A Panel Data Analysis of the Gini Coefficient

Jia Wang

Supervisory Committee:
Professor Lajos Horváth (Committee Chair)
Professor Firas Rassoul-Agha
Professor Jinyi Zhu

A project submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of
Master of Statistics Emphasis Mathematics

Department of Mathematics
University of Utah
2012

Acknowledgement

I would like to thank my family and my friends at the University of Utah for their support. I especially wish to thank Professor Lajos Horváth for his generous help and support. Professor Horváth's insightful guidance and statistical help have been invaluable supports from the beginning through several revisions of this project. Thanks also go to Christopher Robison for proofreading and polishing my work.

Abstract

The purpose of this project is to study common break point in the Gini coefficient data using panel data analysis. We propose two approaches to develop the limiting distribution for the estimated break point and evaluate the approximation via Monte Carlo simulation. The Quasi-maximum likelihood method is shown to be more efficient in detecting the instability in the panels than the Darling-Erdős Limit Results method, especially in the case when the number of panels and sample sizes are small.

Contents

List of Figures

List of Tables

1 Introduction

Panel data refers to the pooling of observations on a cross-section of subjects like households, countries, firms, etc., over several time periods (see [2]). The structure of panel data is different from the classical cross-sectional data, or time series data. With panel data, we observe not only one subject over time, but also multiple subjects at the same time. So panel data analysis is like a marriage of regression and time series analysis. Hence, it allows us to study dynamic and cross-sectional aspects of a problem. For example, in measuring unemployment, cross-sectional data can estimate what proportion of the population is unemployed at a point in time. Repeated cross-sections can further reveal how this proportion changes over time. However, only panel data can estimate what proportion of those are unemployed in one period, but remain unemployed in another period.

In the U.S., two famous panel data sets are the National Longitudinal Surveys of Labor Market Experience (NLS) and the Panel Study of Income Dynamics (PSID), conducted by the University of Michigan. Starting in the mid-1960s, the NLS consists of five distinct segments of the labor force: older men aged between 45 and 49 in 1966, young men aged between 14 and 24 in 1966, mature women aged between 30 and 44 in 1967, young women aged between 14 and 21 in 1968, and youth of both sexes aged between 14 and 24 in 1979. Thousands of variables were collected with emphasis on the supply side of the labor market, including demographic information, training investments, child care usage, as well as drug and alcohol use, etc. The PSID, on the other hand, is concerned with exploring the relationship between household income and socioeconomic characteristics of each family. A large number of studies have used the NLS and the PSID data sets.

Panel data possess some advantages over cross-sectional or time series data, such as controlling for individual heterogeneity, increasing the number of data points along with the degrees of freedom, and reducing the collinearity among explanatory variables. Therefore, panel data are well suited to study economic problems like income, unemployment, poverty level, and so on. An illustrative empirical example is given by Hajivassiliou (1997), who studies the external debt repayment problems, using a panel of 79 developing countries observed over the period 1970-1982. The model considers country specific variables such as their colonial history, financial institutions, religious affiliations and political regimes. Because all these factors affect the attitudes that these countries have with regards to borrowing and defaulting and the way they are treated by the lenders; ignoring this country heterogeneity will run the risk of obtaining biased results (see [2]).

2 Data Source and Description

The Gini coefficient is a widely accepted measurement of income inequality in economics study. Roughly speaking, a Gini coefficient of zero represents perfect income equality, i.e., everyone gets exactly the same income. On the other hand, a Gini coefficient close to one implies high

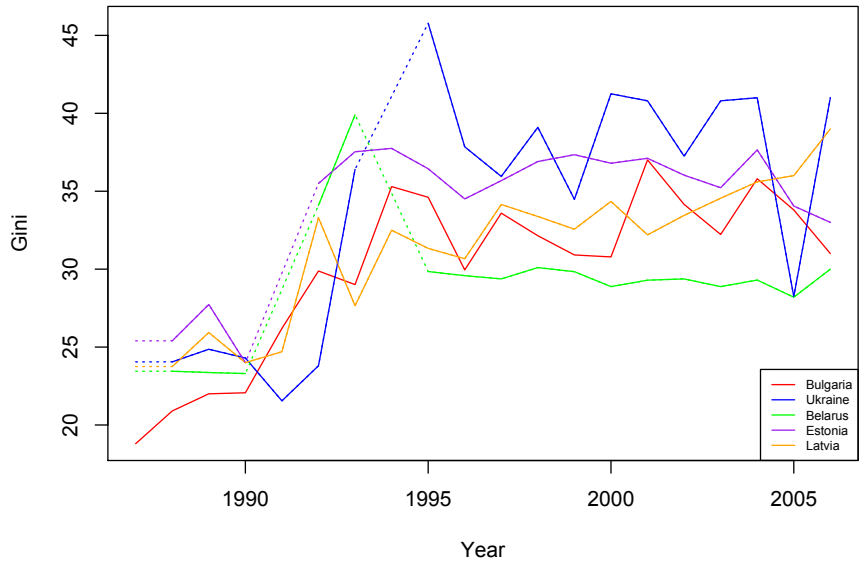
inequality, in other words, wealth is concentrated among only a few people while others are all in extreme poverty. According to the U.S. Census Bureau, most European developed countries tend to have Gini indices between 0.24 and 0.36.

The purpose of this project is to examine if there is a common change in means over the certain time period using a panel data analysis model. Common breaks in panel data are a wide spread phenomena. For example, an oil price shock may affect almost every country's economic growth. The data set we used in this project is extracted from the World Income Inequality Database (WIID2C), which can be accessed from UNU-WIDER website (see `http://www.wider.unu.edu/research/Database/en_GB/wiid/`). The original data set contains the Gini coefficients in percentage points and relative source information of 159 countries, with some records even reaching back as far as 1860. The variable, Gini coefficient, was calculated by WIDER using methods developed by Shorrocks and Wan to estimate the Gini coefficient from decile data (see [8]).

The first step in our analysis procedure is to reshape the raw data set (WIID2C) from long version to wide version in SAS, thus making it more convenient to use. Note, if several Gini coefficients are reported on the same year, then we will use the mean of these numbers as the Gini coefficient of that year. It becomes evident that lots of the data points are missing for most of the countries, especially before the 1980s, so a truncation in time is necessary. In order to ensure enough number of countries and data points, we decide to choose a twenty-year span from 1987 to 2006. Hence, a total number of 33 countries, including 24 European countries (United Kingdom, Germany, Spain, and so on), Australia, United States, China, Taiwan and South American countries (Brazil, Argentina, and so on), each with at least 16 data points during this time period, are extracted from the reshaped data set using the statistical software package R. To replace missing values, we consider the following two cases:

- If the first data point, or the last data point are missing, then replace it with the first following data point, or the last preceding data point, respectively.
- Otherwise, replace the missing value(s) by linear interpolation.

Figure 2.1 to Figure 2.4 are plots for the data set by regional groups, with dotted lines representing linear interpolation. The raw data set and the data set after replacing missing values are included in the Appendix.



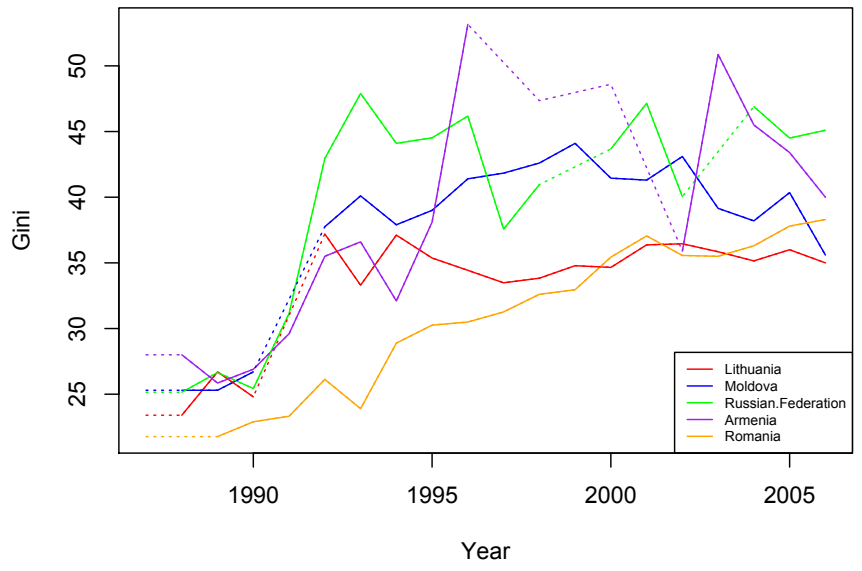
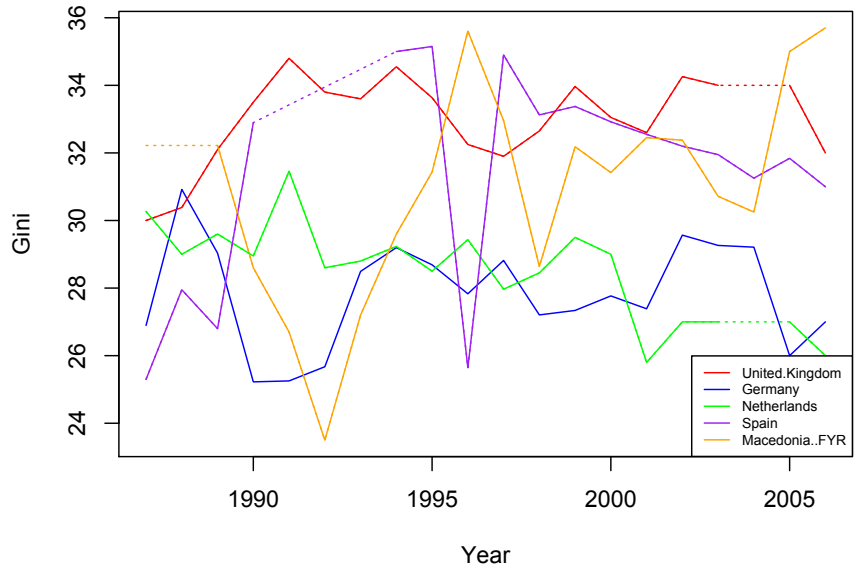


Figure 2.1: Plots for Gini coefficients of selected Countries (Group 1 and 2).



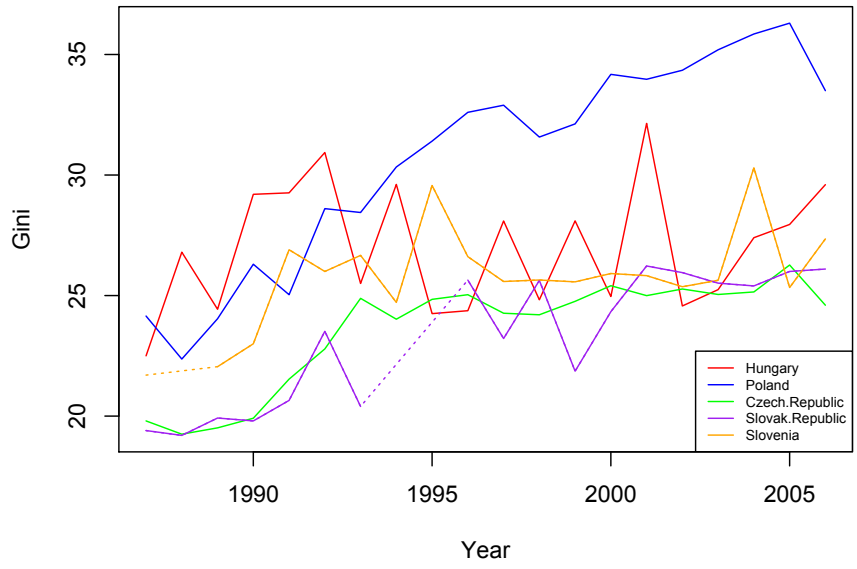
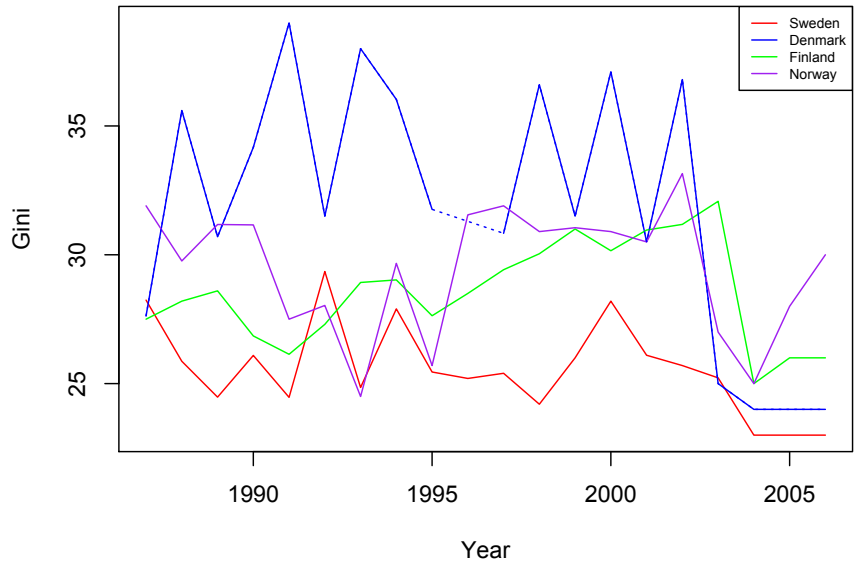


Figure 2.2: Plots for Gini coefficients of selected Countries (Group 3 and 4).



