \cdots, J_{nK_0}$ to be exclusively detected with large probability as implied by (R.3), we can guarantee that each J_{ni} is much more likely to be correctly detected than its being incorrectly ignored. This provides an asymptotic justification for Algorithms 1 and 2 where we use marginal empirical probabilities to identify change-points.

3.3 More Remarks on Applying Algorithms 1 and 2

In general it is important to monitor the convergence of the Markov chain generated by an MCMC algorithm. In particular, one needs to determine the length of a burn-in period so that the Markov chain generated after the burn-in period can be safely regarded as becoming stationary and be used for making inference. Many graphic and numeric diagnostic methods have been developed for dealing with this issue in literature.

However, determining the burn-in sequence to be removed is not critical on applying Algorithms 1 and 2 for detecting change-points. This is because we use either the minimizer of the generated sequence of the criterion values (emBIC or tBIC) or the marginal empirical probabilities ($p_i^{(L)}$ or $\breve{p}_i^{(L)}$, $i = 1, \cdots, n-1$) against p^* to determine the optimal change-points configuration. In the former case, removing the initial burn-in sequence reduces the search space, hence does not improve on finding a better minimizer. In the latter case, an upper bound of the standard error of $p_i^{(L)}$ or $\breve{p}_i^{(L)}$ is $(2\sqrt{L})^{-1}$, which is not affected by the actual values generated in the Markov chain. The effect of ignoring the burn-in becomes negligible when L is set moderate and p^* is set neither too close to 0 nor too close to 1. Nevertheless, we set the length of burn-in period to be $L_0 = 5$ or a similar small number to remove any adverse initialization effect of $\mathbf{V}_n^{(0)}$. Our simulation study indicates that this small L_0 is sufficient.

Clearly, it is still important to specify a proper number for L, the number of samples to be generated in Algorithms 1 and 2. As $n-1$ emBIC or tBIC values need to be evaluated in generating each $\mathbf{V}_n^{(\ell)}$, $\ell = 1, \cdots, L$, it is computationally very intensive if both n and L

are large. But a larger L definitely will give a more reliable result. On the other hand, we will suggest running Algorithm 1 or 2 twice in detecting the true change-points configuration. In the first run we want to identify those V_i's not being the true change-points, and remove them from being generated again in the sequel. In the second run we want to identify the true change-points as accurately as possible from those V_i's left from the first run. This suggests using a relatively small L in the first run and a relatively large L in the second run. We propose $L = 50$ and $p^* = 0.15$ in the first run, and $L = 100$ and $p^* = 0.5$ in the second run.

The following calculations may help on understanding these choices. Let $p_i = \Pr(V_i = 1)$ and $T_i^{(L)}$ be the number of times V_i is generated to be 1 in L i.i.d. Bernoulli trials. Then $\Pr(T_i^{(50)} \leq 50 \times 0.15 = 7.5) < 1.05 \times 10^{-7}$ if $p_i \geq 0.5$, and $< 2.42 \times 10^{-11}$ if $p_i \geq 0.6$. Also $\Pr(T_i^{(50)} > 50 \times 0.15 = 7.5) < 3.69 \times 10^{-8}$ if $p_i \leq 0.05$. So many V_i's not being the true change-points are very likely to be correctly excluded, and any true change-point is very unlikely to be wrongly excluded after the first run when set $L = 50$ and $p^* = 0.15$. On the other hand, $\Pr(T_i^{(100)} \leq 100 \times 0.5 = 50) < 2.21 \times 10^{-5}$ if $p_i \geq 0.7$, and $< 2.14 \times 10^{-11}$ if $p_i \geq 0.8$; $\Pr(T_i^{(100)} > 100 \times 0.5 = 50) < 9.04 \times 10^{-6}$ if $p_i \leq 0.3$, and $< 5.18 \times 10^{-12}$ if $p_i \leq 0.2$. So any V_i being a true change-point is very unlikely to be undetected, any remaining V_i not being a true change-point is very unlikely to be selected as a change-point when $L = 100$ and $p^* = 0.5$ are set in the second run.

The turning parameter $\tau > 0$ in both $\text{PemBIC}_{\tau\gamma}$ and $\text{PtBIC}_{\tau\alpha}$ is used to adjust the number of distinct change-points configurations to be generated by the Gibbs sampler. If τ is set small, the $\text{emBIC}_{\tau\gamma}$ (or $\text{tBIC}_{\tau\alpha}$) values sequence being generated may be very slow to progress into the neighbourhood of the minimum $\text{emBIC}_{\tau\gamma}$ (or $\text{tBIC}_{\tau\alpha}$) value. If τ is set large, the $\text{emBIC}_{\tau\gamma}$ (or $\text{tBIC}_{\tau\alpha}$) values sequence may bypass the minimum $\text{emBIC}_{\tau\gamma}$ (or $\text{tBIC}_{\tau\alpha}$) too often to ever reach it. We suggest to set τ to such a value that the generated L configurations contain roughly $0.3L$ but not smaller than $0.05L$ distinct configurations. In the simulation study we have done, we set $\tau = 1$ and have not encountered any situations where adjusting τ is required.

The hyperparameter γ adjusts the penalty in emBIC_γ which becomes BIC when $\gamma = 0$. Asymptotic study shows that the log-likelihood term, i.e. the first term in emBIC_γ (3.6) plays a dominant role in including true change-points in the selection by emBIC, while the penalty terms in (3.6) have a dominant role in excluding redundant change-points from the selection. A

proper data-adaptive specification of the γ value is therefore important in the latter case. Note that the value reduction of the first term in (3.6) needs to be smaller than the value increase of the other terms in (3.6) when preventing a redundant change-point from being selected. Our empirical study suggests this can be effectively achieved by setting $\gamma = q_\nu \log\log n$ where q_ν is the level ν sample quantile of certain standardized data sequence deemed not containing any change-point. Such standardized sequence can be constructed in the following way: First, execute the first run of Algorithm 1 with $\gamma = 0$ or say $\gamma = 2$ to produce a partition of the original data by the estimated change-points (which are likely to contain redundant ones but unlikely to miss the true ones). Second, each segment in the partition is standardized by subtracting its sample mean and being divided by its sample standard deviation, i.e., the z-scores of each segment are calculated. The absolute values of the z-scores from all segments give the referred standardized sequence. Common values of level ν are 0.90, 0.95 and 0.99.

The significance level α controls the penalty part in tBIC_α through the critical value $c_{n\alpha}$ given by (3.15) and (3.16). The commonly chosen values of α are 0.1, 0.05 and 0.01 by [Chen and Gupta, 1997]. We suggest using $\alpha = 0.1$ in the first run of Algorithm 2 and use $\alpha = 0.05$ in the second run of Algorithm 2. Note our $c_{n\alpha}$ is of half size of c_α given by [Chen and Gupta, 1997, eq.(8)] because our BIC is half of their SIC.

A computing complication may occur occasionally when using Gibbs sampler to update certain V_i from $V_i^{(\ell-1)}$ to $V_i^{(\ell)}$. For example, in determining mean-variance change-points, occurrence of $V_i^{(\ell)} = V_{i-1}^{(\ell)} = 1$ will result in zero estimate of σ_i^2 and the criterion value being $-\infty$ (3.11) and (3.12)). Then a modification on the computing procedure is necessary to avoid this complication. In the case of mean-variance change-points, we estimate σ_i^2 based on Y_i and its closest neighboring observations where not all the associated $V_i^{(\ell)}$ values equal 1.

By running Algorithm 1 or 2, not only we obtain an empirical distribution for each V_i of the latent indicator process $(V_1, \cdots, V_{n-1})$, from which we obtain an optimal estimate of the change-points, we can also find the empirical distribution of the number of change-points and obtain an optimal estimate of it. This is done, for example, by computing $K_n^{(\ell)} = \sum_{i=1}^{n-1} V_i^{(\ell)}$, $\ell = 1, \cdots, L$, tallying $(K_n^{(1)}, \cdots, K_n^{(L)})$, and finding the associated mode.

A post-selection inference may be made to validate the change-points detected by Algorithms 1 and 2 to improve the selection performance. This is done by performing a BIC

induced test of [Chen and Gupta, 1997], in each data segment spanned by 3 consecutive change-points already detected, to validate/update the middle change-point. Our simulation study suggests such post-selection calibration can improve the selection precision considerably.

Considering all the remarks in this section, we propose the following three-step procedures to implement Algorithms 1 and 2.

Procedure 1. *Three-step implementation of Gibbs sampler + emBIC*

- **Step 1.**

  1-1. By default set $L_0=5; L=50; \tau=1; \gamma=0$ or 2; $\nu=0.90$, 0.95 or 0.99; and $p^*=0.15$.

  1-2. Apply the Gibbs sampler part of Algorithm 1 to generate L_0+L candidate change-points configurations; then remove the first L_0 ones. The remaining L configurations are denoted as $\{\mathbf{V}_n^{(1)}, \cdots, \mathbf{V}_n^{(L)}\}$.

  1-3. Use the emBIC part of Algorithm 1 to determine the optimal change-points configuration from $\{\mathbf{V}_n^{(1)}, \cdots, \mathbf{V}_n^{(L)}\}$, denoted as $\mathbf{V}_n^{*(L)} = (V_1^{*(L)}, \cdots, V_{n-1}^{*(L)})$, or equivalently $(j_1^{*(L)}, \cdots, j_{K_n^{*(L)}}^{*(L)}; K_n^{*(L)})$. That is, $K_n^{*(L)} = \sum_{i=1}^{n-1} V_i^{*(L)}$; and $V_i^{*(L)} = 1$ if $i \in \{j_1^{*(L)}, \cdots, j_{K_n^{*(L)}}^{*(L)}\}$ and $V_i^{*(L)} = 0$ if $i \notin \{j_1^{*(L)}, \cdots, j_{K_n^{*(L)}}^{*(L)}\}$, $i = 1, \cdots, n-1$.

  1-4. Compute the sample quantile q_ν from the absolute z-scores described above.

- **Step 2.**

  2-1. By default set $L_0 = 0; L = 100; \tau = 1; \gamma = q_\nu \log\log n$; and $p^* = 0.5$.

  2-2. Let $\mathbf{W} = (W_1, \cdots, W_{K_n^{*(L)}}) = (V_i : i = j_1^{*(L)}, \cdots, j_{K_n^{*(L)}}^{*(L)})$ be the subset of the latent indicator process $(V_1, \cdots, V_{n-1})$ determined by the optimally estimated change-points configuration obtained from Step 1.

  2-3. Define a new probability mass function $\text{PemBIC}'_{\tau\gamma}(\mathbf{W})$ on the subspace spanned by $\mathbf{W}$, using the original data $Y_1, \cdots, Y_n$. Then apply Algorithm 1 to $\text{PemBIC}'_{\tau\gamma}(\mathbf{W})$ with the given L_0, L, τ, γ and p^*. The resultant optimal change-points estimates $(W_1^{*(L)}, \cdots, W_{K_n^{*(L)}}^{*(L)})$ can be equivalently written as $(\hat{j}_{n1}^{(L)}, \cdots, \hat{j}_{n\hat{K}_n^{(L)}}^{(L)}; \hat{K}_n^{(L)})$, where $\{\hat{j}_{n1}^{(L)}, \cdots, \hat{j}_{n\hat{K}_n^{(L)}}^{(L)}\}$ must be a subset of $\{j_1^{*(L)}, \cdots, j_{K_n^{*(L)}}^{*(L)}\}$ and $\hat{K}_n^{(L)} \le K_n^{*(L)}$.

