2022

MULTIPLE TESTING CORRECTION IN TIME SERIES ROLLING WINDOW ANALYSIS WITH APPLICATION OF GWAS METHODS

Siyu Wang
Michigan Technological University, siywang@mtu.edu

MULTIPLE TESTING CORRECTION IN TIME SERIES ROLLING WINDOW ANALYSIS WITH APPLICATION OF GWAS METHODS

By

Siyu Wang

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Statistics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2022

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Statistics.

Department of Mathematical Sciences

Thesis Advisor:	*Dr. Yeonwoo Rho*
Committee Member:	*Dr. Qiuying Sha*
Committee Member:	*Dr. Kui Zhang*
Department Chair:	*Dr. Jiguang Sun*

Contents

List of Figures

Acknowledgments

First of all, I would like to express my deepest gratitude to Dr. Yeonwoo Rho, who provides me with invaluable guidance on my master's thesis and academic writing. I also appreciate her understanding of my personal situation and her encouragement for me to complete this thesis remotely. Without her dedicated support and help, this thesis cannot be completed.

I am also grateful for my committee members Dr. Qiuying Sha and Dr. Kui Zhang for their help and advice. I would like to thank faculty, staff and graduate students at the Department of Mathematics for their instructions and encouragements. I cannot name them all but must thank Dr. Melissa Keranen, Dr. Jiguang Sun and Dr. Jianping Dong.

Finally, I would like to thank my friends who offered me enormous help when I was at Houghton with my toddler alone last year. I also appreciate all the support from my family. Without their unconditional love and support, this work would not have been completed.

List of Abbreviations

AR	Autoregressive
BIC	Bayesian Information Criterion
BH	Benjamini-Hochberg procedure
FDR	False Discovery Rate
FWER	Familywise Error Rate
GATES	Gene-Based Association Test Using Extended Simes
GWAS	Genome-Wide Association Study
HMP	Harmonic Mean p-value
MinP	Minimum p-value
MinpBt	Bootstrap Minimum p-value
OB	O'Brien's method
SNPs	Single nucleotide polymorphisms
GATES	Gene-based Association Test that uses Extended Simes

Abstract

Rolling window is a popular tool in time series analysis. When conducting hypothesis testing on each window simultaneously, multiple testing problem occurs. In the literature in rolling window analysis, it appears that bootstrap is the most frequently used, if not only, method to address the multiple testing issue. This thesis aims to adapt multiple testing correction methods that are popular in genome-wide association study to the time series rolling window context. In particular, some of these methods require the knowledge of the correlation structure of test statistics. In genetics, this structure can be obtained from an external source, which may not exist in time series. To overcome this difficulty, we adopt the AR sieve idea, which enables the computation of correlation structure based on the estimated AR coefficients. We also present the finite sample simulation to illustrate the performance of these methods.

Chapter 1

Introduction

Rolling window has been widely used in time series analysis and econometrics. However, multiple testing issue occurs when hypothesis testing is performed on a number of windows at the same time. It seems that econometrics scholars usually employ bootstrap-based methods to handle the multiple testing problem in the rolling window setting, and this problem has not gotten enough attention yet [29].

Multiple testing is a challenging problem in Genome-wide association studies (GWAS). Researchers in GWAS are interested in determining the relationship between single nucleotide polymorphisms (SNPs) and phenotypes; in other words, how variations in gene sequences affect individuals' observable traits. For example, studies are conducted to determine if the differences in the lipid metabolic genes affect

the lipid metabolism pathways[47]. Since the tests, which detect the associations between SNPs and phenotypes, are performed on multiple phenotypes against SNPs, the problem of multiple comparisons happens, making it necessary to adjust p-values to reduce the chance of making type I errors. In addition, sharing individual level data has many restrictions with different institutions for the purpose of research. Therefore, statistical methods, to combine p-values using the summarized statistics, are also important when the whole data set is less accessible [37]. Researchers in GWAS have been aware of these problems for a long time, and a large number of methods have been proposed. These methods can be divided in two categories. One involves p-values without considering the dependence structure of test statistics, and the other specifies the dependence structure of test statistics.