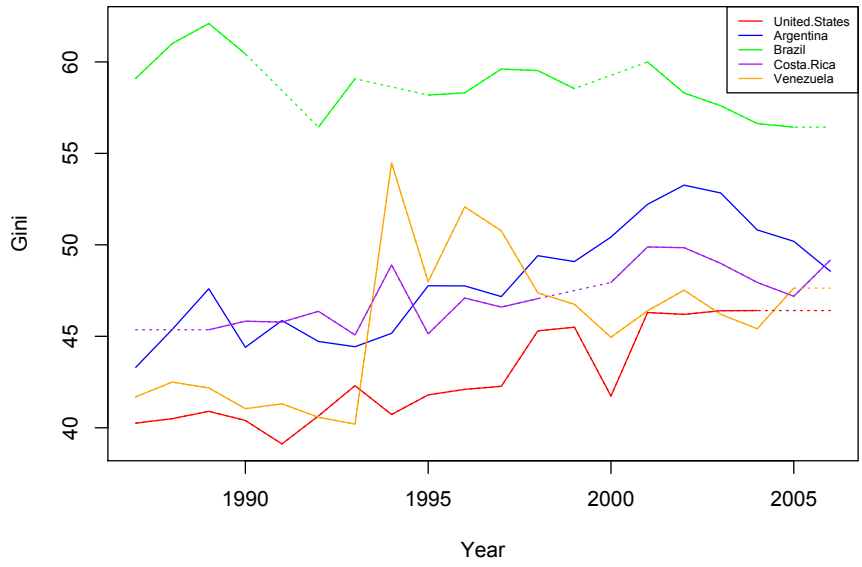


Figure 2.3: Plots for Gini coefficients of selected Countries (Group 5 and 6).

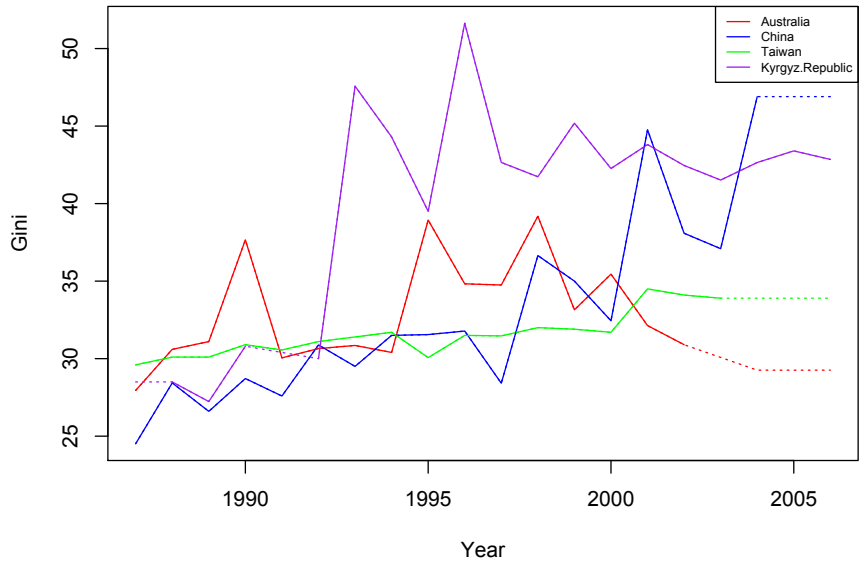


Figure 2.4: Plots for Gini coefficients of selected Countries (Group 7).

3 The Model

The panel data model is defined as

$$X_{i,j} = \mu_i + \delta_i I_{\{j>k_0\}} + e_{i,j}, \quad 1 \le i \le N, 1 \le j \le T. \tag{3.1}$$

where $E[e_{i,j}] = 0$ for all i and j. We call the series $\{X_{i,j}, 1 \le j \le T\}$ the i^{th} panel. In this model, each panel has a common break point at k_0, where k_0 is unknown. δ_i represents the magnitude of the break, which can be either random or nonrandom and is assumed to be independent of the error process $e_{i,j}$. The pre-break mean of the i^{th} panel is μ_i and the post-break mean is $\mu_i + \delta_i$. We refer to N as the number of panels or the number of series, and refer to T as the number of observations or sample size. For panel data, N is usually large relative to T. We are interested in testing the null hypothesis:

$$H_0 : \delta_i = 0 \text{ for all } 1 \le i \le N. \tag{3.2}$$

This means under the null hypothesis, there is no change in the location parameter (mean) μ_i throughout the whole observation period. Under the alternative hypothesis, there is an integer k_0, $1 \le k_0 < T$, such that

$$\mu_{i,1} = \mu_{i,2} = \cdots = \mu_{i,k_0} \neq \mu_{i,k_0+1} = \cdots = \mu_{i,T} \text{ for } 1 \le i \le N.$$

The alternative implies that the means have changed at the same time in each panel. It is obvious that if $k_0 = T$, then there is no change over the time.

In Bai's paper (see [1]), he used a quasi-maximum likelihood approach to estimate the time of change k_0, by the location of the maximum of the absolute value of

$$\bar{V}_{N,T}(t) = \frac{1}{N^{1/2}} \sum_{i=1}^{N} \left\{ \frac{1}{\sigma_i^2} Z_{T,i}^2(t) - \frac{\lfloor Tt \rfloor (T - \lfloor Tt \rfloor)}{T^2} \right\}, \quad 0 \le t \le 1, \tag{3.3}$$

where

$$Z_{T,i}(t) = \frac{1}{T^{1/2}} \left(S_{T,i}(t) - \frac{\lfloor Tt \rfloor}{T} S_{T,i}(1) \right), \quad 0 \le t \le 1,$$

and

$$S_{T,i}(t) = \sum_{j=1}^{\lfloor Tt \rfloor} X_{i,j}, \quad 0 \le t \le 1,$$

with σ_i^2's being some suitably chosen standardization constants satisfying

$$\lim_{T \to \infty} \frac{1}{T} E \left(\sum_{j=1}^{T} e_{i,j} \right)^2 = \sigma_i^2, \quad 1 \le i \le N. \tag{3.4}$$

Note, here $\lfloor \cdot \rfloor$ means taking the integer part. σ_i^2 is also known as the long-run variance of $e_{i,j}, 1 \le j \le T$, that is, the spectral density at frequency zero.

Definition 3.1. *A **Linear Process**, x_t, is defined to be a linear combination of white noise variates w_t, and is given by*

$$x_t = \mu + \sum_{j=-\infty}^{\infty} \psi_j \omega_{t-j}, \qquad \sum_{j=-\infty}^{\infty} |\psi_j| < \infty. \tag{3.5}$$

For the linear process in Definition 3.1, we may show that $E[x_t] = \mu$ and the autocovariance function

$$\gamma(h) = \mathrm{Cov}(x_t, x_{t+h}) = \sigma_w^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h};$$

where $\sigma_w^2 = \mathrm{Var}(w_t)$. Note, $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ implies $\sum_{j=-\infty}^{\infty} \psi_j \psi_{j+h} < \infty$, for fixed h. Hence, for a linear process with white noise errors, its mean and autocovariance function are well defined. In this project we follow a model where the error terms form a linear process with mean 0:

$$e_{i,j} = \sum_{\ell=0}^{\infty} c_{i,\ell} \varepsilon_{i,j-\ell}, \quad 1 \le i \le N, 1 \le j \le T. \tag{3.6}$$

Definition 3.2. *A **weakly stationary** time series, x_t, is a finite variance process such that:*

(i) the mean function, μ_t, is a constant and does not depend on time t, and

(ii) the autocovariance function, $\gamma(s,t)$, depends on s and t only through their difference $|s-t|$.

Assumption 3.1. *$\varepsilon_{i,j}$'s in (3.6) are assumed to satisfy the following regularity conditions:*

(i) the sequences $\{\varepsilon_{i,j}, -\infty < j < \infty\}$ are independent of each other, hence the panels are independent of each other;

(ii) for every i the variables $\{\varepsilon_{i,j}, -\infty < j < \infty\}$ are independent and identically distributed.

It is easy to see that under these conditions the process $\bar{V}_{N,T}$ does not depend on $\mathrm{Var}(\varepsilon_{i,0})$ so it can be assumed that the variance of the innovations is 1, and that the higher moments exist:

$$E\varepsilon_{i,0} = 0,\ E\varepsilon_{i,0}^2 = 1 \text{ and } E|\varepsilon_{i,0}|^\kappa < \infty. \tag{3.7}$$

Besides, we require the average of the high moments of $\varepsilon_{i,0}$'s to be bounded:

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} E|\varepsilon_{i,0}|^\kappa < \infty. \tag{3.8}$$

The choice of κ will be specified in Theorems 3.1 and Theorem 4.1 as $\kappa > 4$ and $\kappa = 8$. The errors in each panel are stationary linear processes and their distributions depend on the panels. The coefficients satisfy the following properties:

Property 3.1. *In Definition 3.1, the coefficients of the linear sequences in (3.6) have the following properties:*

(i) $|c_{i,\ell}| \le c_0(\ell+1)^{-\alpha}$ for all $1 \le i \le N$, $1 \le \ell < \infty$ with some c_0 and $\alpha > 2$;

(ii) there is $\delta > 0$ such that $a_i^2 \ge \delta^2$ with $a_i = \sum_{\ell=0}^{\infty} c_{i,\ell}$ for all $1 \le i \le N$.

Notice that under Property 3.1 (ii), together with (3.4) and (3.7), we get

$$E[e_{i,j}^2] = \sum_{\ell=0}^{\infty}\sum_{k=0}^{\infty} c_{i,\ell}c_{i,k}E[\varepsilon_{i,j-\ell}\varepsilon_{i,j-k}] = \left(\sum_{\ell=0}^{\infty} c_{i,\ell}\right)^2,$$

which implies that $a_i^2 = \sigma_i^2$. Thus,

$$\sigma_i^2 \ge \delta^2 \text{ for all } 1 \le i \le N, \tag{3.9}$$

i.e., we have a common lower bound, δ^2, for the long-run variances of each panel.

Assumption 3.2. *The number of panels (N) and the length of the observed time series in each panel (T) are assumed to satisfy:*

$$\frac{N}{T^2} \to 0. \tag{3.10}$$

Note, this assumption allows the number of panels to be larger than the number of the observations in each panel.