- **Step 3. Post-selection Calibrations.** The statistical significance of each detected change-point $\hat{j}_{nk}^{(L)}$, $k = 1, \cdots, \hat{K}_n^{(L)}$, is assessed by testing "H_0: no change-point in $\widehat{Y}_{nk} = (Y_{\hat{j}_{n(k-1)}^{(L)}+1}, \cdots, Y_{\hat{j}_{n(k+1)}^{(L)}})$" versus "$H_1$: one change-point exists in $\widehat{Y}_{nk}$". Specifically, the BIC induced test of [Chen and Gupta, 1997] with properly specified critical value $c_{n\alpha}$ is used for each test. The change-point $\hat{j}_{nk}^{(L)}$ is removed from the selection if H_0 is accepted. Otherwise, $\hat{j}_{nk}^{(L)}$ is replaced by the new estimate associated with this test.

Procedure 2. *Three-step implementation of Gibbs sampler + tBIC*

- **Step 1.**

  1-1. By default set $L_0 = 5, L = 50, \tau = 1, \alpha = 0.1$ and $p^* = 0.15$.

  1-2. Apply the Gibbs sampler part of Algorithm 2 to generate $L_0 + L$ candidate change-points configurations; then remove the first L_0 ones. The remaining L configurations are denoted as $\{\mathbf{V}_n^{(1)}, \cdots, \mathbf{V}_n^{(L)}\}$.

  1-3. Use the tBIC part of Algorithm 2 to determine the optimal change-points configuration from $\{\mathbf{V}_n^{(1)}, \cdots, \mathbf{V}_n^{(L)}\}$, denoted as $\mathbf{V}_n^{+(L)} = (V_1^{+(L)}, \cdots, V_{n-1}^{+(L)})$, or equivalently $(j_1^{+(L)}, \cdots, j_{K_n^{+(L)}}^{+(L)}; K_n^{+(L)})$. That is, $K_n^{+(L)} = \sum_{i=1}^{n-1} V_i^{+(L)}$; and $V_i^{+(L)} = 1$ if $i \in \{j_1^{+(L)}, \cdots, j_{K_n^{+(L)}}^{+(L)}\}$ and $V_i^{+(L)} = 0$ if $i \notin \{j_1^{+(L)}, \cdots, j_{K_n^{+(L)}}^{+(L)}\}$, $i = 1, \cdots, n-1$.

- **Step 2.**

  2-1. By default set $L_0 = 0, L = 100, \tau = 1, \alpha = 0.05$ and $p^* = 0.5$.

  2-2. Let $\mathbf{W} = (W_1, \cdots, W_{K_n^{+(L)}}) = (V_i : i = j_1^{+(L)}, \cdots, j_{K_n^{+(L)}}^{+(L)})$ be the subset of the latent indicator process $(V_1, \cdots, V_{n-1})$ determined by the optimally estimated change-points configuration obtained from Step 1.

  2-3. Define a new probability mass function $\mathrm{PtBIC}'_{\tau\alpha}(\mathbf{W})$ on the subspace spanned by $\mathbf{W}$ and based on the original data $Y_1, \cdots, Y_n$. Then apply Algorithm 2 to $\mathrm{PtBIC}'_{\tau\alpha}(\mathbf{W})$ with the given L_0, L, τ, α and p^*. The optimal change-points estimates $(W_1^{+(L)}, \cdots, W_{K_n^{+(L)}}^{+(L)})$ can be equivalently written as $(\breve{j}_{n1}^{(L)}, \cdots, \breve{j}_{n\breve{K}_n^{(L)}}^{(L)}; \breve{K}_n^{(L)})$, where $\{\breve{j}_{n1}^{(L)}, \cdots, \breve{j}_{n\breve{K}_n^{(L)}}^{(L)}\}$ must be a subset of $\{j_1^{+(L)}, \cdots, j_{K_n^{+(L)}}^{+(L)}\}$ and $\breve{K}_n^{(L)} \le K_n^{+(L)}$.

- **Step 3. Post-selection Calibrations.** Assess the statistical significance of each $\breve{j}_{nk}^{(L)}$, $k = 1, \cdots, \breve{K}_n^{(L)}$, and update the $\breve{j}_{nk}^{(L)}$ value in the same way as in Step 3 of Procedure 1.

Finally, note that the finite-sample results from Procedures 1 and 2 in theory vary with the randomness involved in the Gibbs sampling. But this variation mostly disappears after applying the post-selection calibrations step.

3.4 Simulation Study and Real Data Examples

3.4.1 Simulation Study

We assess the finite sample performance of the proposed emBIC+Gibbs sampling method for multiple change-points detection. The performance was also compared with that of tBIC+Gibbs which was developed based on SIC [Chen and Gupta, 1997] but different (equations (3.14) and (3.21)), and with that of MPC in section 3.1.2. We developed Matlab code for the two procedures given in section 3.3, which together with the available code from [Lavielle, 2005] for MPC was used in our simulation study.

A properly designed simulation setup is necessary in order for the simulation results to be informative. Six factors were considered in generating the data for our simulation study. These are the base probability distribution for the data being simulated, the sample size n, the number of true change-points K, the locations of the change-points, the sizes of the changes, and the number of simulation times for each case. It was sufficient to use 1000 for the number of simulation times.

For given n and K the data of $(Y_1, \cdots, Y_n)$ were so generated that they can be partitioned into $K+1$ equal consecutive segments with the K internal nodes being the change-points. We assume each Y_i is of the form $Y_i = \mu_j + \sigma_j Z_i$, $i = 1, \cdots, n$ and $j = 1, \cdots, K+1$, where the mean and standard deviation of Y_i equal μ_j and σ_j respectively if Y_i falls into the jth segment. Here each Z_i has mean 0 and standard deviation 1, and is obtained by i) generating a random number from one of the following base distributions: N(0,1), Exp(1) and Pareto(5,1); ii) standardizing the generated number by subtracting the mean and being divided by the standard deviation. Note the density function of a Pareto(α, β) distribution is $\alpha\beta^{\alpha}x^{-(\alpha+1)}I(x > \beta)$ with $\alpha > 0$ being the scale parameter and $\beta > 0$ being the location parameter. The three base distributions used provided a good representation of distributions in terms of symmetry, skewness and tail length.

Although we were able to vary values of the other four factors to obtain more simulation results, we decided to use the setup of these four factors as was used in [Lavielle, 2005], in order to properly control the length of the dissertation. Namely,

1. choose $n = 500$;
2. in cases of having change-points, set $K = 4$ and construct the data by generating 5 equal consecutive segments each containing 100 sample numbers from the same distribution;
3. in the case of changes in mean, set $\sigma_1 = \cdots = \sigma_5 = 1$ and the 5 segment means $(\mu_1, \cdots, \mu_5) = (0, 1, 0, 2, 0)$;
4. in cases of changes in variance, set $\mu_1 = \cdots = \mu_5 = 0$ and the 5 segment standard deviations $(\sigma_1, \cdots, \sigma_5) = (1, 2, 1, 3, 1)$;
5. in cases of changes in mean and variance, set $(\mu_1, \cdots, \mu_5) = (0, 1, 0, 2, 0)$ and $(\sigma_1, \cdots, \sigma_5) = (1, 2, 1, 3, 1)$, respectively.

The simulation results under the setup given above are summarized in Tables 6 to 11, where Tables 6 to 8 are for the cases of having no change-points in mean, variance and (mean, variance), respectively; and Tables 9 to 11 are for cases of having 4 change-points in mean, variance and (mean, variance), respectively. The tables contain the results from using 8 different computing procedures labeled by the Method column: emBIC1, emBIC3 and emBIC5 refer to the three-step implementation of Gibbs sampler + emBIC with $\nu = 0.90$, 0.95 and 0.99 respectively; emBIC2, emBIC4 and emBIC6 refer to using the first two steps of the above three-step implementation (i.e. no post-selection calibrations) with $\nu = 0.90$, 0.95 and 0.99 respectively; tBIC refers to the three-step implementation of Gibbs sampler + tBIC; and MPC refers to the procedure of [Lavielle, 2005]. In each table, each number in section K gives the number of times that a specified K change-points were detected in the 1000 simulations; and each number in the section "Correction Detection" gives the number of times the change-point at a specific location in the data (i.e. 100/500, 200/500, 300/500 or 400/500) was identified in the 1000 simulations. Note that a change-point location detected within a distance of 5 from j was regarded as the location being at j in our simulation study, in order to remove the over-expression of the variation in change-points locations.

Tables 6 to 11 inform the following findings at least:

1. Consider the results of Table 6 where changes in mean are the focus but no change-points in mean exists (i.e. $K = 0$). Here emBIC1, emBIC3, emBIC5 and tBIC have excellent performance with the minimum frequency of correct detection being 916. Procedures emBIC2, emBIC4 and emBIC6 mostly work very well except when the distribution is Pareto(5,1). MPC does not work well except when the distribution is N(0,1).

2. Consider the results of Table 7 where changes in variance are the focus but no change-points in variance exists (i.e. $K = 0$). Now MPC works well only in the case of N(0,1) distribution. The other 7 procedures work mostly very well in the cases of N(0,1) and Exp(1) except that emBIC2 and tBIC do not do well in the case of Exp(1). No procedure does well when the distribution is Pareto(5,1).

3. Consider the results of Table 8 where no change-points in (mean, variance) exists (i.e. $K = 0$). All 8 procedures work mostly very well in the cases of N(0,1) and Exp(1) except that tBIC does badly in the case of Exp(1). No procedure does well when the distribution is Pareto(5,1) except emBIC5 and emBIC6. That tBIC does poorly at Exp(1) is most likely because the insufficient penalty in tBIC produces false change-points from the first two steps that cannot be reconciled by the post-selection calibrations step.

4. Consider the results of Table 9 where each data sample has $K = 4$ mean change-points. Procedures emBIC1-4, tBIC and MPC work mostly very well in the cases of N(0,1) and Exp(1) except that MPC over-estimates the number of change-points. Procedures emBIC5-6 mostly under-estimate the number of change-points. No procedure does well when the distribution is Pareto(5,1) except emBIC1 and emBIC3 which work surprisingly well, most likely due to the post-selection calibrations.

5. Consider the results of Table 10 where each data sample has $K = 4$ variance change-points. All procedures work very well at N(0,1) except that emBIC5-6 tend to under-estimate and MPC tends to over-estimate. Procedure emBIC6 performs worse than emBIC5 in terms of frequency of correct detection. All procedures' performance is fair at Exp(1) with emBIC3-4 performing marginally better. No procedure performs well at Pareto(5,1).