First, we discuss the methods without modeling the dependency structure. The Bonferroni correction [6] is a classical method to control the familywise error rate (FWER). This method constructs the cut off value $\frac{\alpha}{n}$ at the significance level α divided by the number of tests n. However, the Bonferroni correction is known to be conservative typically when the number of tests is very large [32]. Simes [43] proposes an improvement of the Bonferroni correction. The Simes method rejects the null hypothesis if $p_{(i)} \leq \frac{i\alpha}{n}$ for at least one $i = 1, 2, \ldots, n$. The Simes method is more powerful than the Bonferroni correction. Harmonic mean p-value [19] controls the FWER. Wilson [51] mentions that by the assumptions of the generalized central limit theorem, an asymptotically exact p-value can be computed from the harmonic mean

p-value. Although Wilson [51] did not prove the generalized central limit theorem under dependent setting, according to [51]'s simulations, HMP has reasonable power, better than the Bonferroni correction, even when the tests are dependent. Besides the methods to control FWER, Benjamini and Hochberg [4] propose controlling the False Discovery Rate (FDR) when conducting multiple comparisons. Benjamini and Yekutieli [5] also prove the Benjamini-Hochberg procedure can control FDR when tests are positively dependent.

Now, we explore some methods enhancing the statistical power by modeling the dependency structure of tests. Gene-based association test using extended Simes procedure (GATES) [26] and trait-based association test that uses extended Simes procedure (TATES) [48] are inspired by the Simes method. Both GATES and TATES work well in combining univariate p-values to an overall p-value under the presence of possible dependency among p-values. The Minimum p-value (MinP) is another method requires the covariance matrix of test statistics. MinP is constructed to determine the most significant test statistics. The test statistics asymptotically follows a multivariate normal distribution under the null [37]. The key step for MinP is finding out the covariance matrix of test statistics. The O'Brien's method (OB) considers a linear combination of test statistics. However, OB may have a disadvantage when dealing with heterogeneous means. Yang et al. [52] propose a random splitting and cross-validation methods can overcome the heterogeneous mean limitation.

The above mentioned techniques have been successfully applied in GWAS. In this thesis, we take the initiative to explore how these methods perform in the time series rolling window setting to address the multiple testing issue. We select methods from the two categories of GWAS methods. OB, GATES and MinP are the three methods involving the correlation matrix or the covariance matrix of test statistics . Benjamini-Hochberg procedure (BH) and the harmonic mean p-value proposed by Wilson [51] (HMP) are the methods only need the original p-values. We also present residual bootstrap minimum p-value (MinpBt) [41], which is a commonly used method in time series to address the multiple testing problem. These methods will be elaborated in times series context in Chapter 3.

In GWAS, the dependency structure can be obtained from an external source, which does not exist in time series. To overcome this difficulty, we propose to approximate the unknown time series data structure to an autoregressive (AR) model by adopting the idea of AR sieve. We derive the theoretical covariance and correlation matrices of test statistics using the estimated autoregressive coefficients from the estimated model, which can be applied in methods that use the dependency structure of tests. We also use the AR sieve idea to compute the standard error if each test statistic, rather than using the conventional heteroskedastistic and autocorrelation consistant (HAC) esitmatiors [33]. The results from our simulations indicate that the methods we adopt can effectively work in the time series rolling window setting when the dependency is not too strong.

This thesis is constructed as follows. In Chapter 2, we introduce the three popular p-value combination methods in GWAS that utilizes covarinace structure among the tests: GATES, MinP and OB. In Chapter 3, we describe the major hurdles in implementing all the methods to the time series rolling window setting, and propose a remedy based on an AR sieve idea. The finite sample simulation results are presented in Chapter 4. In Chapter 5, we summarize the simulation results and discuss the future work. Additional simulation results are presented Appendices A – D.

Chapter 2

Methods in GWAS

In this chapter, we briefly review GATES, OB, and MinP in GWAS. Detecting relationships between SNPs and phenotypes is a prominent study in genetics. SNPs are the differences in gene sequences among individuals, and phenotypes are traits or observed gene expressions such as eye colors and blood types. For instance, [24] investigates if 2.6 million SNPs with a more than $100,000$ individuals of European ancestry, are associated with any of four lipid phenotypes: cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol and triglycerides. Since multiple phenotypes are probably related to SNPs, it is hard to detect the whole associations between SNPs and phenotypes if performing single test on one phenotype against each SNP. Therefore, conducting a test on multiple phenotypes against

each SNP can increase the statistical power. In this example, it is necessary to adjust p-values since millions of tests are performed. Numerous techniques have been developed to address the multiple comparison in GWAS. We choose three popular methods, GATES, MinP and OB in GWAS, following [24]'s setting.