Definition 3.3. *A process, $\{x_t\}$, is said to be a* ***Gaussian process*** *if the n-dimensional vectors $\mathbf{x} = (x_{t_1}, x_{t_2}, \ldots, x_{t_n})'$, for every collection of time points, $t_1, t_2, \ldots, t_n$, and every positive integer n, have a multivariate normal distribution.*

Definition 3.4. *The Skorokhod space $\mathcal{D}[0,1]$ is the set of all right continuous and with left limit functions (RCLL) from $[0,1]$ in $\mathbb{R}$.*

Let $\xrightarrow{\mathcal{D}[0,1]}$ denote the weak convergence of stochastic processes in the Skorokhod space $\mathcal{D}[0,1]$.

Theorem 3.1. *Under H_0, if Assumption 3.1, Property 3.1, (3.10) hold and (3.7), (3.8) are satisfied with some $\kappa > 4$, then*

$$\bar{V}_{N,T}(t) \xrightarrow{\mathcal{D}[0,1]} \Gamma(t)$$

where $\Gamma(t)$ is a Gaussian process with $E\Gamma(t) = 0$ and $E\Gamma(t)\Gamma(s) = 2t^2(1-s)^2$, if $0 \le t \le s \le 1$.

Theorem 3.1 provides a possible way to study the asymptotic behavior of $\bar{V}_{N,T}(t)$ by investigating the Gaussian process $\Gamma(t)$. More precisely, $\sup_{1\le t\le 1} |\bar{V}_{N,T}(t)|$ and $\int_0^1 \bar{V}_{N,T}^2(t)dt$ converge in distribution to $\sup_{0\le t\le 1} |\Gamma(t)|$ and $\int_0^1 \Gamma^2(t)dt$. A formal, detailed proof of Theorem 3.1 can be found in Horváth and Hušková (2012) [6].

Definition 3.5. *A continuous time process $\{W(t); t \ge 0\}$ is called a* ***standard Brownian motion****, or a* ***Wiener process****, if it satisfies the following conditions:*

(i) $W(0) = 0$;

(ii) $\{W(t_2) - W(t_1), W(t_3) - W(t_2), \ldots, W(t_n) - W(t_{n-1})\}$ *are independent for any collection of points,* $0 \le t_1 < t_2 \cdots < t_n$, *and integer* $n > 2$;

(iii) $W(t + \Delta t) - W(t) \sim N(0, \Delta t)$ *for* $\Delta t > 0$.

Proposition 3.1. *If $\{W(t); t \ge 0\}$ is a Wiener process, then*

$$E[W(t)] = 0,\ \mathrm{Var}[W(t)] = t,\ \mathrm{Cov}(W(s), W(t)) = \min(s,t).$$

The results for the expectation and variance follow immediately from definition (3.5), $W(t) = W(t) - W(0) \sim N(0,t)$. To check the results for the covariance, suppose $0 \le t \le s \le 1$, since non-overlapping increments are independent, it is easy to see:

$$\begin{aligned}
\mathrm{Cov}(W(t), W(s)) &= E[(W(t) - E[W(t)]) \cdot (W(s) - E[W(s)])] \\
&= E[W(t) \cdot W(s)] \\
&= E[W(t) \cdot ((W(s) - W(t)) + W(t))] \\
&= E[W(t) \cdot (W(s) - W(t))] + E[W(t)^2] \\
&= E[(W(t) - W(0)) \cdot (W(s) - W(t))] + \mathrm{Var}[W(t)] + E^2[W(t)] \\
&= t
\end{aligned}$$

Corollary 3.1. *Under Theorem 3.1, we have*

$$\{\Gamma(t), 0 \le t \le 1\} \stackrel{\mathcal{D}}{=} \{\sqrt{2}(1-t)^2 W\left(t^2/(1-t)^2\right), 0 \le t \le 1\}, \tag{3.11}$$

where $\{W(t), 0 \le t \le 1\}$ *is a Wiener process.*

Proof: Since a Gaussian process is uniquely determined by its mean and covariance structure, checking the mean and covariance functions of $\Gamma(t)$ in Theorem 3.1 will be sufficient. It is easy to see the mean is zero. Suppose $0 \le t \le s \le 1$, the covariance function for the right-hand-side Wiener process is

$$\begin{aligned}
&\operatorname{Cov}\left(\sqrt{2}(1-t)^2 W\left(\frac{t^2}{(1-t)^2}\right), \sqrt{2}(1-s)^2 W\left(\frac{s^2}{(1-s)^2}\right)\right) \\
&= 2(1-t)^2(1-s)^2 \operatorname{Cov}\left(W\left(\frac{t^2}{(1-t)^2}\right), W\left(\frac{s^2}{(1-s)^2}\right)\right) \\
&= 2(1-t)^2(1-s)^2 \min\left(\frac{t^2}{(1-t)^2}, \frac{s^2}{(1-s)^2}\right) \\
&= 2t^2(1-s)^2 \\
&= E[\Gamma(t)\Gamma(s)].
\end{aligned}$$

It is well known that the CUSUM process (standardized by the long-run variance) converge weakly to a Brownian bridge assuming weak dependence. Under the condition of Theorem 3.1, for each i the process $Z_{T,i}$ converges to a Brownian bridge, so it is interesting to compare (3.11) to $(1-t)W(t/(1-t))$ which defines a Brownian bridge.

Theorem 3.2. *If Assumption 3.1, Proposition 3.1, (3.10) hold and (3.7), (3.8) are satisfied with some* $\kappa > 4$, *and*

$$0 < \liminf_{T\to\infty} \frac{k_0}{T} \le \limsup_{T\to\infty} \frac{k_0}{T} < 1,$$

$$\frac{T}{N^{1/2}} \sum_{i=1}^{N} \delta_i^2 \longrightarrow \infty,$$

as $N, T \to \infty$, *then*

$$\sup_{0\le t\le 1} |\bar{V}_{N,T}(t)| \xrightarrow{\mathcal{P}} \infty.$$

Theorem 3.2 implies the null hypothesis is rejected when $\sup_{0\le t\le 1} |\bar{V}_{N,T}(t)|$ is large. We can compute the asymptotic critical values from Corollary 3.1, thus developing a test based on the value of $\sup_{0\le t\le 1} |\bar{V}_{N,T}(t)|$. The test is sensitive to fixed changes in relatively few panels, and at the same time it is also sensitive to relatively smaller changes in a large number of panels. However, in order to implement the test based on $\sup_{0\le t\le 1} |\bar{V}_{N,T}(t)|$, we have to estimate the long-run variance σ_i^2 first.

4 Estimation of Long-Run Variance for Panel Data

Since the long-run variances σ_i^2's in (3.3) are unknown, we need to find some suitable estimators for them. Hence we define

$$V_{N,T}(t) = \frac{1}{N^{1/2}} \sum_{i=1}^{N} \left\{ \frac{1}{\hat{\sigma}_i^2} Z_{T,i}^2(t) - \frac{\lfloor Tt \rfloor (T - \lfloor Tt \rfloor)}{T^2} \right\}, \quad 0 \le t \le 1, \tag{4.1}$$

with the long-run variance of $T^{-1/2} S_{T,i}(1)$ estimated by $\hat{\sigma}_T^2(i)$. We consider the following two cases:

(1) If the innovations $\{e_{i,j}, 1 \le j \le T\}$ are independent and identically distributed for all i, then we can use the sample variance

$$\hat{\sigma}_T^2(i) = \frac{1}{T-1} \sum_{j=1}^{T} (X_{i,j} - \bar{X}_T(i))^2 \tag{4.2}$$

$$\bar{X}_T(i) = \frac{1}{T} \sum_{j=1}^{T} X_{i,j}$$

to estimate the long-run variance of $T^{-1/2} S_{T,i}(1)$ in the i^{th} panel.

(2) If the independence cannot be assumed, then we use a kernel estimator (see [6]):

$$\hat{\sigma}_T^2(i) = \frac{1}{T} \sum_{j=1}^{T} (X_{i,j} - \bar{X}_T(i))^2 + 2 \sum_{l=1}^{T-1} K\left(\frac{\ell}{h}\right) \hat{\gamma}_{T,\ell}(i), \tag{4.3}$$

where

$$\hat{\gamma}_{T,\ell}(i) = \frac{1}{T-\ell} \sum_{j=1}^{T-\ell} (X_{i,j} - \bar{X}_T(i))(X_{i,j+\ell} - \bar{X}_T(i))$$

is the sample covariance of lag ℓ in the i^{th} panel. The function K is the kernel in the definition of $\hat{\sigma}_T^2(i)$ in (4.3) and $h = h(T)$ is the bandwidth (window). Later we will see in the next section that one should be careful with the choice of the bandwidth in the simulation study since the bandwidth choice is one of the hardest parts in the estimation.

Definition 4.1. *f is **Lipschitz continuous** on $[a, b]$ if*

$$|f(t) - f(s)| \le C|t - s| \text{ for all } a \le t, s \le b,$$

where $C = C(a, b)$.

Assumption 4.1. *We assume the following conditions on the kernel estimator:*

(i) $K(0) = 1$,

(ii) $K(u) = 0$ *if* $|u| > a$ *and* $K(u)$ *is Lipschitz continuous on* $[-a, a]$ *with some* $a > 0$,

(iii) K *has* ν *bounded derivatives in a neighborhood of 0 and the first* $\nu - 1$ *derivatives of* K *are 0 at 0, where* $\nu \geq 1$ *is an integer,*

(iv) $h = h(T) \to \infty$ *and* $h/T \to 0$ *as* $T \to \infty$.

From the discussions in section 3, we can see that in order to claim $V_{N,T}$ and $\bar{V}_{N,T}$ have the same asymptotic distribution, the estimator $\hat{\sigma}_T^2(i)$ must be very close to σ_i^2. Assumption *(iii)* above is needed to obtain a very small bias of the estimator $\hat{\sigma}_T^2(i)$. Horváth and Hušková (2012) showed in their paper that if the panels are dependent of each other, then even very small changes to the model in (3.1) will alter the asymptotic distribution for $\bar{V}_{N,T}$. Similarly, the long-run variance estimator $\hat{\sigma}_T^2(i)$ must be very close to σ_i^2 to claim that $V_{N,T}$ and $\bar{V}_{N,T}$ have the same asymptotic distribution. Further discussions on kernel estimators can be found in Taniguchi and Kakizawa (2000) and Brockwell and Davis (2006). Notice, that the "flat-top" Bartlett type kernel satisfies Assumption 4.1 for all $\nu \geq 1$. Therefore, we will use a "flat-top" kernel in the next section.

The next condition is on the connection between the number of panels (N), the length of the observed time series in each panel (T) and the bandwidth (h):

$$\frac{Nh^2}{T^2} \to 0 \text{ and } \frac{N^{1/2}}{h^\tau} \text{ where } \tau = \min(\nu, a - 1). \tag{4.4}$$

As in (3.10), assumption (4.4) allows for a short time series in a much larger number of panels.

Theorem 4.1. *If* H_0*, Assumption 3.1, Property 3.1, Assumption 4.1 hold and (3.7) and (3.8) are satisfied with* $\kappa = 8$*, then*

$$V_{N,T}(t) \xrightarrow{\mathcal{D}[0,1]} \Gamma(t), \tag{4.5}$$

where $\Gamma(t)$ *is defined in Theorem 3.1.*

5 Autoregressive Panels

Definition 5.1. *An* ***Autoregressive Model*** *of order p, abbreviated* ***AR(p)****, is of the form*

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t, \tag{5.1}$$

where x_t *is stationary, and* $\phi_1, \phi_2, \ldots, \phi_p$ *are constants* ($\phi_p \neq 0$).