6. Consider the results of Table 11 where each data sample has $K = 4$ (mean, variance)

Table 6: No Change-point in Mean

Test	Distribution	Method	K			
			0	1	2	≥ 3
MEAN	N(0,1)	emBIC1	1000	0	0	0
		emBIC2	1000	0	0	0
		emBIC3	1000	0	0	0
		emBIC4	1000	0	0	0
		emBIC5	1000	0	0	0
		emBIC6	1000	0	0	0
		tBIC	991	4	5	0
		MPC	744	10	105	141
	Exp(1)	emBIC1	977	9	14	0
		emBIC2	713	6	228	53
		emBIC3	992	6	2	0
		emBIC4	863	0	122	15
		emBIC5	997	1	2	0
		emBIC6	966	0	34	0
		tBIC	933	23	42	2
		MPC	574	233	96	97
	Pareto(5,1)	emBIC1	928	44	26	2
		emBIC2	305	6	348	341
		emBIC3	964	23	12	1
		emBIC4	551	6	317	126
		emBIC5	991	6	2	1
		emBIC6	841	1	137	21
		tBIC	916	45	37	2
		MPC	404	0	279	317

Table 7: No Change-point in Variance

Test	Distribution	Method	K			
			0	1	2	≥ 3
MEAN	N(0,1)	emBIC1	999	0	1	0
		emBIC2	993	0	7	0
		emBIC3	1000	0	0	0
		emBIC4	998	0	2	0
		emBIC5	1000	0	0	0
		emBIC6	1000	0	0	0
		tBIC	978	4	18	0
		MPC	739	12	115	134
	Exp(1)	emBIC1	604	204	141	51
		emBIC2	477	71	279	173
		emBIC3	819	128	45	8
		emBIC4	766	62	145	27
		emBIC5	952	40	7	1
		emBIC6	944	24	30	2
		tBIC	386	230	223	161
		MPC	508	39	173	280
	Pareto(5,1)	emBIC1	60	314	197	429
		emBIC2	24	14	94	868
		emBIC3	164	371	219	246
		emBIC4	137	47	233	583
		emBIC5	480	337	136	47
		emBIC6	470	88	278	164
		tBIC	37	271	170	522
		MPC	372	42	198	388

Table 8: No Change-point in Mean-and-Variance

Test	Distribution	Method	K			
			0	1	2	≥ 3
MEAN	N(0,1)	emBIC1	997	2	1	0
		emBIC2	965	1	33	1
		emBIC3	999	1	0	0
		emBIC4	993	1	6	0
		emBIC5	1000	0	0	0
		emBIC6	999	0	1	0
		tBIC	957	21	22	0
		MPC	792	4	108	96
	Exp(1)	emBIC1	765	119	91	25
		emBIC2	722	66	150	62
		emBIC3	927	55	18	0
		emBIC4	920	40	39	1
		emBIC5	987	13	0	0
		emBIC6	986	13	1	0
		tBIC	187	139	229	445
		MPC	635	46	133	186
	Pareto(5,1)	emBIC1	144	260	177	419
		emBIC2	121	47	146	686
		emBIC3	391	297	187	125
		emBIC4	376	92	257	275
		emBIC5	681	225	77	17
		emBIC6	675	95	184	46
		tBIC	16	158	81	745
		MPC	398	63	192	347

Table 9: Multiple Change-points in Mean

Distribution	Method	K									Correct Detection			
		0	1	2	3	4	5	6	7	≥ 8	1/5	2/5	3/5	4/5
N(0,1)	emBIC1	0	0	13	9	976	2	0	0	0	921	918	999	1000
	emBIC2	0	0	12	10	971	6	1	0	0	806	815	999	1000
	emBIC3	0	0	73	20	906	1	0	0	0	860	855	999	1000
	emBIC4	0	0	73	20	905	2	0	0	0	708	738	1000	1000
	emBIC5	0	0	378	27	595	0	0	0	0	576	574	997	1000
	emBIC6	0	0	378	27	595	0	0	0	0	419	475	999	1000
	tBIC	0	0	0	2	987	8	3	0	0	936	930	999	1000
	MPC	0	0	0	0	745	61	88	40	66	934	927	1000	1000
Exp(1)	emBIC1	0	0	22	14	993	22	9	0	0	905	907	1000	998
	emBIC2	0	0	18	7	600	106	176	28	65	761	796	1000	999
	emBIC3	0	0	174	33	787	6	0	0	0	754	754	1000	998
	emBIC4	0	0	171	25	627	77	17	11	0	584	595	999	1000
	emBIC5	5	0	796	16	183	0	0	0	0	176	175	995	993
	emBIC6	5	0	787	13	171	18	6	0	0	123	135	977	988
	tBIC	0	0	3	10	938	31	17	1	0	919	926	1000	997
	MPC	0	0	155	2	456	75	162	29	121	770	779	1000	999
Pareto(5,1)	emBIC1	0	4	26	44	877	41	7	1	0	879	901	996	994
	emBIC2	0	0	4	2	245	89	239	105	316	851	856	1000	1000
	emBIC3	0	4	112	55	811	15	3	0	0	791	816	996	993
	emBIC4	0	0	82	17	402	119	208	66	106	648	691	1000	1000
	emBIC5	23	4	725	42	206	0	0	0	0	205	218	975	970
	emBIC6	23	0	659	18	214	40	31	2	0	141	166	957	952
	tBIC	0	4	25	52	856	45	16	2	0	871	902	996	994
	MPC	15	0	269	16	273	77	150	43	157	543	554	981	982

Table 10: Multiple Change-points in Variance

Distribution	Method	K									Correct Detection			
		0	1	2	3	4	5	6	7	≥ 8	1/5	2/5	3/5	4/5
N(0,1)	emBIC1	0	0	5	12	976	5	2	0	0	908	915	990	994
	emBIC2	0	0	5	11	962	12	10	0	0	715	782	987	988
	emBIC3	0	0	32	34	933	1	0	0	0	866	879	990	994
	emBIC4	0	0	32	33	931	2	2	0	0	653	707	971	987
	emBIC5	0	0	172	83	745	0	0	0	0	706	746	984	994
	emBIC6	0	0	172	83	745	0	0	0	0	470	565	900	962
	tBIC	0	0	0	2	972	14	12	0	0	913	921	990	994
	MPC	0	0	1	0	742	70	99	25	63	915	919	990	991
Exp(1)	emBIC1	0	1	26	54	576	228	90	17	8	620	630	903	921
	emBIC2	0	0	18	32	454	195	180	61	60	608	571	902	926
	emBIC3	1	3	93	103	654	118	26	2	0	547	583	895	923
	emBIC4	1	0	84	84	604	133	77	12	5	501	523	860	917
	emBIC5	4	11	338	152	477	18	0	0	0	404	409	833	918
	emBIC6	4	9	328	145	478	30	6	0	0	290	330	735	838
	tBIC	0	1	4	32	446	284	156	60	17	642	650	895	921
	MPC	9	1	91	36	486	112	100	58	107	593	596	899	924
Pareto(5,1)	emBIC1	1	22	33	68	184	196	226	139	131	473	463	698	754
	emBIC2	0	0	1	8	48	71	128	114	370	562	547	824	853
	emBIC3	1	26	70	116	310	240	151	62	24	434	409	677	736
	emBIC4	0	1	21	38	189	161	204	140	246	466	472	760	801
	emBIC5	6	45	211	192	370	130	39	7	0	338	308	624	717
	emBIC6	5	8	166	124	381	145	103	39	29	313	332	636	712
	tBIC	1	24	26	50	148	160	215	171	215	486	476	701	758
	MPC	90	14	172	54	271	93	116	59	121	356	362	641	707

change-points. All procedures work very well at N(0,1) except that emBIC5-6 tend to under-estimate and MPC tends to over-estimate. Procedure emBIC6 performs worse than emBIC5 in terms of frequency of correct detection. All procedures' performance is fair at Exp(1) with emBIC3-4 performing marginally better; also emBIC5-6 tend to under-estimate and emBIC1-2, tBIC and MPC tend to over-estimate at Exp(1). No procedure performs well at Pareto(5,1) with low frequencies of $K = 4$; but all procedures except emBIC5 and emBIC6 mostly over-estimate the number of change-points.

7. Overall, the Gibbs sampler + emBIC based procedures and the Gibbs + tBIC one are competitive, with the former performing better more often than the latter. MPC seems somewhat less competitive than the other 7 procedures.

3.4.2 Example 1. Change-points in IBM Stock Historical Prices

IBM common stock daily closing prices from 17/May/1961 to 2/Nov/1962, which are recorded in [Box and Jenkins, 1976], are analyzed here for detecting possible historical volatility change-points. From an exploratory analysis it suggests the first differences of the logarithm of these prices, consisting of 368 numbers as displayed in Figure 3, be used for the detection. Using the default setting in Procedures 1 and 2 (i.e. $\gamma = 2$, $\nu = 0.95$ and $\alpha = 0.05$ etc.) we apply both Gibbs+emBIC and Gibbs+tBIC methods and find two variance change-points at locations 235 and 279; see Figure 3. Using MPC gives the same result.

Our result conforms to that of [Inclan and Tiao, 1994] who used their iterative cumulative sums of squares method. [Baufays and Rasson, 1985] found two variance changes at locations 235 and 280 using a maximum likelihood method. [Wichern *et al.*, 1976] assumed a piecewise ARIMA(1,1,0) model to the data and found two variance changes at 180 and 235. Point 180 does not seem to be a variance change-point based on Figure 3. [Tsay, 1988] assumed piecewise ARIMA(0,1,1) and detected a variance change only at point 237.

Assuming changes in both mean and variance are possible and using the default setting, Gibbs+emBIC finds two change-points also at 235 and 279, while Gibbs+tBIC finds only one change-point at 235. In contrast MPC finds more change-points at 95, 98, 204, 207, 235, 279.