Consider a linear regression model, where the number of individuals is m, SNPs are explanatory variables $\boldsymbol{x}$, and n phenotypes are response variables $\boldsymbol{y}$. In general, covariates, such as age and sex, appear in GWAS to adjust the strength of associations between SNPs and phenotypes [30]. For simplicity, we present a model, without covariates, testing the associations of the i-th phenotype with one SNP for $i = 1, \ldots, n$.

$$y_{it} = \beta_{0i} + \beta_{1i} x_t + \epsilon_{it} \quad \text{for} \quad t = 1, \ldots, m,$$

where x_t is an explanatory variable (SNP), β_{0i} and β_{1i} are the intercept and slope of the linear regression model. In GWAS, the observations are usually assumed independent across different individuals but not necessarily across different response variables (phenotypes). That is, for each $i = 1, \ldots, n$, ϵ_{it} are assumed independent and identically distributed (iid) with mean zero and variance $\sigma^2_{\epsilon i} > 0$.

To investigate the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$, existing studies focus on testing whether the regression coefficients are zero. For each i, one can test $H_{0i} : \ \beta_{1i} = 0$ versus $H_{ai} : \ \beta_{1i} \neq 0$ to see if the given explanatory variable has any significant

effect on the ith response variable. The test statistic to test H_{0i} can be a z-score $Z_i = \frac{\widehat{\beta}_{1i}}{se(\widehat{\beta}_{1i})}$, where $\widehat{\beta}_{1i}$ is the least square estimator of β_{1i} and $se(\widehat{\beta}_{1i})$ is its standard error under the iid error assumption. The global null and alternative hypotheses are H_0 : all H_{0i} are true and H_a : at least one H_{0i} is false. To test the global null hypothesis, understanding the behavior of the vector $\boldsymbol{Z} = (Z_1, \cdots, Z_n)'$ of test statistics for all n response variables is necessary. Note that $\boldsymbol{Z}$ approximately follows a multivariate normal distribution with mean zero and $n \times n$ covariance matrix Σ under the null hypothesis H_0 [24, 52]. The p-value p_i for each test can be calculated from this normal distribution. The correlation matrix among p-values is denoted as Φ_p. We refer the p-value for the global null H_0, combining p_i, as the overall p-value.

Since the observations can be dependent across $i = 1, \ldots, n$, Σ and Φ_p are not necessarily diagonal. Estimating Σ_z or Φ_p is one of the key steps in the GWAS methods we will review in this sections. In particular, $\widehat{\Phi}_p$, the estimated Φ_P, is one of the key elements in GATES, which will be introduced in Section 2.1. The estimated Σ, denote as $\widehat{\Sigma}$, for OB and MinP will be discussed in Section 2.2 and Section 3.3.

2.1 Gene-based Association Test that uses Extended Simes Procedure

We start by ranking p-values p_i in an ascending order $p_{(1)}, p_{(2)}, p_{(3)}, \ldots, p_{(n)}$. The Simes method [43] combines p-values by rejecting the null hypothesis when $p_{(i)} \leq i\alpha/n$ for at least one i among n tests, where α is the significance level. Even though the Simes method is more powerful than the Bonferroni method, it may still be too conservative when p-values are highly correlated [26]. The Simes method is a foundation for many methods that utilize the correlation structure of p-values. Li et al. [26] proposes an extended Simes method, named GATES, to evaluate SNP-based p-values and combines the overall p-value by selecting the minimum ordered i-th SNP-based p-value. GATES is different from phenotype based p-value combination method. GATES tests multiple SNPs versus one phenotype whereas phenotype based p-value method tests multiple phenotypes to one SNP. Van der Sluis et al. [48] borrows the GATES idea and develops the phenotype p-value combining technique known as TATES. Since this approach is mainly based on [26]'s idea, we refer this approach GATES, rather than TATES.

For GATES, one of the crucial processes is estimating the correlation matrix of SNP-based p-values. Li et al. [26] approximates the correlation matrix of p-values by

fitting a sixth order polynomial. To generate the response variables (SNPs), [26] proposes to use SNP's variant components alleles. In general, each SNP has two alleles. The correlation coefficient of alleles, denoted as r, can be obtained from an external resources in GWAS. We simplify the steps of achieving this sixth order polynomial are listed below.