Consider an AR(1) process defined by a recursion formula

$$x_t = \rho x_{t-1} + \varepsilon_t. \tag{5.2}$$

Iterating (5.2) backwards k times, we get

$$x_t = \rho x_{t-1} + \varepsilon_t$$

$$
\begin{aligned}
&= \rho^2 x_{t-2} + \rho x_{t-1} + \varepsilon_t \\
&\vdots \\
&= \rho^k x_{t-k} + \sum_{\ell=0}^{k-1} \rho^\ell \varepsilon_{t-\ell}.
\end{aligned}
$$

This suggests that if $|\rho| < 1$ and $E(\log(1+|\varepsilon_0|) < \infty$, the above formula has a stationary solution such that x_t is a function of $\varepsilon_t, \varepsilon_{t-1}, \ldots$, i.e., x_t is a linear process defined in (3.5)

$$
x_t = \sum_{\ell=0}^{\infty} \rho^\ell \varepsilon_{t-\ell}. \tag{5.3}
$$

Since $E(\log(1 + |\varepsilon_0|)) < \infty$ implies for all $c > 0$

$$
\limsup_{\ell \to \infty} \frac{|\varepsilon_\ell|}{c^\ell} = 0 \quad \text{a.s.},
$$

that is, $\forall \delta > 0, \exists \ell > \ell_0(\delta)$, such that $|\varepsilon_\ell| < c^\ell$, hence, choose c such that $|\rho c| < 1$

$$
\sum_{\ell=0}^{\infty} \rho^\ell \varepsilon_\ell = \sum_{\ell=0}^{\ell_0} \rho^\ell \varepsilon_\ell + \sum_{\ell=\ell_0+1}^{\infty} \rho^\ell \varepsilon_\ell \le \sum_{\ell=0}^{\ell_0} \rho^\ell \varepsilon_\ell + \sum_{\ell=\ell_0+1}^{\infty} |\rho|^\ell |\varepsilon_\ell| < \sum_{\ell=0}^{\ell_0} \rho^\ell \varepsilon_\ell + \sum_{\ell=\ell_0+1}^{\infty} |\rho c|^\ell < \infty. \tag{5.4}
$$

This implies (5.3) exists with probability 1, and it is straightforward to check this is a stationary solution. We can see that the process in (5.3) does not depend on the future, we say the process is causal. Furthermore, if we assume the error terms are independent identically distributed random variables, then

(i) $E[\varepsilon_i]$ exists implies $E[x_t]$ exists;

(ii) $E[\varepsilon_i^2]$ exists implies $E[x_t^2]$ exists.

Proof: From (5.3) and $|\rho| < 1$, we get

$$
E[x_t] = \sum_{\ell=0}^{\infty} \rho^\ell E[\varepsilon_{t-\ell}] < \infty
$$

and by the Lebesgue dominated convergence theorem,

$$
E[x_t^2] = E\left(\sum_{\ell=0}^{\infty} \rho^\ell \varepsilon_{t-\ell} \right)^2 = \sum_{\ell=0}^{\infty} \sum_{k=0}^{\infty} \rho^\ell \rho^k E[\varepsilon_{t-\ell} \varepsilon_{t-k}] < \infty.
$$

Next, we claim that if the error terms ε_i's are independent and identically distributed random variables with mean zero and $\mathrm{Var}(\varepsilon_i) = \sigma^2$, then by Donsker's Theorem (see [4]),

$$
\frac{1}{T^{1/2}} \sum_{\ell=0}^{\lfloor Tt \rfloor} \varepsilon_\ell \xrightarrow{\mathcal{D}[0,1]} \sigma W(t), \quad 0 \le t \le 1
$$

and

$$\frac{1}{T^{1/2}}\sum_{j=0}^{\lfloor Tt\rfloor} x_j \xrightarrow{\mathcal{D}[0,1]} \frac{\sigma}{1-\rho}W(t),$$

where $W(t)$ is the Wiener process.

6 A Simulation Study

In this section, we investigate the asymptotic behavior of $\sup_{0\le t\le 1}|V_{N,T}(t)|$ under Theorem 3.1 and Theorem 4.1, seeking to obtain the 90%, 95% and 99% critical values for this test statistic using Monte Carlo simulation. We are also interested to see how well the approximation will be if the number of panels (N) and sample size (T) are small or moderate. Similar results can be obtained for $\sup_{0\le t\le 1}|\bar{V}_{N,T}(t)|$, where $\bar{V}_{N,T}(t)$ is defined in (3.3).

Table 6.1 contains the asymptotic critical values z_α generated from standard Brownian Motion using the equation,

$$P\{\sup_{0\le t\le 1}|\Gamma(t)|\le z_\alpha\}=\alpha,$$

where $\alpha = 0.9, 0.95$ and 0.99 (see [6]).

Table 6.1: Asymptotic critical values of $\sup_{0\le t\le 1}|\Gamma(t)|$

90%	95%	99%
0.7956	0.8942	1.1452

Next, we use the following procedure to calculate the critical values of $\sup_{0\le t\le 1}|V_{N,T}(t)|$ in case of the errors being independent identically distributed standard normal, χ^2_5-5, and t_5, as well as they are AR(1) processes with $\rho = 0, 0.1, 0.3$ and 0.5, respectively.

- **Step 1:** Generate the panels under the null hypothesis, i.e., let $X_{i,j} = e_{i,j}$ for $\le i \le N, 1 \le j \le T$, where the $e_{i,j}$ follows the underlying distribution.
- **Step 2:** Estimate the long-run variances $\hat{\sigma}_i^2$ using the sample variance in (4.2) if the errors are i.i.d., or using the kernel estimator in (4.3) if the errors are dependent of each other.
- **Step 3:** Compute the maximum of the absolute value of $V_{N,T}(t)$ defined in (3.3) using the long-run variances $\hat{\sigma}_i^2$ estimated in Step 2.
- **Step 4:** Repeat Step (2) and (3) 1000 times, then compute the empirical distribution function

$$\hat{F}_n(t)=\frac{\text{number of elements in the sample}\le t}{n}=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}\{x_i\le t\}$$

based on the simulated sample. Thus, we obtain the limiting distribution of $\sup_{0\leq t\leq 1}|V_{N,T}(t)|$.

- **Step 5:** Obtain the critical values of the limiting distribution.

Table 6.2 shows the outcomes of the Monte Carlo experiments when the distribution of the errors are independent identically distributed standard normal, χ^2_5-5, and t_5, respectively. The long-run variances σ_i's are estimated by the sample variances in each case. The last two distributions allow us to see the effects of skewness and heavy-tailness on the limiting distribution. The critical values obtained from the simulations indicate that those effects are negligible. From Table 6.2, we can see the simulated critical values are very close to the asymptotic critical values in Table 6.1. Comparing Table 6.2 and Table 6.3 with $\rho = 0$ illustrates that the estimation of the long-run variance reduces the accuracy of the limiting results.

Table 6.2: Simulated critical values of $\sup\limits_{0<t<1}|V_{N,T}(t)|$ based on independent and identically distributed $N(0,1)$, χ^2_5-5 and t_5 errors.

	$N(0,1)$			χ^2_5-5			t_5		
N/T	90%	95%	99%	90%	95%	99%	90%	95%	99%
50/50	0.811	0.904	1.104	0.799	0.910	1.102	0.800	0.961	1.159
100/50	0.799	0.864	1.146	0.793	0.860	1.066	0.777	0.873	1.043
100/100	0.836	0.956	1.200	0.857	0.949	1.153	0.813	0.919	1.127
200/100	0.834	0.927	1.141	0.820	0.900	1.038	0.831	0.932	1.162

According to the discussions in Section 4, if the innovations are dependent, a "flat top" kernel function can be used to estimate $\hat{\sigma}^2_T(i)$. Hence, we can take the kernel function of the form

$$K(x) = \begin{cases} 1, & 0 \leq |x| < 1 \\ 2-|x|, & 1 \leq x < 2 \\ 0, & 2 \leq |x|. \end{cases} \tag{6.1}$$

After a couple of trials and errors, we choose $h = 1.5$ for $T = 50$ and $h = 1.7$ for $T = 100$ as the optimal windows, and estimate the long-run variance using (4.3) with the kernel function defined in (6.1). Table 6.3 summarizes the resulting critical values of AR(1) model with the choice of the parameter $\rho = 0$, $\rho = 0.1$, $\rho = 0.3$, and $\rho = 0.5$, respectively. Upon inspecting Table 6.3, we can notice that the simulated critical values will be farther from the asymptotic critical values with the increasing of dependence in the AR(1) model, as the series tends to be non-stationary if ρ is large.

The power of the test are given in Table 6.4, Table 6.5 and Table 6.6. The magnitude of the change is set to be independent uniform on $[-1/2, 1/2]$ or $[-1, 1]$ in all the panels as well as in 50% of the panels. We choose the change point at $\lfloor T/4 \rfloor$. The observed percentage of rejections

are reported at the 5% significance levels for independent standard normal errors (Table 6.4) and AR(1) process with $\rho = 0.1$ and $\rho = 0.3$ with independent standard normal innovations (Table 6.5 and Table 6.6). As we have discussed in the previous section, the empirical results indicate that our test is sensitive to small changes in a large number of panels and relatively large changes in only several panels.

Table 6.3: Simulated critical values of $\sup_{0<t<1} |V_{N,T}(t)|$ in case of AR(1) processes with standard normal errors.

	$\rho = 0$			$\rho = 0.1$			$\rho = 0.3$			$\rho = 0.5$		
N/T	90%	95%	99%	90%	95%	99%	90%	95%	99%	90%	95%	99%
50/50	0.926	1.105	1.274	0.866	0.989	1.250	0.894	0.998	1.247	1.194	1.299	1.581
100/50	1.003	1.151	1.517	0.945	1.071	1.306	0.956	1.063	1.272	1.318	1.467	1.752
100/100	0.887	1.037	1.222	0.848	0.960	1.162	0.983	1.131	1.397	1.497	1.647	1.998
200/100	0.904	1.044	1.235	0.877	0.961	1.136	1.065	1.187	1.379	1.728	1.876	2.153

Table 6.4: Empirical rejection percentage for $\sup_{0<t<1} |V_{N,T}(t)|$ at 5% significance level in case of independent standard normal errors when $k_0 = \lfloor T/4 \rfloor$.

	$U[-1/2, 1/2]$		$U[-1, 1]$	
N/T	50%	100%	50%	100%
50/50	20.4	56.6	92.9	100
100/50	31.9	85.6	99.8	100
100/100	88.2	100	100	100
200/100	99.6	100	100	100

Table 6.5: Empirical rejection percentage for $\sup_{0<t<1} |V_{N,T}(t)|$ at 5% significance level in case of AR(1) process with $\rho = 0.1$ and standard normal innovations when $k_0 = \lfloor T/4 \rfloor$.

	$U[-1/2, 1/2]$		$U[-1, 1]$	
N/T	50%	100%	50%	100%
50/50	20.5	57.4	92.2	100
100/50	25.2	78.1	99.9	100
100/100	90.4	100	100	100
200/100	99.7	100	100	100

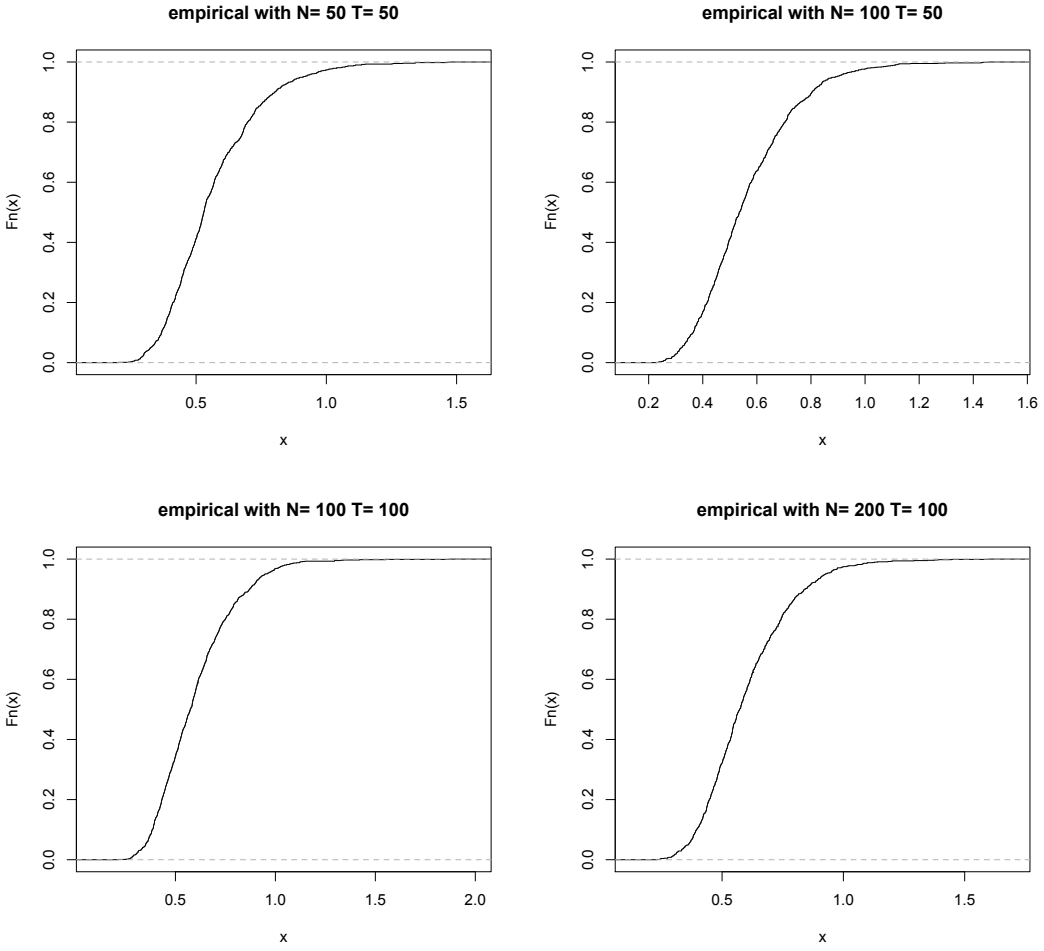


Figure 6.1: Empirical Distribution Function of $\sup_{0<t<1} |V_{N,T}(t)|$ with i.i.d. standard normal errors.