Table 11: Multiple Change-points in Mean-and-Variance

Distribution	Method	K									Correct Detection			
		0	1	2	3	4	5	6	7	≥ 8	1/5	2/5	3/5	4/5
N(0,1)	emBIC1	0	0	12	14	972	2	0	0	0	939	946	998	999
	emBIC2	0	0	11	14	887	46	40	1	1	653	758	674	986
	emBIC3	0	0	59	45	894	2	0	0	0	869	893	996	999
	emBIC4	0	0	59	45	864	18	14	0	0	552	686	548	981
	emBIC5	0	1	380	73	546	0	0	0	0	562	565	986	999
	emBIC6	0	1	380	73	544	1	1	0	0	309	480	514	924
	tBIC	0	0	0	2	959	28	10	1	0	956	960	998	1000
	MPC	0	0	1	0	770	68	85	25	51	955	959	998	999
Exp(1)	emBIC1	0	0	19	38	640	196	85	15	7	658	671	919	921
	emBIC2	0	0	19	26	544	219	124	38	30	594	577	740	879
	emBIC3	0	2	128	100	683	73	12	2	0	590	585	886	919
	eBmIC4	0	2	127	93	660	90	21	6	1	448	488	596	824
	emBIC5	17	20	440	111	409	3	0	0	0	420	343	793	892
	emBIC6	17	19	440	110	409	4	1	0	0	286	274	454	767
	tBIC	0	0	0	4	192	203	232	158	211	894	905	993	992
	MPC	2	0	48	26	613	143	75	35	58	890	906	986	992
Pareto(5,1)	emBIC1	0	6	14	28	177	222	195	152	206	576	566	832	853
	emBIC2	0	0	4	9	99	100	158	142	488	610	579	784	870
	emBIC3	0	8	53	82	408	226	136	60	27	531	498	791	829
	emBIC4	0	1	39	55	339	184	173	105	104	476	486	638	794
	emBIC5	5	27	254	164	433	97	15	4	1	430	358	707	817
	emBIC6	5	16	228	137	430	123	46	7	8	337	310	463	711
	tBIC	0	6	6	13	38	52	103	141	641	607	626	847	852
	MPC	31	6	115	64	391	143	111	50	89	490	490	796	822

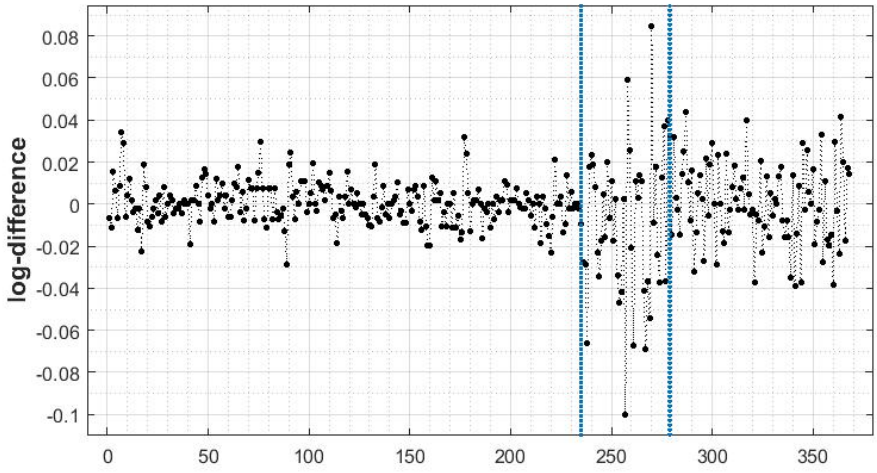

Figure 3: *HV* Change-points Found in the IBM Stock Price Data by Gibbs+emBIC

3.4.3 Example 2. Change-points in DNA Copy Number Data

Array DNA copy number data, resulted from array comparative genomic hybridization (CGH) studies of DNA sequences, can be analyzed by a change-points finding method which determines the locations where the underlying copy number logratio changes its mean. [Olshen *et al.* (2004)] developed a circular binary segmentation method for doing this and implemented it in an R-package called DNAcopy [Seshan and Olshen, 2011]. The data analyzed here correspond to the GM05296 array CGH study of fibroblast cell strains [Snijder *et al.*, 2001], and can be found in the DNAcopy package. After removing the missing values the GM05296 data contain 2112 values of logratio of copy numbers selected from chromosomes 1-22 and X.

Using the default setting Gibbs+emBIC finds 6 mean change-points: 114, 1127, 1168, 1251, 1266 and 2062 shown in Figure 4, while Gibbs+tBIC finds 8 mean change-points: 114, 1127, 1168, 1251, 1266, 1478, 1570 and 2062. MPC returns 5 mean change-points: 1128, 1168, 1252, 1266 and 2062, very close to those from Gibbs+emBIC. As a comparison, the circular binary segmentation method, implemented by the segment function in DNAcopy, returns 30 mean change-points at the default setting: 132, 196, 282, 425, 434, 447, 555, 640, 812, 963, 1074, 1127, 1131, 1168, 1200, 1251, 1266, 1385, 1479, 1536, 1612, 1678, 1744, 1835, 1888, 1925, 2012, 2030, 2045, 2061, far more than those found by the other methods.

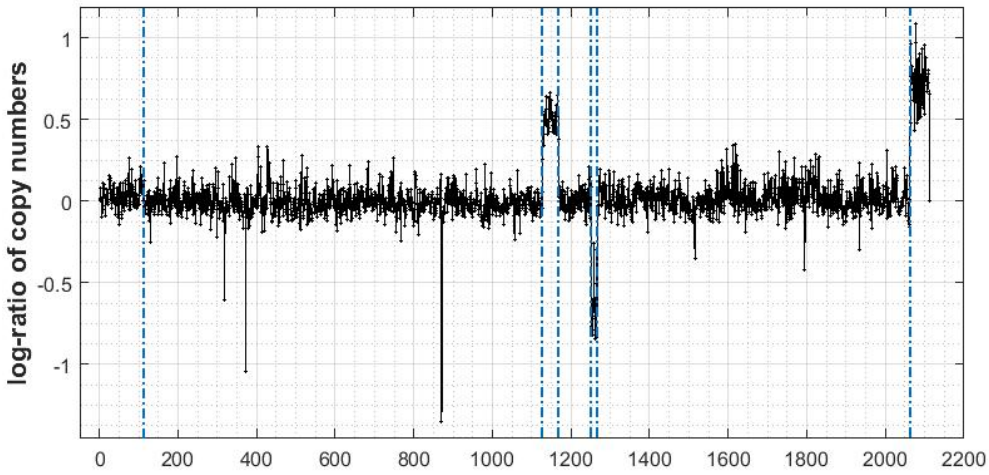

Figure 4: Mean Change-points Found in the DNA Copy Number Data by Gibbs+emBIC

3.5 Conclusions

We have developed a computationally efficient method for multiple change-points estimation. The new method combines an empirical Bayesian information assessment criterion with a Gibbs sampler induced stochastic search algorithm in an innovative and coherent way. The method has been shown to have both fine asymptotic properties and satisfactory finite-sample performance; and has been implemented by a comprehensive computing procedure ready for use in practice. The ideas used in developing our Gibbs sampler + emBIC method has also enabled us to extend a recent change-points testing method, which combines a binary segmentation procedure with the Schwarz information criterion [Chen and Gupta, 1997], to produce the Gibbs sampler + tBIC method. The method of Gibbs + tBIC performs similarly to the Gibbs + emBIC but is still more likely to over-estimate than the latter. Using the Gibbs sampler induced stochastic search greatly reduces the computing load required and is the primary reason for both Gibbs + emBIC and Gibbs + tBIC to have satisfactory performance.

4 Long Term Implied Volatility Behaviour Analysis

The purpose of this chapter is mainly to analyze the long term implied volatility (IV) behaviour of S&P500 index option. We discuss the interest rate risk's impact on the long term asset IV in stock market. We also analyze other factors which would affect long term asset IV.

To capture interest rate risk, we use the widely adopted short interest rate model. However, there is a challenge here. In general, we treat one-month interest rate as short rate to calibrate model, but one-month interest rates have been close to 0 since financial crisis in 2008. Even one-year interest rates are also very low. Under the condition of such low interest rates, no short interest rate models would work well. In order to apply short interest rate model here, we introduce time series change-point concept. We use the change-point detection technique we introduced in chapter 3 to find the optimal time range, then we calibrate the CIR model[Cox, Ingersolla and Ross, 1985] by the market data in such optimal time range. After obtaining all parameters in the CIR model, we put all of them into the BS-CIR model to reproduce the call option price for the S&P500 index. After that, we add capital charge costs to the final option price and then reproduce all terms of IVs by the BS model. Finally, we obtain that the long term IV with incorporated interest rate risk and capital charge does replicate the observed shape of the long term IV from market, i.e. IV increases as term increases. When capital charge is added, we can see the IV from our working model levels up and closely match the observed IV term structure. This analysis indicates further that HV gives us past information about the equity, while IV shows present information. The big gap between them tells us that market IV contains information about not only equity itself, but also other information about market like interest rate risk and capital risk charge.

4.1 Model Frame

4.1.1 European Call Option

The European call option price can be expressed as:

$$C = E[(S_T - K, 0)^+]$$

where S_T is the price of asset at time T and K is the strike price.

By BS model (See Appendix A-2 for details), the call option price can be calculated by

$$C = S_0 N(d_1) - K e^{-rT} N(d_2)$$

where S_0 is the asset price at time 0, r is the risk free interest rate, T is the expired time, $N(\cdot)$ is the cdf of the standard Normal distribution,

$$d_1 = \frac{\log(S_0/K) + (r + \sigma^2/2)/T}{\sigma\sqrt{T}} \quad \text{and}$$

$$d_2 = d_1 - \sigma\sqrt{T},$$

where σ is the volatility of an asset.

For a call option, if $S_0 = K$, it is called at-the-money (ATM) option, if $S_0 > K$, it is called in-the-money (ITM) option and otherwise, it is called out-of-the-money (OTM) option. If one uses the quoted implied volatility, one should be able to obtain the option price by BS model. In this chapter, we only study the ATM IV behaviour. The study presented here is applicable for either ITM or OTM option IV.

4.1.2 CIR Model

All short interest rate models can be written as

$$dr_t = \mu_r(t)dt + v_r(t)dB,$$

where r_t is the interest rate, $\mu_r(t)$ is the drift term, and $v_r(t)$ is the volatility term, B is the Brownian motion.

There are some popular interest rate models like the Vassicek model [Vasicek, 1977], the Hull and White (HW) model [Hull and White, 1993] and the CIR model [Cox, Ingersolla and Ross, 1985], etc. We choose the CIR model as our working model, since it can always guarantee interest rate to be positive if some conditions hold and its closed form solution makes the calibration straightforward. It has some weaknesses like it can never reproduce some very special yield curve no matter how to adjust the parameters and it cannot reproduce the yield

curve precisely. However, we discuss the long term volatility behaviour and we don't need to obtain very precise option prices.

The CIR model is

$$dr_t = \kappa(\theta - r_t)dt + \sigma_r\sqrt{r_t}dB,$$

where θ is the long run average interest rate, κ is the reverting rate at r_t to θ, σ_r is the volatility of the short interest rate r_t.

[Rebonato, 1998] introduces one calibration method for the CIR model:

$$\min(\text{LSDIF}) = \sum_i [P_{obs,i} - P_{mod,i}(\phi_1, \phi_2, \phi_3)].$$

where $P_{obs,i}$ is the bond price on the market with expired time T_i, and $P_{mod,i}$ is the bond price by the CIR model with the same expired time T_i, and

$$\phi_1 = \sqrt{(\kappa + \lambda)^2 + 2\sigma_r^2}, \tag{4.1}$$

$$\phi_2 = (\kappa + \lambda + \phi_1)/2, \tag{4.2}$$

$$\phi_3 = 2\kappa\theta/\sigma_r^2, \tag{4.3}$$

where λ is market price of risk.

The volatility term σ_r can be obtained directly by (4.1) and (4.2), i.e.,

$$\sigma_r = \sqrt{2(\phi_1 - \phi_2)\phi_2} \tag{4.4}$$

By far we can obtain ϕ_1, ϕ_2 and ϕ_3. It is easy to see that as long as we obtain any one of κ, θ and λ, we can obtain others by (4.1), (4.2) and (4.3).