1. Simulate two SNPs, based on r and allele frequencies under Hardy-Weinberg equilibrium in genetics, m times. Perform association tests on alleles for each SNP, resulting in two p-values.

2. Conduct the simulation $10,000$ times and calculate the correlation coefficients of the p-values.

3. Increase allele frequencies and r by 0.05 each time from minimum to maximum. Repeat Steps 1 and 2 for each allele frequencies and r and then a series of correlation coefficients of the p-values for each condition are generated.

4. Fit a sixth order polynomial, regressing the p-value correlation coefficients on r.

Using the external information r and this sixth order polynomial, we can approximate the $n \times n$ correlation matrix $\widehat{\Phi}_p$ among p-values. We calculate eigenvalues from the full correlation matrix $\widehat{\Phi}_p$ and from the top $i \times i$ submatrices of $\widehat{\Phi}_p$. After applying an indicator function $I(x)$, the effective number of all p-values m_e and the effective

number of top i independent p-values $m_{e,i}$ are obtained. GATES proposes the overall p-value as follows

$$P_{GATES} = \min_{i=1,\ldots,n} \frac{m_e p_{(i)}}{m_{e,i}}.$$

The effective number of top i independent p-values are expressed as

$$m_{e,i} = i - \sum_{b=1}^{i} [I(\lambda_b > 1)(\lambda_b - 1)] > 0 \quad \text{for} \quad i = 1, \ldots, n,$$

where λ_b is the b^{th} eigenvalue of the top $i \times i$ submatrix of the correlation matrix $\widehat{\Phi}_p$, and $I(x)$ is an indicator function. When i equals to n, m_e is computed using eigenvalues from the entire correlation matrix $\widehat{\Phi}_p$.

GATES is a useful tool for computation, statistical power enhancement, and type I error control when dependency exists. Furthermore, since GATES does not need the raw data for computation, it can be employed even when the data structure is complicated. However, in time series, we do not have the external information r. We adapt GATES in Section 3.3 to a time series rolling window setting, proposing an AR sieve approach to compensate the lack of external information.

2.2 Minimum p-value Method

MinP, or minimum p-value, is equivalent to the maximum absolute value of the n test statistics $\boldsymbol{Z}$:

$$T_{MinP} = \max_{1 \leq i \leq n} |Z_i|.$$

To obtain the overall p-value, one of the most popular methods is developed by Conneely and Boehnke [12] when test statistics are correlated. They propose a rapid numerical integration to calculate the overall p-value, and the p-value of MinP can be shown as

$$p_{MinP} = 1 - \int_{-T_{MinP}}^{T_{MinP}} \cdots \int_{-T_{MinP}}^{T_{MinP}} f(Z_1, \ldots, Z_n; 0, \widehat{\Sigma}) dZ_1 \ldots dZ_n,$$

where $f(Z_1, \ldots, Z_n; 0, \widehat{\Sigma})$ is the density function of a multivariate normal distribution with mean zero and covariance matrix $\widehat{\Sigma}$. Note that this method requires an estimation $\widehat{\Sigma}$ of the covariance matrix Σ among the test statistics. In GWAS, this information is often available from an external source. For the rolling window setting, we propose to use the AR sieve approach to estimate $\widehat{\Sigma}$. See Section 3.4 on how the MinP is implemented in a rolling window setting.

2.3 O'Brien's method

O'Brien's method is a rank-sum-type test to combine p-values, which is widely used in clinical research [23]. However, OB may have a disadvantage when dealing with heterogeneous means. [52]. OB suggests the linear combination of $Z_1, Z_2, \cdots, Z_n$, with $(\widehat{\Sigma})^{-1}$ and $e = (1, 1, \ldots, 1)'$ of length n. The global OB test statistic is

$$T_{OB} = e^{'}(\widehat{\Sigma})^{-1}\boldsymbol{Z}.$$

It is easy to see that the OB test statistics T_{OB} approximately follows a normal distribution with mean zero and variance $e'(\widehat{\Sigma})^{-1}e$ based on the central limit theorem or if $\boldsymbol{Z}$ is normal. To implement OB, covariance matrix $\widehat{\Sigma}$ of test statistics needs to be estimated. Again, we can use the AR sieve idea to estimate $\widehat{\Sigma}$. See 3.5 for more details on OB method in rolling window.