Table 6.6: Empirical rejection percentage for $\sup_{0<t<1} |V_{N,T}(t)|$ at 5% significance level in case of AR(1) process with $\rho = 0.3$ and standard normal innovations when $k_0 = \lfloor T/4 \rfloor$.

	$U[-1/2, 1/2]$		$U[-1, 1]$	
N/T	50%	100%	50%	100%
50/50	39.0	65.5	96.1	100
100/50	50.2	89.1	99.7	100
100/100	91.2	100	100	100
200/100	99.5	100	100	100

7 Results and Discussions

We are interested in whether or not there exists a common change in mean for all the panels using the test derived in previous sections. In case of 33 countries over 20 years in the dataset, we simulated the critical values for $\sup_{0<t<1} |V_{N,T}(t)|$ with N=33 and T=20. Since T=20 is small, the long-run variance estimator in (4.3) might not be very accurate. So we used the sample variance estimator in (4.2) as well as the kernel estimator in (4.3) with several choices of bandwidths. The resulted maximum values of $|V_{N,T}(t)|$ from our Gini dataset are 8.522 and 5.750, respectively. This means the null hypothesis is rejected at 0.01 significance level in our case; hence, we conclude there is a common change in the mean of each panel. Furthermore, the test detects that the change point, i.e., the location of the maximum value of $|V_{N,T}(t)|$, occurred at year 1992. A visual inspection on Figure 2.1 to Figure 2.4 supports the conclusion that there is an increase in the Gini coefficients in almost every country in our dataset.

Table 7.1: Simulated critical values of $\sup_{0<t<1} |V_{N,T}(t)|$ with N=33 and T=20 based on $N(0, 1)$, $\chi^2_5 - 5$ and t_5 errors.

Errors	90%	95%	99%
N(0, 1)	0.711	0.817	1.017
$\chi^2_5 - 5$	0.707	0.816	1.105
t_5	0.719	0.823	1.113

Table 7.2: Simulated critical values of $\sup_{0<t<1}|V_{N,T}(t)|$ with N=33 and T=20 based on AR(1) processes, the choice of the bandwidth for kernel estimator is 0.6.

AR(1) Process	90%	95%	99%
$\rho = 0$	0.735	0.831	1.053
$\rho = 0.1$	0.938	1.047	1.328
$\rho = 0.3$	1.455	1.612	1.914
$\rho = 0.5$	2.263	2.416	2.797

8 Darling-Erdős Limit Results

Consider the model in (3.1), the quasi-maximum likelihood approach of Bai (see [1]), yields the test statistic $\max_{1\leq k<T}|A_{N,T}(k)|$, where

$$A_{N,T}(k) = \frac{T}{k(T-k)}\frac{1}{N^{1/2}}\sum_{i=1}^{N}\left\{\frac{1}{\sigma_i^2}Z_{T,i}^2(k) - \frac{k(T-k)}{T}\right\}, \tag{8.1}$$

with

$$Z_{T,i}(k) = \sum_{j=1}^{k}(X_{i,j} - \bar{X}_{i,T}) \tag{8.2}$$

and $\bar{X}_{i,T}$ is sample mean of the i^{th} panel. Note that under the null hypothesis (3.2), $A_{N,T}(k)$ does not depend on the unknown $\mu_i's$, so under H_0 we have

$$A_{N,T}(k) = H_{N,T}(k) \text{ for all } 1 \leq k < T, \tag{8.3}$$

where

$$H_{N,T}(k) = \frac{T}{k(T-k)}\frac{1}{N^{1/2}}\sum_{i=1}^{N}\left\{\frac{1}{\sigma_i^2}\left(S_i(k) - \frac{k}{T}S_i(T)\right)^2 - \frac{k(T-k)}{T}\right\}, \tag{8.4}$$

with

$$S_i(k) = \sum_{j=1}^{k} e_{i,j}. \tag{8.5}$$

Assumption 8.1. *In this section, we consider the case when the panels are based on the independent observations with the following regularity conditions:*

(1) $e_{i,j}, (1 \leq i \leq N, -\infty < j < \infty)$ *are independent.*

(2) $Ee_{i,0} = 0,\ Ee_{i,0}^2 = 1.$

(3) for every i, $1 \leq i \leq N$, *the variables* $\{e_{i,j}, -\infty < j < \infty\}$ *are identically distributed.*

(4) $\limsup_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} Ee_{i,0}^4 < \infty$.

Since $H_{N,T}(k)$ does not depend on $\mathrm{Var}(e_{i,0})$, the second assumption is not a restriction on the model and therefore can be assumed without loss of generality. The distribution of the $e_{0,j}$'s can be very different and we only require the average of the fourth moments is bounded. Let

$$A(x) = (2\log x)^{1/2} \tag{8.6}$$

and

$$D(x) = 2\log x + \frac{1}{2}\log\log x - \frac{1}{2}\log\pi. \tag{8.7}$$

Due to (8.3) it is enough to consider $H_{N,T}(k)$ under the null hypothesis. Chan, Horváth and Hušková (2012) proved the following Darling-Erdős type results in their paper (see [3]). They also showed that the normalizing centering sequences $A(\log T^2)$ and $D(\log T^2)$ might not work well if the sample sizes are small or moderate via simulations. It is feasible to replace them with other sequences and the limit is still the double exponential extreme value distribution.

Theorem 8.1. *If Assumption 8.1 holds, then*

$$\lim_{\min(N,T)\to\infty} P\left\{\frac{A(\log T^2)}{\sqrt{2}} \max_{1\le k<T} |A_{N,T}(k)| \le t + D(\log T^2)\right\} = \exp(-2e^{-t}) \tag{8.8}$$

for all t.

Theorem 8.2. *If Assumption 8.1 holds, then we have for all* $\alpha \ge 0$

$$\lim_{\min(N,T)\to\infty} P\left\{\frac{A(\log(T^2/(\log T)^\alpha))}{\sqrt{2}} \max_{1\le k<T} |H_{N,T}(k)| \le t + D(\log(T^2/(\log T)^\alpha))\right\} = \exp(-2e^{-t}) \tag{8.9}$$

for all t.

Remark 8.1. *In order to provide more accurate approximations for the asymptotic distribution of* $\max_{1\le k<T}|H_{N,T}(k)|$ *under the null hypothesis in case of small sample sizes, we suggest to use the following formula. The tail approximation of the maximum of the absolute value of the Ornstein-Uhlenback process (cf. Theorem A.3.3 in Csörgő and Horváth (1997)) gives for all fixed T*

$$P\left\{\frac{1}{\sqrt{2}} \max_{1\le k<T} |H_{N,T}(k)| > x\right\} \approx \frac{x\exp(-x^2/2)}{\sqrt{\pi/2}}\left\{2\log(T/2) - \frac{2}{x^2}\log(T/2) + \frac{4}{x^2} + O\left(\frac{1}{x^4}\right)\right\}.$$

The other possibility is to compute the maxima of $H_{N,T}(k)$ on restricted intervals. We assume that

$$1 \le h_T \le T/2 \text{ and } h_T/T \to 0 \text{ as } T \to \infty. \tag{8.10}$$

Theorem 8.3. *If Assumption 8.1 and condition (8.10) hold, then we have*

$$\lim_{\min(N,T)\to\infty} P\left\{\frac{A(\log(T^2/h_T^2))}{\sqrt{2}}\max_{h_T\le k\le T-h_T}|H_{N,T}(k)|\le t+D(\log(T^2/h_T^2))\right\}=\exp(-2e^{-t}) \tag{8.11}$$

for all t.

Using Theorem 8.1 - 8.3 together with equation (8.3), we obtain the limiting distribution of the test statistic $\max_{1\le k<T}|A_{N,T}(k)|$ under H_0.

Theorem 8.4. *If Assumption 8.1 holds with* $\min(N,T)\to\infty$*,*

$$0<\liminf_{T\to\infty}\frac{k_0}{T}<\limsup_{T\to\infty}\frac{k_0}{T}<1 \tag{8.12}$$

and

$$\frac{T}{(N\log\log T)^{1/2}}\sum_{i=1}^{N}\frac{\delta_i^2}{\sigma_i^2}\longrightarrow\infty, \tag{8.13}$$

then we have

$$\frac{A(\log T^2)}{\sqrt{2}}\max_{1\le k<T}|A_{N,T}(k)|-D(\log T^2)\xrightarrow{\mathcal{P}}\infty. \tag{8.14}$$

We note that assumption (8.12) implies the change cannot occur too early or too late. This is a standard assumption in change point analysis. If the test statistic computed from the right-hand-side of (8.14) is large, then the null hypothesis is violated. Theorem 8.4 establishes the consistency of the testing procedure based on $\max_{1\le k<T}|A_{N,T}(k)|$.

Remark 8.2. *In applications, the value of the variances* σ_i^2*'s are unknown but they can be estimated with the sample variance* $\hat{\sigma}_i^2$*. The results in this section remain valid if* $A_{N,T}(k)$ *is replaced with*

$$\hat{A}_{N,T}(k)=\frac{T}{k(T-k)}\frac{1}{N^{1/2}}\sum_{i=1}^{N}\left\{\frac{1}{\hat{\sigma}_i^2}Z_{T,i}^2(k)-\frac{k(T-k)}{T}\right\}.$$

In several applications it is unreasonable to assume that the panels are based on independent observations. If the error terms are assumed to form a linear process,

$$e_{i,j}=\sum_{l=0}^{\infty}c_{i,l}\varepsilon_{i,j-l},\quad 1\le i\le N, 1\le j\le T, \tag{8.15}$$

where the variables $\varepsilon_{i,j}$ satisfy Assumption 8.1. In addition to the requirement that the average of the fourth moments is bounded, we also assume

$$\limsup_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}(E|\varepsilon_{i,0}|^{\gamma})^{1/2}<\infty \text{ with some } \gamma\ge 4. \tag{8.16}$$

The errors in each panel are stationary linear processes and their distributions depend on the panels. The coefficients in the definition of linear processes satisfy the same properties as in Property (3.1).

(a) $|c_{i,\ell}| \le c_0(\ell+1)^{-\alpha}$ for all $1 \le i \le N$, $1 \le j \le T$ with some c_0 and $\alpha > 2$,

(b) there is $\delta > 0$ such that $a_i^2 \ge \delta^2$ with $a_i = \sum_{\ell=0}^{\infty} c_{i,\ell}$ for all $1 \le i \le N$.