[Torosantucci *et al.*, 2007] introduces an empirical evaluation method to estimate parameters in the CIR model. In their paper, they introduced dynamic implementation of the CIR model process to estimate CIR parameters by resorting to the time series of interest rates as a proxy for the short rate. We follow their idea but use [Kladivko, 2007]'s maximum likelihood method (See Appendix A-3 for details). [Torosantucci *et al.*, 2007] also proved that for time step parameter $\Delta = 1/250$, at least 550 daily data are necessary. Furthermore, they mentioned that using a large set of data can obtain reliable estimates. We implement the log-likelihood function in Matlab by using the command *ncx2pdf* to estimate κ. Once we obtain κ, we put it into (4.1), (4.2) and (4.3) to get other parameters.

4.1.3 Change-point Detection in Historical Volatility to Find Optimal Time Range

We calibrate the CIR model by using historical interest rates. Figure 5 shows 1-month, 1-year, 5-year, 10-year and 30-year term interest rates from Jan 2, 1990 to Mar 27, 2014. We can see the interest rates fluctuate largely in those years. If we calibrate CIR model with different time ranges, parameters in CIR model could vary a lot. Therefore, a proper time range is critical to do the CIR calibration.

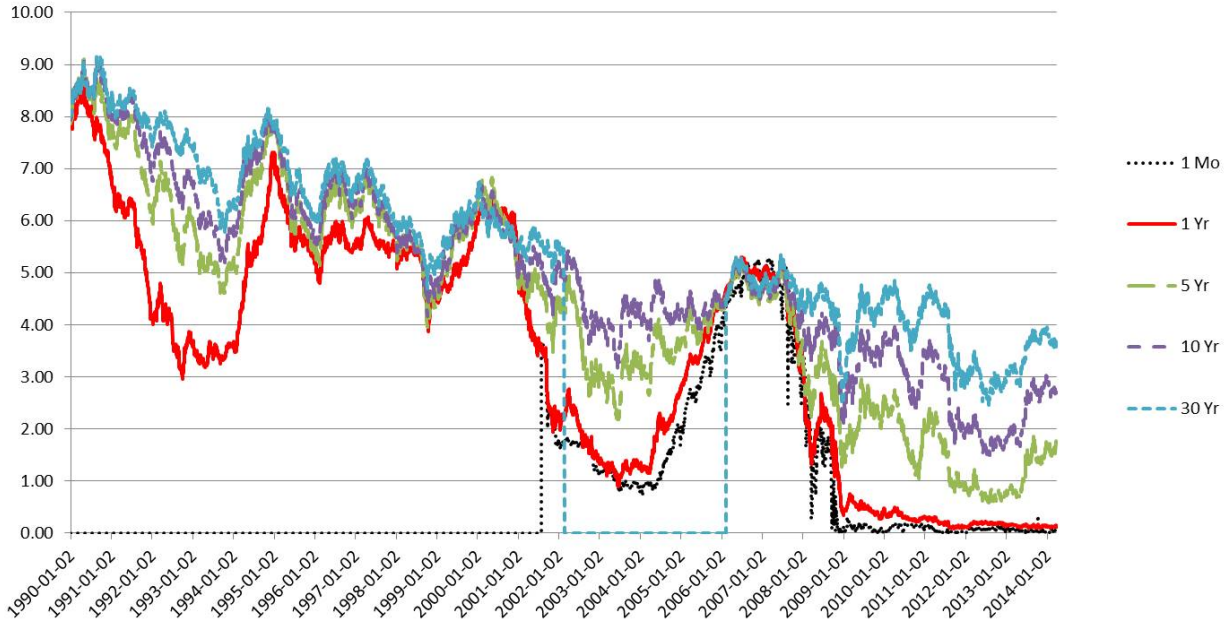

Figure 5: Historical Interest Rate

There are two ways to choose the time range. One way is choosing the whole time range from 1990 to 2008. The second way is choosing the time range which has similar situation as the time range we are interested in. Here we use change-points detection method emBIC introduced in chapter 3 to find change-points in the HV of 1-year interest rate. After finding change-points, we choose the time range in which the situation is closest to the situation we are interested in. Finally, we calibrate the CIR model by this optimal time range data. After comparing the results between two ways, we can see the second one has much better results.

4.1.4 BS-CIR Model

The BS model with the stochastic interest rate model can be expressed as:

$$\begin{aligned} dS_t &= r_t S_t dt + \sigma_s S_t dB_1, \\ dr_t &= \mu_r(t)dt + v_r(t)dB_2, \\ dB_1 dB_2 &= \rho dt, \end{aligned}$$

where S_t is the asset price, r_t is the interest rate, σ_s is the asset volatility, dB_1 and dB_2 are both Brownian motions with correlation ρ, and $\mu_r(t)dt$ and $v_r(t)$ are based on a specific interest rate model.

When BS and CIR model are combined, it is called Black-Scholes-CIR (BS-CIR) model. [Kim, 2002] derived the closed form solution for the European call option price by the BS-CIR model. Since the CIR model guarantees positive interest rates when some conditions hold, and we can observe really low short interest rate, around 0, in the market, we think the BS-CIR model is appropriate here.

The BS-CIR model can be expressed as

$$\begin{aligned} dS_t &= r_t S_t dt + \sigma_s S_t dB_1 \\ dr_t &= \kappa(\theta - r_t)dt + \sigma_r \sqrt{r_t} dB_2 \\ dB_1 dB_2 &= \rho dt, \end{aligned}$$

where θ is the long run average short interest rate, κ is the reverting rate at r_t to θ, σ_r is the volatility of the short interest rate r_t.

[Kim, 2002] derived the closed form of the European call option price for the BS-CIR model:

$$\begin{aligned} \text{C(BS-CIR)} &= E[(S_T - K)^+] \\ &= S_0 N(d_1) - K \exp(-\int_0^T r_t^* dt) N(d_2) \\ &\quad + \sigma_r C_0 [S_0 \phi(d_1) - K \exp(-\int_0^T r_t^* dt)(\phi(d_2) - \sigma_s \sqrt{T} N(d_2))] \\ &\quad + \sigma_r C_1 [d_2 S_0 \phi(d_1) - d_1 K \exp(-\int_0^T r_t^* dt)\phi(d_2)] + o(\sigma r), \end{aligned}$$

where $N(\cdot)$ is the cdf of standard Normal distribution, $\phi(\cdot)$ is the pdf of standard Normal distribution, and

$$\begin{aligned} r_t^* &= r_0 e^{-\kappa t} + \theta(1 - e^{-\kappa t}) \\ d_1 &= \frac{1}{\sigma_s \sqrt{T}}[\log(\frac{S_0}{K}) + \frac{r_0 - \theta}{\kappa}(1 - e^{-\kappa T}) + (\theta + \frac{\sigma_s^2}{2})T] \\ d_2 &= d_1 - \sigma_s \sqrt{T} \end{aligned}$$

and

$$\begin{aligned} C_0 &= \frac{1}{\sigma_s \sqrt{T}} \left[\frac{\lambda(r_0 - \theta)}{\kappa}(\frac{1 - e^{-\kappa T}}{\kappa} - T e^{-\kappa T}) + \frac{\lambda \theta T}{\kappa}(1 - \frac{1 - e^{-\kappa T}}{\kappa}) \right], \\ C_1 &= -\frac{\rho}{\sigma_s T} C_{11}, \end{aligned}$$

where r_0 is the risk free interest rate,

$$\begin{aligned} C_{11} &= \frac{2\sqrt{\theta}((1 + 2e^{\kappa T})\sqrt{r_0} - 3e^{\kappa T/2}\sqrt{r_0 - \theta(1 - e^{\kappa T})}) + (\theta(1 + 2e^{\kappa T}) - r_0)\psi}{2e^{\kappa T}\kappa^2\sqrt{\theta}}, \\ \psi &= \log \left[\frac{\theta(2e^{\kappa T} - 1) + r_0 + 2e^{\kappa T/2}\sqrt{\theta^2(e^{\kappa T} - 1) + \theta r_0}}{(\sqrt{r_0} + \sqrt{\theta})^2} \right] \end{aligned}$$

The only unknown parameter ρ can be estimated by the method proposed on the paper of [Kim, 2002]. Once all parameters are obtained, option prices can be reproduced with different terms. Furthermore, once we have option prices, corresponding IVs are easily obtained.

4.1.5 Capital Charge for Index Options

In 2010, the Basel Committee on Banking Supervision (BCBS) released the Basel III framework, which contains global regulatory standards on banks' capital requirement. The rules of capital adequacy require banks to set special deposits aside against the market risk. Market risk is the risk of losses in on- and off-balance sheet positions arising from movements in market prices.

For index options, there are two components of capital charges:

- Specific risk charge: the risk of loss caused by an adverse price movement of a debt instrument or security due principally to factors related to the issuer.

The specific risk charge for an option based on an index of equities is calculated by multiplying the market value of the equity index by 2%. For example, if the S&P500 index is 1500, then the specific risk charge for the index call option is $1500 \times 2\% = 30$.

- General market risk charge: the risk of loss arising from adverse changes in market prices.

  For the general market risk charge, we just introduce the scenario approach here, since financial institutions writing options must use the scenario method.

  The scenario method is introduced in Basel III framework: The scenario method uses simulation techniques to calculate changes in the value of an option's portfolio for changes in the level and volatility of the prices of its associated underlying instruments. Under this approach, the general market risk charge is determined by the scenario "matrix" that produces the largest loss. The first dimension of each matrix requires the institution to evaluate the portfolio over a specified range above and below the current value of the underlying instrument, commodity, or index. The range for index option is $\pm 8\%$. The second dimension of the matrix entails a change in the volatility of the underlying price equal to $\pm 25\%$ of the current volatility. See Table 12 in details.

Table 12: Scenario matrix

Index	-8%	-5.33%	-2.67%	Current Value	$+2.67\%$	$+5.33\%$	$+8\%$
-25%	gain/loss	gain/loss	gain/loss	gain/loss	gain/loss	gain/loss	gain/loss
Current Value	gain/loss	gain/loss	gain/loss	0	gain/loss	gain/loss	gain/loss
$+25\%$	gain/loss	gain/loss	gain/loss	gain/loss	gain/loss	gain/loss	gain/loss

The general market risk charge for index options should be calculated together with the associated hedging positions. The associated hedging position totally depends on the hedging strategy. In theory, the BS delta continuous hedging strategy with proportional transaction costs has infinite costs. Even in discrete models, transaction costs for hedging are substantial ([Figlewski, 1989]). [Soner *et al.*, 1994] pointed out that using the trivial strategy of buying one share of the underlying stock and holding to maturity is the least expensive method of writing a European call in BS model with proportional transaction costs. Here we use the hedging strategy above, i.e., we sell a call index option and hold

an index product. We change the index value and equity volatility according to the scenario matrix. Finally we can obtain the gain/loss matrix.