Chapter 3

Multiple Testing in Rolling Window

In this section, we adapt the methods described in chapter 2 in the context of time series rolling window and discuss BH, MinpBt and HMP. When applying GWAS-inspired methods in a rolling window setting, we face two major obstacles. One challenge is the difficulty of approximating the standard error. The conventional approach is the heteroskedasticity and autocorrelation consistent (HAC) estimation [53], but it is well-known that the HAC estimation suffers from severe size distortion when the dependence is strong [33]. The other obstacle is the lack of external information for the estimation of the covariance matrix of test statistics. In this chapter, we propose

the AR-sieve idea. This idea is originally from the AR sieve bootstrap [6] to approximate stationary time series data to an autoregressive (AR) model. The number of lag can be chosen by an information criterion. In this thesis, we use the Bayesian Information Criterion (BIC), but other information criteria can also be considered.

Based on the fitted AR model, we can compute the standard errors for each test statistic as well as the correlation covariance matrices of the test statistics across the windows. GATES, OB, and MinP can then be applied to the time series rolling window setting. In this thesis, we consider a simple mean test setting for brevity.

3.1 Zero Mean Test

Consider a set of time series data $X_t = \mu_t + u_t$ for $t = 1, \ldots, T$, where T is the total data length, μ_t is the mean and u_t is the error process. We assume that the error u_t is a stationary linear process. We consider overlapping rolling windows with a fixed window size m that moves forward in time. The number of windows is $n = T - m + 1$. We consider the following simple mean:

$$H_{0i} : \ \mu_i = 0 \text{ for all } i = 1, \ldots, n \quad \text{versus} \quad H_{ai} : \ \mu_i \neq 0 \ \text{ for some } i.$$

Recall that in the linear model for GWAS in Chapter 2, the number of individuals m in GWAS plays a similar role as the length of a fixed rolling window. The number of phenotypes n corresponds to the number of windows in time series.

We construct a zero mean test statistic of the i-th rolling window

$$Z_i = \overline{X}_i, \quad \text{for} \quad i = 1, \ldots, n,$$

where $\overline{X}_i = \frac{1}{m} \sum_{t=i}^{i+m-1} X_t$ is the sample mean of the i-th rolling window. It is worth mentioning that in genetics, $s.e.(\widehat{\beta}_i)$, in the z-score calculation, is obtained from the iid assumption, which is not appropriate for time series. In the next section, we introduce AR sieve approach to approximate $\{X_t\}_{t=1}^{T}$ to an AR($\widehat{p}$) model, where $\widehat{p}$ is the number of lags in the AR model chosen by the information criteria. Based on the fitted AR($\widehat{p}$) model, we can compute the standard error for Z_i, and therefore, p-value for each rolling window using the standard normal distribution function. The AR sieve modeling step also allows to calculate the correlation matrix of test statistics across window, which is necessary for GATES, OB and MinP.

3.2 AR Sieve Modeling

Bühlmann [8] proposes the AR sieve bootstrap, which approximates time series data

to an AR($\widehat{p}$) process. A large body of literature has shown that AR sieve produces a good approximation of the errors process [1, 2, 7, 8, 9, 21, 25]. We use this approach to fit the original time series data $\{X_t\}_{t=1}^T$ to calculate the standard error of test statistic as well as the variance of test statistics for rolling windows. The steps are outlined below.

1. Fit an AR(p) model to $\{X_t\}_{t=1}^T$ for each $p = 1, \ldots, p_{max}$:

$$X_t - \overline{X} = \sum_{i=1}^{p} \rho_i (X_{t-i} - \overline{X}) + \epsilon_t, \quad t = 1, \ldots, T,$$

   where $\overline{X} = T^{-1} \sum_{t=1}^{T} X_t$ is the sample mean of X_t and ϵ_t are iid random variables with mean zero and finite variance $\sigma_t^2 > 0$.

2. Determine the lag $\widehat{p}$ by choosing the minimum BIC such that $\widehat{p} = \underset{p \in \{1,2,\ldots,p_{max}\}}{\arg\min} \{BIC(AR(p))\}$.