In the definition of $Z_{N,T}(k)$ now σ_i^2 is the long-run variance of the i^{th} panel, given by

$$\lim_{T\to\infty} \frac{1}{T} E\left(\sum_{j=1}^{T} e_{i,j}\right)^2 = \sigma_i^2, \quad 1 \le i \le N.$$

Under assumption 8.1 and (8.16) together with Property (3.1) σ_i^2 exists, $a_i^2 = \sigma_i^2$ and

$$\sigma_i^2 \ge \delta^2 \text{ for all } 1 \le i \le N, \tag{8.17}$$

i.e., we have a common lower bound, δ^2, for the long-run variances of each panel. The last condition is on the connection between the number of panels (N), the length of the observed time series in each panel (T) and the interval where the maximum is computed:

$$N^{1/2} h_T^{(2-\gamma)/(2\gamma)} (\log h_T)^{1+2/\gamma} (\log\log T)^{1/2} \to 0, \tag{8.18}$$

where γ is from assumption (8.16). Next, we show that Theorem 8.3 holds for $\max_{h_T \le k \le T-h_T} |A_{N,T}(k)|$, the proof of the theorem is based on Philips-Solo (1992) representation (see [3]). In applications the long-run variance σ_i^2 is unknown, we can estimate σ_i^2 with a Bartlett type estimator $\hat{\sigma}_i^2$ (see [6]).

Theorem 8.5. *We assume model (3.1) holds. If H_0, Assumption 8.1, (8.16), Property (3.1) and (8.18) are satisfied, then we have*

$$\lim_{\min(N,T)\to\infty} P\left\{ \frac{A(\log(T^2/h_T^2))}{\sqrt{2}} \max_{h_T \le k \le T-h_T} |A_{N,T}(k)| \le t + D(\log(T^2/h_T^2)) \right\} = \exp(-2e^{-t}) \tag{8.19}$$

for all t.

In this section we investigate, via Monte Carlo simulations, to see how well the empirical distribution

$$F_{N,T;\alpha}(t) = P\{L_{N,T;\alpha} \le t\}$$

can be approximated by the limiting distribution $\exp(-2\exp(-t))$, where functions A and D are defined in (8.6) and (8.7) and

$$L_{N,T;\alpha} = \frac{A(\log(T^2/(\log T)^\alpha))}{\sqrt{2}} \max_{1 \le k < T} |A_{N,T}(k)| - D(\log(T^2/(\log T)^\alpha)).$$

It has been observed that if the normalization and centering of Theorem 8.2 is used, then the approximation with the limit $\exp(-2\exp(-t))$ is better on the upper tail in case of small

and moderate sample sizes (see [3]). Modifying the normalizing and centering sequences as in Theorem 8.2 were suggested by Csörgő and Horváth (1997, Chapter 1).

We consider the power of the test very briefly here. The size of the changes are independent uniform on $[-1/2, 1/2]$ or $[-1, 1]$ in all the panels as well as in 50% of the panels. The powers of the test are reported at 5% significance level in case of independent standard normal errors when the change point $k_0 = \lfloor T/4 \rfloor$ (Table 8.1) and $k_0 = \lfloor T/2 \rfloor$ (Table 8.2). Upon inspecting Table 8.1 and 8.2, we notice that the test has more power when $k_0 = \lfloor T/2 \rfloor$ than $k_0 = \lfloor T/4 \rfloor$.

Table 8.1: Empirical rejection percentage for $L_{N,T;3}$ at 5% significance level in case of independent standard normal errors when $k_0 = \lfloor T/4 \rfloor$.

	$U[-1/2, 1/2]$			$U[-1, 1]$		
N/T	25%	50%	100%	25%	50%	100%
50/50	7.0	20.2	61.3	59.2	96.0	100
100/50	11.2	32.9	90.6	83.6	99.9	100
100/100	38.9	90.3	100	99.6	100	100
200/100	67.0	99.5	100	100	100	100

Table 8.2: Empirical rejection percentage for $L_{N,T;3}$ at 5% significance level in case of independent standard normal errors when $k_0 = \lfloor T/2 \rfloor$.

	$U[-1/2, 1/2]$			$U[-1, 1]$		
N/T	25%	50%	100%	25%	50%	100%
50/50	13.1	38.2	86.3	78.9	99.6	100
100/50	18.1	60.3	99.2	95.8	100	100
100/100	40.5	89.9	100	99.6	100	100
200/100	65.3	99.5	100	100	100	100

We now test the null hypothesis on our Gini index dataset using the Darling-Erdős Limit Results developed in the previous sections. Not surprisingly, the test also rejects the null hypothesis at 0.01 significance level in favor of the alternative that there is a common change in the Gini index. It is interesting to compare the powers obtained in Table 6.4 and Table 8.1, under the same conditions when the change point $k_0 = \lfloor T/4 \rfloor$. The test derived from Darling-Erdős limit results approach is less efficient in detecting the instability than the test obtained from the quasi-maximum likelihood method, especially in case of small changes with small or moderate sample sizes.

9 Appendix

9.1 Code Written for Data Extraction and Simulation

9.1.1 Data reshaping

This code was written in SAS and takes care of data reshaping. It converts the dataset from long version to wide version.

```
libname income 'C:\Users\u0637206\Desktop\Income study\WIID2C.xls';
options validvarname=any;

data all;
set income."WIID2C$"n; *read dataset into SAS;
run;

proc print data=all;
where year=.;
var country3 year gini;
run;

data country3_yr_gini;
set all;
keep country3 year gini;*keep 3 variables;
run;

proc print data=country3_yr_gini;
run;

proc sort data=country3_yr_gini out=gini_sorted;
by country3 year;
run;

data income1;
set gini_sorted;
by country3 year;*must sort first;
if year=. then year=.z;*missing year;
firstyr=first.year;
lastyr=last.year;
firstcountry3=first.country3;
lastcountry3=last.country3;
run;
```

```
proc print data=income1;
var country3 year gini firstyr lastyr;
run;

***find the average gini within each year group;
data income2;
set income1;
if firstyr then do;
totalgini=0;
count=0;
end;
count+1;
totalgini+gini;
if lastyr then do;
average_gini=totalgini/count;
output;
end;
run;

***reshape data from long to wide;
proc transpose data=income2 out=income_final(drop=_name_);
by country3;
id year;
var gini;
run;

***export to csv file;
ods csv file="C:\Users\u0637206\Desktop\Income study\Gini50.csv";

proc print data=income_final;
run;

ods csv close;
```

9.1.2 Data extraction and cleaning

The following R scripts did the job of extraction and replacing missing values in the dataset.

```
mydata=read.csv("/users/jessica/Desktop/MSTAT PROJECT/giniall.csv",
header=T,sep=',',na.strings=".")
```

```
new=mydata[,2:161]

gini=ts(new[60:79,],start=c(1987),end=c(2006),frequency=1)

#count number of non-missing values
n=rep(0,ncol(gini))
for (i in 1:ncol(gini)){
count=0
for (j in 1:nrow(gini)){
if (is.na(gini[j,i])==FALSE) count=count+1
}
n[i]=count
}

totnumber=data.frame(colnames(gini),n)
index=which(n>=16,arr.ind=T)
final.dataset=gini[,index]
final=t(final.dataset)
dim(final.dataset)
colnames(final)=c(seq(1987,2006,1))
final

##American Countries, replace missing values##
library(zoo)

american=c(7,9,16,19,20)
z=window(final.dataset[,american])
#United States, Argentina, Brazil, Costa.Rica, Venezuela##
plot(z,plot.type=c("single"),col=c("red","blue","green","purple",'orange'),
ylab="Gini",xlab="Year")
#text(z,colnames(z),pos=4)
z[,1]=na.locf(z[,1])##US
lines(z[,1],type="l",lty=3,col="red")
z[,3]=c(na.approx(z[,3]),z[19,3])##Brazil
lines(z[,3],type="l",lty=3,col="green")
z[,4]=c(rep(z[3,4],2),na.approx(z[,4]))##Costa.Rica
lines(z[,4],type="l",lty=3,col="purple")
z[,5]=na.locf(z[,5])#Venezuela
lines(z[,5],type="l",lty=3,col="orange")
legend("topright",colnames(z),col=c("red","blue","green","purple",'orange'),
```

```
lty=1,cex=0.6)

##Asian &Oceanian Countries, replace missing values##
asian.oceanian=c(6,10,11,25)
z2=window(final.dataset[,asian.oceanian])
plot(z2,plot.type=c("single"),col=c("red","blue","green","purple",'orange'),
ylab="Gini",xlab="Year")
z2[,1]=c(na.approx(z2[,1]),rep(z2[18,1],2))##Australia
lines(z2[,1],type="l",lty=3,col="red")
z2[,2]=na.locf(z2[,2])##China
lines(z2[,2],type="l",lty=3,col="blue")
z2[,3]=na.locf(z2[,3])##Taiwan
lines(z2[,3],type="l",lty=3,col="green")
z2[,4]=c(z2[2,4],na.approx(z2[,4]))##Kyrgyz.Republic
lines(z2[,4],type="l",lty=3,col="purple")
legend("topright",colnames(z2),col=c("red","blue","green","purple",'orange'),
lty=1,cex=0.6)

##European Countries##

european=c(1,2,3,4,5,8,12,13,14,15,17,18,21,22,23,24,26,27,28,29,30,31,32,33)

eastern.european1=c(14,22,23,24,26)

z3=window(final.dataset[,eastern.european1])
plot(z3,plot.type=c("single"),col=c("red","blue","green","purple",'orange'),
ylab="Gini",xlab="Year")
z3[,2]=c(z3[2,2],na.approx(z3[,2]))#Ukraine
lines(z3[,2],typ="l",lty=3,col="blue")
z3[,3]=c(z3[2,3],na.approx(z3[,3]))#Belarus
lines(z3[,3],typ="l",lty=3,col="green")
z3[,4]=c(z3[2,4],na.approx(z3[,4]))#Estonia
lines(z3[,4],typ="l",lty=3,col="purple")
z3[,5]=c(z3[2,5],na.approx(z3[,5]))#Latvia
lines(z3[,5],typ="l",lty=3,col="orange")
legend("bottomright",colnames(z3),col=c("red","blue","green","purple",'orange'),
lty=1,cex=0.6)

eastern.european2=c(27,28,29,30,33)
z4=window(final.dataset[,eastern.european2])
```

```
plot(z4,plot.type=c("single"),col=c("red","blue","green","purple",'orange'),
ylab="Gini",xlab="Year")
z4[,1]=c(z4[2,1],na.approx(z4[,1]))
lines(z4[,1],typ="l",lty=3,col="red")
z4[,2]=c(z4[2,2],na.approx(z4[,2]))
lines(z4[,2],typ="l",lty=3,col="blue")
z4[,3]=c(z4[2,3],na.approx(z4[,3]))
lines(z4[,3],typ="l",lty=3,col="green")
z4[,4]=c(z4[2,4],na.approx(z4[,4]))
lines(z4[,4],typ="l",lty=3,col="purple")
z4[,5]=c(rep(z4[3,5],2),na.approx(z4[,5]))
lines(z4[,5],typ="l",lty=3,col="orange")
legend("bottomright",colnames(z4),col=c("red","blue","green","purple",'orange'),
lty=1,cex=0.6)

western.european=c(1,3,4,21,32)
z5=window(final.dataset[,western.european])
plot(z5,plot.type=c("single"),col=c("red","blue","green","purple",'orange'),
ylab="Gini",xlab="Year")
z5[,1]=na.approx(z5[,1])
lines(z5[,1],typ="l",lty=3,col="red")
z5[,3]=na.approx(z5[,3])
lines(z5[,3],typ="l",lty=3,col="green")
z5[,4]=na.approx(z5[,4])
lines(z5[,4],typ="l",lty=3,col="purple")
z5[,5]=c(rep(z5[3,5],2),na.approx(z5[,5]))
lines(z5[,5],typ="l",lty=3,col="orange")
legend("bottomright",colnames(z5),col=c("red","blue","green","purple",'orange'),
lty=1,cex=0.6)

central.european=c(12,13,17,18,31)
z6=window(final.dataset[,central.european])
plot(z6,plot.type=c("single"),col=c("red","blue","green","purple",'orange'),
ylab="Gini",xlab="Year")
z6[,4]=na.approx(z6[,4])
lines(z6[,4],typ="l",lty=3,col="purple")
z6[,5]=na.approx(z6[,5])
lines(z6[,5],typ="l",lty=3,col="orange")
legend("bottomright",colnames(z6),col=c("red","blue","green","purple",'orange'),
lty=1,cex=0.6)
```

```
northern.european=c(2,5,8,15)
z7=window(final.dataset[,northern.european])
plot(z7,plot.type=c("single"),col=c("red","blue","green","purple",'orange'),
ylab="Gini",xlab="Year")
z7[,2]=na.approx(z7[,2])
lines(z7[,2],type="l",lty=3,col="blue")
legend("topright",colnames(z7),col=c("red","blue","green","purple",'orange'),
lty=1,cex=0.6)

fixed.dataset=final.dataset
fixed.dataset[,northern.european]=z7
fixed.dataset[,central.european]=z6
fixed.dataset[,western.european]=z5
fixed.dataset[,eastern.european2]=z4
fixed.dataset[,eastern.european1]=z3
fixed.dataset[,asian.oceanian]=z2
fixed.dataset[,american]=z
fixed=t(fixed.dataset)
colnames(fixed)=c(seq(1987,2006,1))
fixed
```