After considering the capital charge, the options price can be written as:

$$\begin{aligned} \text{Market option price with market } IV &= \text{Option price with model based} \\ &+ \text{Potential loss rate} \times \text{Total capital charge} \end{aligned}$$

It means the market IV includes not only the information about equity itself, but also other information like capital charges for the market risk.

4.2 Real Data Analysis

We collect market data from Apr 11, 2011 to Dec 13, 2011. They include 168 days of closed S&P500 index, ATM implied volatility of S&P500 index options from Barclays Capital Incorporated, and U.S. Daily Treasury Yield Curve Data from U.S. Department of the Treasury's Data Center. We also collect S&P500 index from Jan 2, 1980 to Mar 27, 2014. The terms of ATM IV of S&P500 index options are from 3 month to 15 years.

Figure 6 shows the marker IV surface from Apr 11, 2011 to Dec 13, 2011. It is easy to see there are 2 types of shapes: upward sloping curve (from Apr 11, 2011 to Aug 4, 2011) and concave curve (from Apr 11, 2011 to Dec 8, 2011). No matter what shape it has, IV rises as the maturity is longer than 7 years. Here we are going to explain this phenomenon.

4.2.1 CIR Model Parameters Estimation

Figure 5 shows interest rates after point 4715 (Sep 12, 2008) is much lower than those before point 4715. Also, we find that data between points 2800 and 4000 are the closest to the data after 4715.

We transform 1-year interest rate by the first order difference of logarithm. After that, we run emBIC to detect change-points in HV between Jan 2, 1990 and Sep 12, 2008. We totally find 12 change-points, 496, 803, 1715, 2187, 2239, 2772, 2948, 3701, 3983, 4427, 4549 and 4593. We choose time range from 2949 to 3701 (from Aug 31, 2001 to Sep 2, 2004) as

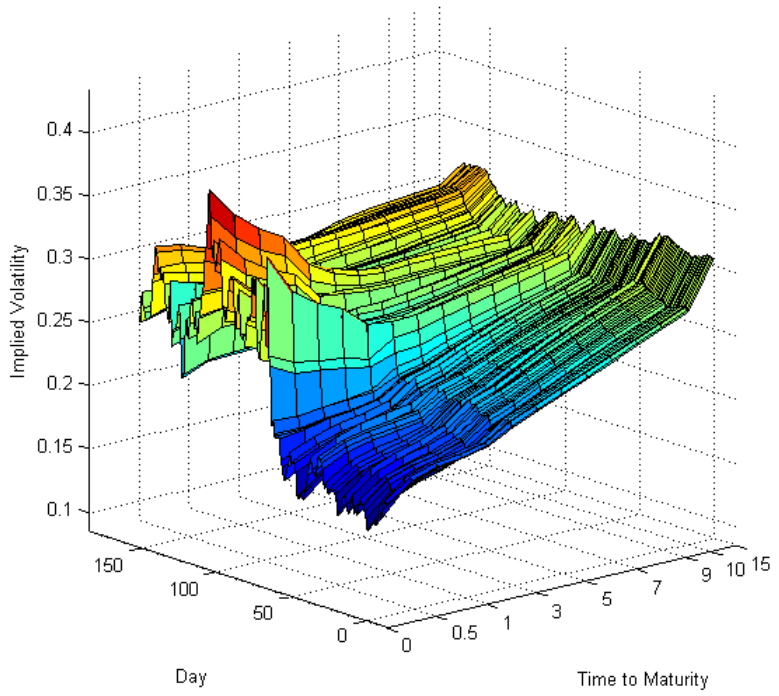

Figure 6: Market IV Surface

our optimal time range to estimate the parameters in the CIR model since they have closest behaviour to the interest rate after Sep 12, 2008.

We set up 2 scenarios:

- S1: All parameters in the CIR model are estimated based on interest rates from Jan 2, 1990 - Sep 12, 2008.
- S2: All parameters in the CIR model are estimated based on interest rates from Aug 31, 2001 to Sep 2, 2004.

Table 13 shows the average interest rates for 2 scenarios. We don't consider the interest rates after Sep 12, 2008 since the financial crisis makes the interest rate extremely low.

After calibrating the CIR model with data in Table 13, we obtain ϕ_1, ϕ_2 and ϕ_3 for 2

Table 13: Average Interest Rate Under Different Scenarios (%)

Scenarios	1/12 Y	1/4 Y	1/2 Y	1 Y	2 Y	3 Y	5 Y	7 Y	10 Y	20 Y	30 Y
S1	2.57	4.13	4.29	4.43	4.78	4.98	5.32	5.57	5.72	5.78	6.42
S2	1.34	1.37	1.46	1.71	2.24	2.70	3.47	3.97	4.36	5.21	5.36

scenarios in Table 14. The volatility of the interest rate σ_r can be obtained by (4.4). The relatively small σ_r implies the long run interest rate is very stable.

Table 14: CIR Parameters Estimated I

Scenarios	ϕ_1	ϕ_2	ϕ_3	σ_r
S1	1.7650	1.7644	93.7544	0.0467
S2	1.5346	1.5343	142.7318	0.0320

Next step, we estimate κ by dynamic implementation of the CIR model process introduced by [Torosantucci *et al.*, 2007].

The only question here is how to choose the time series of the short rate. The common choice for the short rate is the shortest maturity rate in the dataset. However, after year 2008 financial crisis, the 1-month interest rate is almost 0. Even 1 year rate is much lower than 1%. Under this kind of situation, we can choose 1-month interest rate from July 31, 2001 to Dec 31, 2007 (1-month interest rate before July 31, 2001 cannot be found in U.S. Daily Treasury Yield Curve Data Center). Totally there are 1613 short interest rates (see Figure 7 (a)).

The whole data set almost makes up a perfect interest rate circle, middle - low - middle - high - middle. Noticeably, data between 1517 and 1524 (between Aug 15 and Aug 24, 2007) has big fluctuation, which makes the estimated κ much larger than others. Also there was a big drop after 29 (Sep 10, 2001). Therefore we collect data from Sep 21, 2001 to Dec 31, 2007 with deducting 8-day rates from Aug 15 to Aug 24, 2007. There are totally 1571 data.

[Torosantucci *et al.*, 2007] proved that for time step parameter 1/250, at least 550 daily data are necessary. Furthermore, they mentioned that using a large set of data can obtain reliable estimates. 600 daily data are used here to obtain maximum likelihood estimators. That is, starting any day in our data set, we estimate CIR parameters by using the last 600

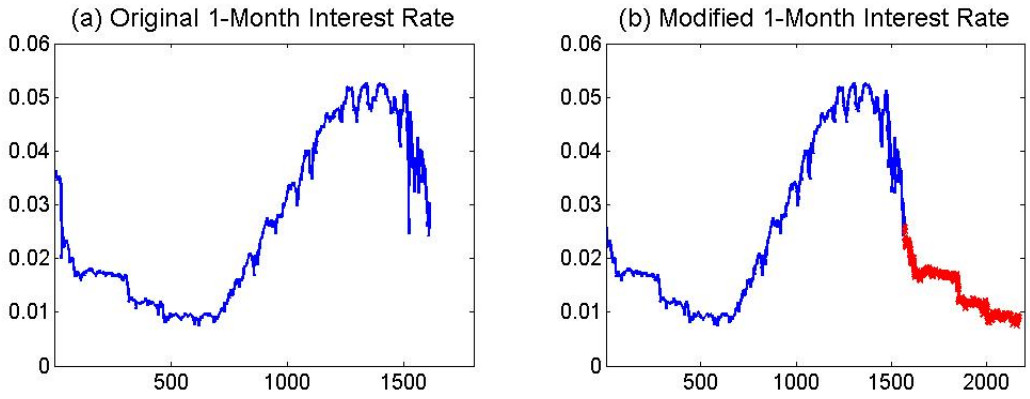

Figure 7: 1-month Interest Rate

data. In order to obtain estimators for all days we choose, we copy the first 599 of 1571 data to the end, so we have totally 2170 data. Figure 7 (b) shows modified one-month interest rate. Next, we estimate each day's κ for Sep 21, 2001 to Dec 31, 2007. We totally obtain 1571 maximum likelihood estimators for κ. However, 485 of 1571 estimators are extremely small (less than 0.0001). We don't think these estimators are useful. Finally we have 1086 estimators.

We use the mean of these κ's above as the value of κ. We insert it into ϕ_1, ϕ_2, ϕ_3 of the CIR model. Finally, we obtain long run average parameters for the CIR model in Table 15.

Table 15: CIR Parameters Estimated II

	κ	θ	σ_r	λ
S1	1.1403	0.0897	0.0467	0.6235
S2	1.1403	0.0642	0.0320	0.3936

Remark

- Under S1, the long run average $\theta = 0.0897$. Obviously, it is too high. The reason is: interest rates between 1990 and 1994 were too high (See Figure 5). We estimate CIR parameters by average historical interest rates; however, interest rates between 1990 and 1994 make the average too high.

- Under S2, the long run average $\theta = 0.0642$, which is normal. We think it is appropriate to analyze the long term behaviour.

4.2.2 IV Based on the BS-CIR Model and Capital Charge

We put all CIR parameters under S2 into the BS-CIR model. We also use modified one-month interest rate and corresponding S&P500 index to estimate ρ. We compute one ρ with 600 one-month interest rates and corresponding S&P500 indexes by applying [Kim, 2002]'s method. Totally we obtain 1571 ρ's with average -0.0015. We set $\rho = -0.0015$. Without loss of generality, we use average HV under the same period in S2 (from Aug 31, 2001 to Sep 2, 2004) as equity volatility (σ_s).

First of all, we compute a call option price by the BS-CIR model, then we obtain IV_{BS-CIR} based on the BS model. Second, we calculate capital charge in terms of average HV, and add option price and potential loss of capital charge up to obtain total option price. Finally, we obtain IV_{Total}.

S&P 500 index is 122.78 on Dec 14, 1981 and 1849.04 on Mar 27, 2014 and its annual return is around 8%, so we can assume potential loss rate is 7%. Total capital charges include largest loss in the scenario matrix and 2% specific risk for index option.

Table 16: Volatility Comparison between Market and Model

Term	1	2	3	4	5	6	7
HV_{Equity}	0.2068	0.2206	0.2165	0.2147	0.2110	0.2053	0.1966
IV_{Market}	0.2326	0.2305	0.2406	0.2444	0.2501	0.2557	0.2611
IV_{BS-CIR}	0.2113	0.2330	0.2377	0.2445	0.2488	0.2510	0.2506
IV_{Total}	0.2252	0.2434	0.2470	0.2528	0.2566	0.2583	0.2575
Term	8	9	10	15	20	25	30
HV_{Equity}	0.1875	0.1793	0.1727	0.1688	0.1707	0.1653	0.1629
IV_{Market}	0.2666	0.2714	0.2766	0.3018	N/A	N/A	N/A
IV_{BS-CIR}	0.2501	0.2507	0.2527	0.2789	0.3024	0.3196	0.3350
IV_{Total}	0.2566	0.2568	0.2585	0.2841	0.3074	0.3245	0.3400

Table 16 shows the volatility comparison results between market and our working models.