9.1.3 Simulation for $N(0,1)$, $\chi^2_5 - 5$ and t_5 errors

```
V.fun=function(N,T,t,X){
S1=rowSums(X[,1:T]) # total sum of each row, N by 1 vector
S=rep(0,N)
Z=rep(0,N)
sample.var=rep(0,N)
sum=0
for (i in 1:N) {
S[i]=ifelse(floor(T*t)==0,0,sum(X[i,1:floor(T*t)]))
Z[i]=1/sqrt(T)*(S[i]-floor(T*t)/T*S1[i])
sample.var[i]=var(X[i,])
sum=sum+Z[i]^2/sample.var[i]-floor(T*t)*(T-floor(T*t))/T^2
#use sample variance for sigma^2
}
V=1/sqrt(N)*sum
return(V)
}
```

```
sup.V=function(N,T,times){
sup.V=rep(0,times)
for (j in 1:times) {
X=matrix(,N,T)
for (i in 1:N){
X[i,]=rnorm(T,mean=0,sd=1)
}
for (k in 1:length(t)) {
V[k]=abs(V.fun(N,T,t[k],X))
}
sup.V[j]=max(V)
}
return(sup.V)
}
```

9.1.4 Simulation for AR(1) processes

```
ARsup.V=function(N,T,times,h){  ##compute the sup of V, based N(0,1) errors
sup.V=rep(0,times)
for (j in 1:times) {
X=matrix(,N,T)
sigmahat2=rep(0,N)
for (i in 1:N){
X[i,]=arima.sim(list(ar=rho),n=T)
sigmahat2[i]=kernel.sigmahat2(T,X[i,],h)
##using kernel function
}
for (k in 1:length(t)) {
V[k]=abs(V.fun(N,T,t[k],X,sigmahat2))
}
sup.V[j]=max(V)
}
rm(j,X,sigmahat2,k,V)
return(sup.V)
}

V.fun=function(N,T,t,X,sigmahat2){
##using estimated sigma^2 to compute V, a function of t.##
S1=rowSums(X[,1:T]) # total sum of each row, N by 1 vector
```

```
S=rep(0,N)
Z=rep(0,N)

sum=0
for (i in 1:N) {
S[i]=ifelse(floor(T*t)==0,0,sum(X[i,1:floor(T*t)]))
Z[i]=1/sqrt(T)*(S[i]-floor(T*t)/T*S1[i])
sum=sum+Z[i]^2/sigmahat2[i]-floor(T*t)*(T-floor(T*t))/T^2
#use estimated variance for sigmahat^2
}
V=1/sqrt(N)*sum
return(V)
}

gamma.hat=function(T,l,Xi){
sum=0
mu=mean(Xi)
for (j in 1:(T-l)){
sum=sum+(Xi[j]-mu)*(Xi[j+l]-mu)
}
sum=sum/(T-l)
return(sum)
}

kernel.sigmahat2=function(T,Xi,h){  ###estimeated variance of T^-1/2S_T,i(1)
sum=0
for (l in 1:(T-1)){
sum=sum+kernel(l/h)*gamma.hat(T,l,Xi)
}
sum=sum*2+var(Xi)*(T-1)/T
return(sum)
}

kernel=function(x){   ##a 'flat top' kernel
k=0
if (abs(x)<=1) k=1
else if (x>-2 & x< -1) k=x+2
else if (x>1 & x<2) k=2-x
return(k)
}
```

9.1.5 Estimate the long-run variance

```
T=20
N=33
times=1000
h=0.6
X=fixed;

sigmahat2=rep(0,N)
for (i in 1:N){
sigmahat2[i]=kernel.sigmahat2(T,X[i,],h)
##using kernel function to estimate the long-run variance
}

for (i in 1:length(t)){
ti=t[i]
V[i]=V.fun(N,T,ti,X,sigmahat2)
}
max(abs(V))
which.max(abs(V))
```

9.1.6 The power of the test

```
###power of the test according to Uniform distribution,
assume sigma_i=1 and change point is at T/4###

#generate a Unif[-1,1] sample when all the panels
have a common change in mean at T/4##
#generate a Unif[-1/2,1/2] sample when 50\% of the
 panels have a common change in mean at T/4##

sample.normal=function(min,max,N,T,k0,proportion){
e=matrix(rnorm(N*T,0,1),nrow=N) # random errors distributed as N(0,1)
# magnitude of changes in N panels, a uniformly distributed N by 1 vector
magnitude=c(runif(N,min,max))*rbinom(N,size=1,proportion)
Y=matrix(rep(0,N*T),nrow=N)  #initialization: zero matrix
Y[,1:k0]=e[,1:k0] #no change up to time k0
Y[,(k0+1):T]=matrix(rep(magnitude,(T-k0)),nrow=N,byrow=F)+e[,(k0+1):T]
#change at time k0+1
```

```
return(Y)
}

V.fun=function(N,T,t,X){  #compute V for each given t value and sample
S1=rowSums(X[,1:T]) # total sum of each row, N by 1 vector
S=rep(0,N)
Z=rep(0,N)
sample.var=rep(0,N)
sum=0
for (i in 1:N) {
S[i]=ifelse(floor(T*t)==0,0,sum(X[i,1:floor(T*t)]))
Z[i]=1/sqrt(T)*(S[i]-floor(T*t)/T*S1[i])
sample.var[i]=var(X[i,])
sum=sum+Z[i]^2/sample.var[i]-floor(T*t)*(T-floor(T*t))/T^2
#use sample variance for sigma^2
}
V=1/sqrt(N)*sum
return(V)
}

##power, use sample variance##
power.normal=function(N,T,times,critical){
rej.count=0 #counter initialization
for (j in 1:times){
# generate the sample each time assuming null hypothesis is false
sample1=sample.normal(min,max,N,T,k0,proportion)
maxV=0 #initialization
for (k in 1:length(t)){
maxV=max(maxV,abs(V.fun(N,T,t[k],sample1)))
#caluate a new |V| and compare
}
rej.count=ifelse(maxV>critical,rej.count+1,rej.count)
}
power=rej.count/times
return(power)
}

sample.ar=function(min,max,N,T,k0,proportion){
e=matrix(arima.sim(list(ar=rho),n=N*T),nrow=N) # random errors distributed as AR(1)
# magnitude of changes in N panels, a uniformly distributed N by 1 vector
```

```
magnitude=c(runif(N,min,max))*rbinom(N,size=1,proportion)
Y=matrix(rep(0,N*T),nrow=N)  #initialization: zero matrix
Y[,1:k0]=e[,1:k0] #no change up to time k0
Y[,(k0+1):T]=matrix(rep(magnitude,(T-k0)),nrow=N,byrow=F)+e[,(k0+1):T]
#change at time k0+1
return(Y)
}

V.fun=function(N,T,t,X,sigmahat2){
##using estimated sigma^2 to compute V, a function of t.##
S1=rowSums(X[,1:T]) # total sum of each row, N by 1 vector
S=rep(0,N)
Z=rep(0,N)
sum=0
for (i in 1:N) {
S[i]=ifelse(floor(T*t)==0,0,sum(X[i,1:floor(T*t)]))
Z[i]=1/sqrt(T)*(S[i]-floor(T*t)/T*S1[i])
sum=sum+Z[i]^2/sigmahat2[i]-floor(T*t)*(T-floor(T*t))/T^2
}
V=1/sqrt(N)*sum
return(V)
}

##power, use sigmahat2=1##
power.ar=function(N,T,times,critical,sigmahat2){
rej.count=0 #counter initialization
for (j in 1:times){
# generate the sample each time assuming null hypothesis is false
sample1=sample.ar(min,max,N,T,k0,proportion)
maxV=0 #initialization
for (k in 1:length(t)){
maxV=max(maxV,abs(V.fun(N,T,t[k],sample1,sigmahat2)))
#caluate a new |V| and compare
}
rej.count=ifelse(maxV>critical,rej.count+1,rej.count)
}
power=rej.count/times
return(power)
}
```

References

[1] Bai, J. (2010). *Common breaks in means and variances for panel data*, Journal of Econometrics, 157 (2010), 78-92.

[2] Baltagi, B. (1988). *Econometric Analysis of Panel Data*, 2nd edition, Springer.

[3] Chan, J., Horváth, L and Hušková, M. (2012). *Darling-Erdös Limit Results for Change-point Detection in Panel Data*

[4] DasGupta, A. (2008). *Asymptotic Theory of Statistics and Probability*, Springer.

[5] Frees, E. (2004) *Longitudinal and Panel Data*, Cambridge University Press.

[6] Horváth, L. and Hušková, M. (2011). *Change-point Detection in Panel Data*, J. Time Series Analysis, To appear (2012+).

[7] Hsiao, C. (2003). *Analysis of Panel Data*, 2nd edition, Cambridge University Press.

[8] *World Income Inequality Database*, User Guide and Data Sources.

[9] Shumway, R. Stoffer, D. (2011) *Time Series Analysis and Its Applications*, 3rd edition, Springer.