HV_{Equity} is average HV from Aug 31, 2001 to Sep 2, 2004. IV_{Market} is the average IV of market from Jun 13 to Nov 8, 2011, totally 168 days. The longest term of market IV is up to 15 years. IV_{BS-CIR} is obtained by the BS-CIR model and IV_{Total} is obtained by combining BS-CIR model and capital requirement charge. Obviously, the long term IV has upward trend with increasing term. Also we find that the capital charge is like a drift term which adjusts IV to go up.

4.2.3 Sensitivity Test

In this section, we discuss sensitivity test for IV. We find that in BS-CIR model, only 2 parameters, θ and σ_s have big impact on the IV. We keep all parameters but changing θ and σ_s. In Table 17, we show the changes in IV when $\sigma_s + 0.05$ and $\theta + 0.005$, respectively.

Table 17: IV sensitivity test for θ and σ_s

Term	1	2	3	4	5	6	7
$\sigma_s + 0.05$	0.0505	0.0500	0.0490	0.0479	0.0466	0.0451	0.0434
$\theta + 0.005$	0.0027	0.0059	0.0089	0.0116	0.0140	0.0163	0.0186
Term	8	9	10	15	20	25	30
$\sigma_s + 0.05$	0.0413	0.0390	0.0367	0.0294	0.0245	0.0192	0.0154
$\theta + 0.005$	0.0209	0.0231	0.0251	0.0312	0.0348	0.0377	0.0396

We find some interesting results from Table 17.

- Long run average rate (θ) affects the IV more and more as term increases. This result matches the common understanding of the interest rate risk on option prices: the interest rate risk should be important for long term options. For short term option, IV changes a little bit as θ changes. Furthermore, if term is shorter than 2 years, interest rate risk can be ignored.

- Equity volatility (σ_s) affects short term IV more than long term IV. When the equity volatility changes 0.05, 1-year IV changes 0.05 as well while 15-year IV just changes 0.03 and 30-year IV just changes 0.015.

4.3 Discussion and Conclusion

1. We successfully explain the IV behaviour after applying the change-points method to find optimal time range.

   It is clear that under the optimal time range S2, all IVs derived from our working model are very close to the market IV. However, if we use data in S1, it would yield much higher IVs than market IV.

2. Interest rate risk dominates long term option IV while equity volatility dominates short term IV.

   HV is stable after 5 years in general. The level is kept around 16% and 17%. In the BS-CIR model, we use HV as equity volatility. The stochastic interest rate make the IV higher and higher with term increasing. In sensitivity test, we add 0.005 to θ, longer term IVs increase more. However, if we add 0.05 to equity volatility, longer term IVs increase less. These findings explain the market phenomenon: When VIX is larger, the shorter term IV is higher, even it can affect 5 year term IV. When VIX is smaller, the shorter term IV is lower. However, no matter how high VIX is, long term IV (5-years or more) don't change a lot. For example, from July to November of 2011, S&P500 dropped a lot made the VIX and short term IV very high, but long term IV just changed a little bit. Actually it can be explained that this way: When index volatility increases, shorter term IV increases much more than longer term IV does.

3. Long term IV contains the information about the market's expectation of long run average of interest rate. It indeed dominates the behaviour of long term IV.

   We change the value of θ to see the changes of IV (see Tabel 18). Noticeably, when $\theta < 0.055$, the IV might not keep upwards. When $\theta > 0.055$, IV increases as term increases apparently.

   We think this phenomenon shows the market expectation of long run average interest rate is higher than 0.055, no matter how low the current interest rate is. If we re-observe the market IV from Jun 13 to Nov 8, 2011, we find the market data imply the market expectation of long run average interest rate is between 0.065 and 0.07.

Table 18: Long run average rate θ's impact on IV

Term	1	2	3	4	5	6	7
$\theta = 0.055$	0.2204	0.2330	0.2308	0.2317	0.2309	0.2281	0.2227
$\theta = 0.06$	0.2230	0.2388	0.2395	0.2432	0.2449	0.2446	0.2417
$\theta = 0.065$	0.2257	0.2446	0.2484	0.2547	0.2589	0.2609	0.2605
$\theta = 0.07$	0.2283	0.2505	0.2573	0.2663	0.2729	0.2772	0.2791
$\theta = 0.075$	0.2310	0.2564	0.2662	0.2779	0.2869	0.2935	0.2976
Term	8	9	10	15	20	25	30
$\theta = 0.055$	0.2171	0.2126	0.2098	0.2211	0.2353	0.2443	0.2543
$\theta = 0.06$	0.2387	0.2369	0.2367	0.2564	0.2761	0.2901	0.3036
$\theta = 0.065$	0.2599	0.2605	0.2625	0.2891	0.3131	0.3306	0.3465
$\theta = 0.07$	0.2808	0.2836	0.2875	0.3201	0.3476	0.3679	0.3856
$\theta = 0.075$	0.3015	0.3062	0.3120	0.3498	0.3803	0.4029	0.4220

4. BS-CIR model can explain long term IV behaviour pretty well. No matter what pattern the short term IV is, the long term IV never decreases. It matches [Tehranchi, 2010] theory that long term implied volatility cannot fall.

5. Market risk charge (or capital requirement charge) and other potential costs make IV higher.

   According to our finding, the cost (or potential loss) of market risk charge should be considered in the option price. For S&P500 index call option, market risk charge increases the IV level around 0.5%. Also we think other extra charges like transaction cost, hedging cost, etc., also increase the price of option so that IV becomes higher and higher.

5 Summary and Future Work

5.1 Summary

In the dissertation, we analyze the volatility in finance by statistical tools.

In Chapter 2, we propose weighted sum of powers of variance (WSPV) to find single change-point in HV. Our simulation results show that the modified WSPV has better performance than the WSPV and BIC when sample size is small.

In Chapter 3, we propose an empirical BIC method to detect multiple change-points in HV simultaneously. Simulation results show better performance than some other multiple change-points methods.

In Chapter 4, we successfully explain the long term index IV behaviour by using change-points method to find the optimal time range. After using data in optimal time range, we obtain much better results to match the market data than results derived from without change-points. We have some important findings:

- Interest rate risk dominates long term option IV while equity volatility dominates short term IV.
- Long term IV contains the information about the market's expectation of long run average of interest rate. In fact, it dominates the behaviour of long term option IV. It makes long term options IV goes up as term increases.
- Furthermore, long term IV also reflects the extra charge in option pricing, like capital charges, transaction cost, hedging cost, and so on.

5.2 Future Work

- HV plays a very important role in the stock market. It is often compared with IV to determine if options prices are correct. Furthermore almost all kinds of risk valuations need HV. Therefore, we believe finding change-points in HV is very important in risk

valuations. Changes in HV may cause big changes in risk. We are going to do some other risk valuations by finding change-points in HV in the future.

- Although our MPSWV method performs very well in finding changes in HV when sample size is small, limitation is that the power λ is estimated by simulations. We are going to develop an algorithm to find the optimal value by the original data. Also, simultaneous detection of multiple change-points by MPSWV is being considered.

- In long term IV analysis, we think the capital charge should have more impact on IV. The reason is simple: the general market risk charge is calculated together with the associated hedging. We follow [Soner *et al.*, 1994]'s trivial strategy of buying one share of the underlying stock and holding to maturity. Actually we cannot ignore cost of hedging and transaction fee in long term option since stock would change a lot in long term and trivial hedging strategy is not enough. Therefore we are considering an optimal hedging strategy which can make the balance of risk and cost. After that we can estimate the real capital charge for long term options.

- We would like to add more factors in long term IV analysis. We think there should be other costs which affect the option price so that the long term IV goes up as term increases.

Bibliography

Aue, A. and Horváth, L. (2013). Structural breaks in time series. *Journal of Time Series Analysis*, **34**, 1-16.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models, *Journal of Applied Econometrics*, **18**, pp. 1-22.

Bakshi, G. and Madan, D. (2006), A Theory of Volatility Spreads. *Management Science* 52, pp.1945-1956.

Bakshi S., Cao C., and Chen Z.(2000). Pricing and hedging long-term options. *Journal of Econometrics* 94, pp.277318.

Barry, D. and Hartigan, J.A. (1993). A Bayesian analysis for change point problems, *Journal of the American Statistical Association*, **88**, pp. 309-319.

Baufays, P. and Rasson, J. (1985). Variance Changes in Autoregressive Models, in *Time Series Analysis: Theory and Practice 7*, ed. O.D. Anderson, North-Holland, pp. 119-127.

Billingsley, P. (1968), Convergence of probability Measures, *New York: John Wildy*

Birge, L. and Massart, P. (2001). Gaussian model selection, *Journal of the European Mathematical Society*, 3, pp. 203-268.

Black, F. and Scholes, M.S. (1973), The pricing of options and corporate liabilities. *J. Polit. Econ.* 81, pp. 637-665.

Bollerslev, T., Gibson, M. and Zhou, H. (2011), Dynamic estimation of volatility risk premia and investor risk aversion from Option-implied and realized volatilities, *Journal of Econometrics* 160, pp. 235-245.*et al.*

Bollerslev, T., Tauchen, G. and Zhou, H. (2008), Expected Stock Returns and Variance Risk Premia, *Review of Financial Studies*, Forthcoming.

Box, G.E.P. and Jenkins, G.(1976). *Time Series Analysis: Forecasting and Control.* San Francisco: Holden-Day.

Casella, G. and George, E.I. (1992). Explaining the Gibbs sampler, *American Statistician*, **46**, 167-174.

Chen, J. and Gupta, A. K. (2012), Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance, 2nd edition, *Birkhauser*.

Chen, J. and Gupta, A. K. (1997). Testing and locating variance change points with application to stock prices, *Journal of the American Statistical Association*, 92, pp. 739-747.

Chen, X. (2013). The limit law of the iterated logarithm. *Journal of Theoretical Probability*, 1-5.

Cox J. C., Ingersolla J. E., and Ross S.A.(1985), A theory of the term structure of interest rates, *Econometrica*, 53, pp. 385-407.

Csörgo, M. and Horváth, L.(1997). *Limit Theorems in Change-point Analysis*, Wiley, Chichester.

Cui, J., Pitt, D. and Qian (2010). Model selection and claim frequency for workers compensation insurance. *ASTIN Bulletin The Journal of the International Actuarial Association*, **40**, No. 2, 779-796.

Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A.R. (2006). Structural break estimation for non-stationary time series models. *Journal of the American Statistical Association* **101**, 223-239.

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm, *Annals of Statistics*, **27**(1), pp. 94-128.

Dybvig, P., Ingersoll J., Ross S. (1996), Long forward and zero-coupon rates can never fall, *Journal of Business*, 60, pp. 1-25.

Eraker, B. (2008) The volatility premium, *Working paper*, Duke University.

Figlewsky, S. (1989) Options arbitrage in imperfect markets, *Journal Finance*, 44, pp.1289-1311.

Figlewsky, S. (1997) Forecasting Volatility, *Financial Markets, Institutions & Instruments*, 6, pp.1-88.

Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression. *Journal of Royal Statistical Society B*, **40**, 190-195.

Heston, S. L. (1993), A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options, *The Review of Financial Studies*, Vol 6, pp. 327-343.

Hull, J. and White, A.(1993), One factor interest rate models and the valuation of interest rate derivative securities. *Journal of Financial and Quantitative Analysis*, 28(2).

Hull, J. C. (2011), Options, Futures and Other Derivatives, 8ed, *Prentice Hall*.

Inclan, C. and Tiao, G.C.(1994). Use of cumulative sums of squares for retrospective detection of changes of variance, *Journal of the American Statistical Association*, **89** No. 427, 913-923.

Inclan, C. (1993), Detection of multiple changes of variance using posterior odds, *Journal of Business and Economics Statistics*, 11, 289-300.

Ito, K. (1951), On stochastic differential equations, *Memoirs, American Mathematical Society*, 4, 1-51.

Jensen,J.L.(1906),Sur les fonctions convexes et les inegalites entre les valeurs moyennes, *Acta Math*, 30 pp. 175-193.

Kim, J. and Cheon, S.(2010). Bayesian multiple change-point estimation with annealing stochastic approximation Monte Carlo, *Comput Stat*, **25**, pp. 215-239.

Kim, Y.-J. (2002), Option Pricing under Stochastic Interest Rates: An Empirical Investigation, Kim, *Asia-Pacific Financial Markets* 9, 23-44.

Kladivko, Kamil (2007), Maximum likelihood estimation of the Cox-Ingersoll-Ross process: the Matlab implementation, *Technical Computing Prague*.

Kurozumi, E. and Tuvaandorj, P. (2011). Model selection criteria in multivariate models with multiple structure changes. *Journal of Econometrics*, **164**, 218-38.

Lavielle, M. and Lebarbier, E. (2001). An application of MCMC methods to the multiple change points problem, *Signal Processing*, **81**, pp. 39-53.

Lavielle, M. and Teyssiere G. (2006). Detection of multiple change points in multivariate time series, *Lithuanian Mathematical Journal*, **46**, No. 3.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem, *Signal Processing*, **85**, pp. 1501-1510.

Liang, F. (2007). Annealing stochastic approximation Monte Carlo for neural network training, *Mach Learn*, **68**(3), pp. 201-223.

Qian, G. and Field, C. (2002). Using MCMC for logistic regression model selection involving large number of candidate models. In Fang, K.T., Hickernell, F.J. and Niederreiter, H. (Eds.) *Selected Proceedings of the 4th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, Springer, Hong Kong, 460-474.

Qian, G. and Zhao, X. (2007). Using Gibbs sampler for time series model selection involving many candidate ARMA models. *Computational Statistics and Data Analysis*, **51**, 6180-6196.

Qian, G.(1999). Computations and analysis in robust regression model selection using stochastic complexity, *Computational Statistics*, **14**, 293-314.

Qian, G., Shi, X. and Wu, Y. (2014). A statistical test of change-point in mean that almost surely has zero error probabilities. *Aust. N. Z. J. Stat.*, **55**(4), 435-454.

Rebonato, R. (1998), Interest-Rate Option Models, *John Wiley and Sons Ltd*, 2nd edition.

Robert, C.P. and Richardson, S. (1998). Markov chain Monte Carlo methods. In Robert, C.P. (Ed.) *Discretization and MCMC Convergence Assessment*, Lecture Notes in Statistics **135**, Springer, New York, 1-25.

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

Seshan, V.E. and Olshen, A. (2011). DNAcopy: DNA copy number data analysis. *R package version 1.16.0.*

Shi, X., Wu, Y. and Miao, B.(2009). Strong convergence rate of estimators of change-point and its application. *Computational Statistics and Data Analysis*, **53**, 990-998.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D., and Albertson, D.G. (2001). Assembly of microarrays for genome-wide measurement of DNA copy number, *Nature Genetics*, **29**, 263-4.

Soner, H. M., Shreve, S. E. and Cvitanic, J.(1994), There Is No Non-trivial Hedging Portfolio for Option Pricing with Transaction Costs.

Tehranchi, M. R. (2010), Implied Volatility: Long Maturity Behaviour, *Encyclopedia of Quantitative Finance*.

Torosantucci, L., Uboldi, A., and Bernaschi, M. (2007), Empirical Evaluation of the Market Price of Risk Using the CIR Model, *International Journal of Theoretical and Applied Finance*.

Tsay, R.S.(1998). Outilers, Level Shifts and Variance Changes in Time Series. *Journal of Forecasting*, **7**, 1-20.

Vasicek, O. (1977), An equilibrium characterization of the term structure, *Journal of Financial Economics*, 5, pp. 177-188.

Vostrikova, L. J. (1981). Detection of 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, **24**, 55-59.

Wichern, D.W., Miller, R.B. and Hsu, D.A. (1976). Changes of Variance in First-Order Autoregressive Time Series Models - With an application, *Applied Statistics*, **25**, 248-256.

Yao, Y. (1988). Estimating the number of change points via Schwarz criterion. *Statistics and Probability Letters*, **6**(3), 181-189.

Zhou, H., (2010), Variance risk premia, asset predictability puzzles, and macroeconomic uncertainty, Working paper, *Federal Reserve Board*.

Appendices

A-1 Introduction of CUSUM and BIC-type Method

[Inclan and Tiao, 1994] introduced CUSUM method. The statistic is

$$D_k = \frac{\sum_{i=1}^{k} x_i^2}{\sum_{i=1}^{n} x_i^2} - \frac{k}{n},$$

and

$$\sqrt{n/2} D_k \xrightarrow{D} B_0,$$

where B_0 is a Brownian bridge.

The distribution of $\sup |B_0|$ is given by [Billingsley, 1968]

$$P\{\sup |B_0| \leq b\} = 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2 b^2}.$$

[Chen and Gupta, 2012] introduced a critical value c_α into their BIC model for single change-point detection. Let $BIC(n)$ denote no change-point, and $BIC(k), 1 < k < n$ denote single change-point at position k, and significance level is α. c_α is defined when H_0 no change-point is accepted if

$$BIC(n) < \min_{1<k<n} BIC(k) + c_\alpha,$$

where

$$BIC(n) = n \log 2\pi + n \log \hat{\sigma}^2 + n + \log n,$$

$$BIC(k) = n \log 2\pi + k \log \hat{\sigma}_{1,k}^2 + (n-k) \log \hat{\sigma}_{n,k}^2 + n + 2 \log n,$$

and

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

A-2 Introduction of BS Model

The Black-Scholes (BS) model [Black and Scholes, 1973] assumes that there is a riskless asset with expected return μ and constant volatility σ. The dynamics of the price S of the underlying asset are

$$dS = \mu S dt + \sigma S dB_t, \tag{A1}$$

where B_t is a standard Brownian motion which satisfies

$$dB_t = \epsilon\sqrt{dt},$$

where ϵ is standard Normal distribution with mean 0 and variance 1. The dynamics of price S can be expressed as

$$S_T = S_0 \exp\{\mu t - \frac{1}{2}\sigma^2 T + \sigma B_T\}.$$

It also implies that S_T has a Lognormal distribution. $\log S_T$ has a normal distribution with mean $\log S_0 + (\mu - \sigma^2/2)T$ and variance $\sigma^2 T$.

Suppose f is the price of derivative based on S, by Ito lemma [Ito, 1951], we have

$$df = (\frac{\partial f}{\partial S}\mu S + \frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial S^2}\sigma^2 S^2)dt + \frac{\partial f}{\partial S}\sigma S dB. \tag{A2}$$

The discrete version of (A1) and (A2) are

$$\Delta S = \mu S dt + \sigma S \Delta B$$

and

$$\Delta f = (\frac{\partial f}{\partial S}\mu S + \frac{\partial f}{\partial t} + \frac{1}{2}\frac{\partial^2 f}{\partial S^2}\sigma^2 S^2)\Delta t + \frac{\partial f}{\partial S}\sigma S \Delta B.$$

By choosing an appropriate portfolio, the standard Brownian motion can be deleted ([Hull, 2011]'s book shows in details). Finally we obtain a differential equation:

$$rf = \frac{\partial f}{\partial t} + rS\frac{\partial f}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 f}{\partial S^2},$$

where r is risk free interest rate. This differential equation has many solutions, corresponding to all different derivatives related to underlying assets.

In case of European call option, the boundary condition is

$$f = \max(S - K, 0) \text{ when } t = T,$$

where K is strike price. The call option price C at time 0 is

$$C = S_0 N(d_1) - K e^{-rT} N(d_2),$$

where

$$d_1 = \frac{\log(S_0/K) + (r + \sigma^2/2)/T}{\sigma\sqrt{T}} \quad \text{and}$$

$$d_2 = d_1 - \sigma\sqrt{T},$$

and $N(\cdot)$ is the cdf for a standard Normal distribution.

In case of European put option, the boundary condition is

$$f = \max(K - S, 0) \text{ when } t = T$$

and the put option price P at time 0 is

$$P = Ke^{-rT}N(-d_2) - S_0N(-d_1).$$

A-3 Maximum Likelihood Estimation of the CIR Process

[Kladivko, 2007] introduces how to do maximum likelihood estimation of the CIR process by Matlab. Here we briefly introduces it.

The CIR model is

$$dr_t = \kappa(\theta - r_t)dt + \sigma_r\sqrt{r_t}dB,$$

where θ is the long run average interest rate, κ is the reverting rate at r_t to θ, σ_r is the volatility of the short interest rate r_t.

Given r_t, the density of $r_{t+\delta t}$ is

$$p(r_{t+\Delta t}|r_t) = ce^{-u-v}(\frac{v}{u})^{\frac{q}{2}}I_q(2\sqrt{uv})$$

where

$$\begin{aligned} c &= \frac{2\kappa}{\sigma_r^2(1-e^{-\kappa\Delta t})}, \\ u &= cr_te^{-\kappa\Delta t}, \\ v &= cr_{t+\Delta t}, \\ q &= \frac{2\kappa\theta}{\sigma_r^2} - 1, \end{aligned}$$

and $I_q(2\sqrt{uv})$ is modified Bessel function of the first kind with order q.

For maximum likelihood estimation of the parameter $(\kappa, \theta, \sigma_r)$, the log-likelihood function of $r_{t+\Delta t}|r_t$ is

$$\log L(\kappa, \theta, \sigma_r) = (N-1)\log c + \sum_{i=1}^{N-1}\{-u_{t_i} - v_{t_{i+1}} + 0.5q\log(\frac{v_{t_{i+1}}}{u_{t_i}}) + \log(I_q(2\sqrt{u_{t_i}v_{t_{i+1}}}))\}$$

where $u_{t_i} = cr_{t_i}e^{-\kappa\Delta t}$ and $v_{t_{i+1}} = cr_{t_{i+1}}$. Thus, we can find maximum likelihood estimation

$$(\hat{\kappa}, \hat{\theta}, \hat{\sigma_r}) = \arg\max \log L(\kappa, \theta, \sigma_r).$$