Table 9.1: Gini coefficients in percentage points of 33 selected countries from 1987 to 2006

	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
United.Kingdom	30.00	30.39	32.10	33.50	34.80	33.80	33.60	34.55	33.63	32.25	31.90	32.65	33.97	33.05	32.60	34.26	34.00		34.00	32.00
Sweden	28.24	25.87	24.48	26.09	24.47	29.35	24.85	27.90	25.45	25.20	25.40	24.20	26.00	28.20	26.10	25.70	25.22	23.00	23.00	23.00
Germany	26.90	30.92	29.03	25.22	25.25	25.67	28.49	29.20	28.69	27.83	28.82	27.21	27.34	27.77	27.39	29.57	29.26	29.21	26.00	27.00
Netherlands	30.27	29.00	29.60	28.95	31.46	28.60	28.80	29.23	28.50	29.43	27.97	28.45	29.50	29.00	25.80	27.00	27.00		27.00	26.00
Denmark	27.63	35.60	30.70	34.17	39.00	31.50	38.00	36.03	31.76		30.83	36.60	31.50	37.10	30.50	36.80	25.00	24.00	24.00	24.00
Australia	27.95	30.60	31.10	37.66	30.05	30.65	30.85	30.40	38.94	34.83	34.75	39.19	33.15	35.45	32.13	30.90		29.26		
United.States	40.25	40.50	40.90	40.40	39.12	40.65	42.30	40.73	41.80	42.10	42.27	45.30	45.50	41.73	46.30	46.20	46.40	46.41		
Finland	27.50	28.20	28.60	26.85	26.14	27.30	28.93	29.02	27.63	28.50	29.42	30.04	31.00	30.16	30.96	31.18	32.08	25.00	26.00	26.00
Argentina	43.30	45.38	47.60	44.40	45.86	44.72	44.43	45.16	47.76	47.76	47.17	49.41	49.09	50.43	52.21	53.26	52.84	50.82	50.19	48.56
China	24.52	28.43	26.60	28.72	27.59	30.88	29.50	31.50	31.55	31.78	28.42	36.65	35.00	32.44	44.76	38.09	37.09	46.90		
Taiwan	29.60	30.10	30.10	30.90	30.56	31.10	31.40	31.70	30.07	31.50	31.47	32.00	31.90	31.70	34.50	34.10	33.90			
Hungary	22.50	26.80	24.43	29.20	29.26	30.93	25.50	29.61	24.25	24.37	28.09	24.82	28.10	24.96	32.14	24.57	25.25	27.40	27.95	29.60
Poland	24.15	22.37	24.04	26.30	25.04	28.61	28.45	30.33	31.41	32.60	32.90	31.58	32.12	34.18	33.97	34.35	35.19	35.85	36.30	33.50
Bulgaria	18.80	20.90	22.00	22.07	26.20	29.88	29.01	35.30	34.61	29.95	33.59	32.14	30.91	30.78	37.01	34.16	32.23	35.80	33.80	31.00
Norway	31.90	29.76	31.18	31.16	27.50	28.03	24.50	29.67	25.70	31.55	31.90	30.90	31.05	30.90	30.50	33.15	27.00	25.00	28.00	30.00
Brazil	59.10	61.00	62.10	60.44		56.44	59.08		58.19	58.31	59.61	59.53	58.54		60.00	58.30	57.60	56.63	56.43	
Czech.Republic	19.80	19.25	19.51	19.91	21.53	22.79	24.88	24.02	24.84	25.04	24.27	24.20	24.76	25.41	25.00	25.27	25.05	25.15	26.27	24.60
Slovak.Republic	19.40	19.20	19.92	19.80	20.65	23.52	20.40			25.64	23.22	25.63	21.86	24.33	26.23	25.95	25.51	25.40	26.00	26.10
Costa.Rica			45.36	45.83	45.78	46.37	45.07	48.90	45.14	47.09	46.60	47.06		47.94	49.88	49.84	48.99	47.95	47.19	49.17
Venezuela	41.69	42.50	42.18	41.05	41.31	40.58	40.20	54.48	47.99	52.08	50.77	47.37	46.75	44.95	46.39	47.52	46.21	45.41	47.63	
Spain	25.30	27.95	26.80	32.90				35.00	35.15	25.64	34.90	33.12	33.38	32.92	32.55	32.20	31.95	31.25	31.84	31.00
Ukraine		24.05	24.86	24.30	21.55	23.80	36.40		45.78	37.85	35.95	39.10	34.48	41.25	40.80	37.25	40.80	41.00	28.24	41.00
Belarus		23.45	23.37	23.30		34.10	39.90		29.85	29.58	29.38	30.10	29.84	28.88	29.29	29.37	28.88	29.30	28.20	30.00
Estonia		25.40	27.73	24.00		35.50	37.53	37.75	36.44	34.51	35.68	36.91	37.34	36.80	37.11	36.03	35.23	37.65	34.05	33.00
Kyrgyz.Republic		28.50	27.23	30.80		30.00	47.58	44.30	39.50	51.63	42.65	41.73	45.19	42.26	43.81	42.46	41.52	42.65	43.40	42.85
Latvia		23.75	25.93	24.00	24.70	33.30	27.65	32.50	31.33	30.67	34.15	33.38	32.57	34.35	32.20	33.45	34.55	35.60	36.00	39.00
Lithuania		23.40	26.70	24.80		37.20	33.30	37.11	35.37	34.43	33.48	33.83	34.79	34.65	36.37	36.45	35.85	35.15	36.00	35.00
Moldova		25.30	25.30	26.70		37.75	40.10	37.90	39.00	41.40	41.83	42.60	44.10	41.45	41.30	43.10	39.15	38.20	40.35	35.60
Russian.Federation		25.15	26.65	25.43	31.15	42.95	47.90	44.10	44.52	46.17	37.57	40.96		43.67	47.15	40.05		46.90	44.50	45.10
Armenia		28.00	25.85	26.90	29.60	35.50	36.60	32.10	38.10	53.16		47.35		48.60		35.90	50.87	45.50	43.40	40.00
Slovenia	21.70		22.05	23.00	26.90	26.00	26.67	24.72	29.57	26.61	25.58	25.65	25.57	25.91	25.83	25.37	25.63	30.30	25.33	27.35
Macedonia..FYR			32.22	28.60	26.70	23.50	27.20	29.60	31.43	35.61	32.96	28.64	32.18	31.42	32.45	32.38	30.72	30.25	35.00	35.70
Romania			21.77	22.90	23.32	26.13	23.90	28.89	30.26	30.50	31.27	32.61	32.95	35.44	37.05	35.56	35.50	36.30	37.80	38.30

Table 9.2: Gini coefficients in percentage points of 33 selected countries from 1987 to 2006 with missing value replaced

	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006
United.Kingdom	30.00	30.39	32.10	33.50	34.80	33.80	33.60	34.55	33.63	32.25	31.90	32.65	33.97	33.05	32.60	34.26	34.00	34.00	34.00	32.00
Sweden	28.24	25.87	24.48	26.09	24.47	29.35	24.85	27.90	25.45	25.20	25.40	24.20	26.00	28.20	26.10	25.70	25.22	23.00	23.00	23.00
Germany	26.90	30.92	29.03	25.22	25.25	25.67	28.49	29.20	28.69	27.83	28.82	27.21	27.34	27.77	27.39	29.57	29.26	29.21	26.00	27.00
Netherlands	30.27	29.00	29.60	28.95	31.46	28.60	28.80	29.23	28.50	29.43	27.97	28.45	29.50	29.00	25.80	27.00	27.00	27.00	27.00	26.00
Denmark	27.63	35.60	30.70	34.17	39.00	31.50	38.00	36.03	31.76	31.30	30.83	36.60	31.50	37.10	30.50	36.80	25.00	24.00	24.00	24.00
Australia	27.95	30.60	31.10	37.66	30.05	30.65	30.85	30.40	38.94	34.83	34.75	39.19	33.15	35.45	32.13	30.90	30.08	29.26	29.26	29.26
United.States	40.25	40.50	40.90	40.40	39.12	40.65	42.30	40.73	41.80	42.10	42.27	45.30	45.50	41.73	46.30	46.20	46.40	46.41	46.41	46.41
Finland	27.50	28.20	28.60	26.85	26.14	27.30	28.93	29.02	27.63	28.50	29.42	30.04	31.00	30.16	30.96	31.18	32.08	25.00	26.00	26.00
Argentina	43.30	45.38	47.60	44.40	45.86	44.72	44.43	45.16	47.76	47.76	47.17	49.41	49.09	50.43	52.21	53.26	52.84	50.82	50.19	48.56
China	24.52	28.43	26.60	28.72	27.59	30.88	29.50	31.50	31.55	31.78	28.42	36.65	35.00	32.44	44.76	38.09	37.09	46.90	46.90	46.90
Taiwan	29.60	30.10	30.10	30.90	30.56	31.10	31.40	31.70	30.07	31.50	31.47	32.00	31.90	31.70	34.50	34.10	33.90	33.90	33.90	33.90
Hungary	22.50	26.80	24.43	29.20	29.26	30.93	25.50	29.61	24.25	24.37	28.09	24.82	28.10	24.96	32.14	24.57	25.25	27.40	27.95	29.60
Poland	24.15	22.37	24.04	26.30	25.04	28.61	28.45	30.33	31.41	32.60	32.90	31.58	32.12	34.18	33.97	34.35	35.19	35.85	36.30	33.50
Bulgaria	18.80	20.90	22.00	22.07	26.20	29.88	29.01	35.30	34.61	29.95	33.59	32.14	30.91	30.78	37.01	34.16	32.23	35.80	33.80	31.00
Norway	31.90	29.76	31.18	31.16	27.50	28.03	24.50	29.67	25.70	31.55	31.90	30.90	31.05	30.90	30.50	33.15	27.00	25.00	28.00	30.00
Brazil	59.10	61.00	62.10	60.44	58.44	56.44	59.08	58.63	58.19	58.31	59.61	59.53	58.54	59.27	60.00	58.30	57.60	56.63	56.43	56.43
Czech.Republic	19.80	19.25	19.51	19.91	21.53	22.79	24.88	24.02	24.84	25.04	24.27	24.20	24.76	25.41	25.00	25.27	25.05	25.15	26.27	24.60
Slovak.Republic	19.40	19.20	19.92	19.80	20.65	23.52	20.40	22.15	23.89	25.64	23.22	25.63	21.86	24.33	26.23	25.95	25.51	25.40	26.00	26.10
Costa.Rica	45.36	45.36	45.36	45.83	45.78	46.37	45.07	48.90	45.14	47.09	46.60	47.06	47.50	47.94	49.88	49.84	48.99	47.95	47.19	49.17
Venezuela	41.69	42.50	42.18	41.05	41.31	40.58	40.20	54.48	47.99	52.08	50.77	47.37	46.75	44.95	46.39	47.52	46.21	45.41	47.63	47.63
Spain	25.30	27.95	26.80	32.90	33.42	33.95	34.47	35.00	35.15	25.64	34.90	33.12	33.38	32.92	32.55	32.20	31.95	31.25	31.84	31.00
Ukraine	24.05	24.05	24.86	24.30	21.55	23.80	36.40	41.09	45.78	37.85	35.95	39.10	34.48	41.25	40.80	37.25	40.80	41.00	28.24	41.00
Belarus	23.45	23.45	23.37	23.30	28.70	34.10	39.90	34.87	29.85	29.58	29.38	30.10	29.84	28.88	29.29	29.37	28.88	29.30	28.20	30.00
Estonia	25.40	25.40	27.73	24.00	29.75	35.50	37.53	37.75	36.44	34.51	35.68	36.91	37.34	36.80	37.11	36.03	35.23	37.65	34.05	33.00
Kyrgyz.Republic	28.50	28.50	27.23	30.80	30.40	30.00	47.58	44.30	39.50	51.63	42.65	41.73	45.19	42.26	43.81	42.46	41.52	42.65	43.40	42.85
Latvia	23.75	23.75	25.93	24.00	24.70	33.30	27.65	32.50	31.33	30.67	34.15	33.38	32.57	34.35	32.20	33.45	34.55	35.60	36.00	39.00
Lithuania	23.40	23.40	26.70	24.80	31.00	37.20	33.30	37.11	35.37	34.43	33.48	33.83	34.79	34.65	36.37	36.45	35.85	35.15	36.00	35.00
Moldova	25.30	25.30	25.30	26.70	32.23	37.75	40.10	37.90	39.00	41.40	41.83	42.60	44.10	41.45	41.30	43.10	39.15	38.20	40.35	35.60
Russian.Federation	25.15	25.15	26.65	25.43	31.15	42.95	47.90	44.10	44.52	46.17	37.57	40.96	42.31	43.67	47.15	40.05	43.47	46.90	44.50	45.10
Armenia	28.00	28.00	25.85	26.90	29.60	35.50	36.60	32.10	38.10	53.16	50.25	47.35	47.98	48.60	42.25	35.90	50.87	45.50	43.40	40.00
Slovenia	21.70	21.88	22.05	23.00	26.90	26.00	26.67	24.72	29.57	26.61	25.58	25.65	25.57	25.91	25.83	25.37	25.63	30.30	25.33	27.35
Macedonia..FYR	32.22	32.22	32.22	28.60	26.70	23.50	27.20	29.60	31.43	35.61	32.96	28.64	32.18	31.42	32.45	32.38	30.72	30.25	35.00	35.70
Romania	21.77	21.77	21.77	22.90	23.32	26.13	23.90	28.89	30.26	30.50	31.27	32.61	32.95	35.44	37.05	35.56	35.50	36.30	37.80	38.30