3. Denote the autocovariance of X_t, $\widehat{\mathrm{Cov}}(X_i, X_{i+h})$, from the estimated AR($\widehat{p}$) as $\widehat{\gamma}(h)$. For example, if $\widehat{p} = 1$ then $\widehat{\gamma}(h) = \frac{\widehat{\rho}_1^h \widehat{\sigma}_t^2}{1-\widehat{\rho}_1^2}$, where $\widehat{\sigma}$ and $\widehat{\rho}_1$ are the maximum likelihood estimates of σ_t^2 and ρ_1, respectively. In general, $\widehat{\gamma}(\cdot)$, can be written as a function of estimated AR coefficients $\widehat{\rho}_1, \widehat{\rho}_2, \ldots, \widehat{\rho}_{\widehat{p}}$. The estimated variance of the test statistic for one rolling window is

$$\widehat{\mathrm{Var}}(Z_i) = \frac{1}{m^2} \sum_{k=0}^{m-1} \sum_{g=0}^{m-1} \widehat{\gamma}(|g-k|). \tag{3.2}$$

Since the test statistics Z_i approximately follows a normal distribution, the p-value for each rolling window is $2(1-\Phi(|Z_i|))$, where Φ is the standard normal cumulative distribution function.

An alternative approach to find the estimated $s.e.(\widehat{\beta}_i)$ is using a HAC estimation [53]. However, from the simulation results in Appendix B, the HAC estimation does not control the size well. This size distortion is well-known in literature [33]. When smaller width m for rolling windows are chosen, the size distortion would become even worse [33, 50]. We propose to use the AR sieve approximation to alleviate the size distortion issue.

As discussed in Chapter 2, MinP, GATES and OB needs to estimate the correlation and covariance matrices of test statistics. The steps we propose to calculate them based on the AR sieve model are listed:

1. Pick any two windows with distance h: $W_i = (X_i, \ldots, X_{i+m-1})$ and $W_{i+h} = (X_{i+h}, \ldots, X_{i+m-1+h})$, where h ranges from 0 to $T-m$ and $i = 1, \ldots, n$. The corresponding test statistics for these two windows are Z_i and Z_{i+h}.

2. Calculate the theoretical covariance of test statistics for W_i and W_{i+h} as

$$\mathrm{Cov}(Z_i, Z_{i+h}) = \frac{1}{m^2} \sum_{k=0}^{m-1} \sum_{g=0}^{m-1} \mathrm{Cov}(X_{i+k}, X_{i+h+g}).$$

Recall that $\widehat{\gamma}$ is the sample autocovariances of $\{X_t\}_{t=1}^{T}$ based on the estimated AR coefficients. The estimated autocovariance among the tests, $\mathrm{Cov}(Z_i, Z_{i+h})$, can be rewritten as

$$\widehat{\mathrm{Cov}}(Z_i, Z_{i+h}) = \frac{1}{m^2} \sum_{k=0}^{m-1} \sum_{g=0}^{m-1} \widehat{\gamma}(|h+g-k|). \tag{3.3}$$

The (i,j)-th element of the estimated $n \times n$ covariance matrix $\widehat{\Sigma}_z$ of test statistics using AR sieve approximation is $\widehat{\mathrm{Cov}}(Z_i, Z_j)$ obtained from (3.3).

3. The correlation of test statistics using the estimated AR($\widehat{p}$) coefficients between any two windows can be expressed as

$$\widehat{\mathrm{Cor}}(Z_i, Z_{i+h}) = \frac{\widehat{\mathrm{Cov}}(Z_i, Z_{i+h})}{\widehat{\mathrm{Var}}(Z_i)} = \frac{\sum_{k=0}^{m-1} \sum_{g=0}^{m-1} \widehat{\gamma}(|h+g-k|)}{\sum_{k=0}^{m-1} \sum_{g=0}^{m-1} \widehat{\gamma}(|g-k|)}. \tag{3.4}$$

The (i,j)-th element of $\widehat{\Omega}_z$ is $\widehat{\mathrm{Cor}}(Z_i, Z_j)$ obtained from (3.4). Note that the diagonal elements of $\widehat{\Omega}_z$ are exactly 1 indicating the distance between two windows is 0. The off-diagonal elements of $\widehat{\Omega}_z$ are ordered by the distance between two windows from 1 to $T - m - 1$.

3.3 Adjusted Gene-based Association Test that uses Extended Simes Procedure

In this section, we adapt GATES to the rolling window setting. Let $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$ be the ascending ordered p-values for n windows. The overall p-value is computed by combining the individual p-values into one smallest weighted p-value P_{GATES} as follows:

$$P_{GATES} = \min_{1 \leq i \leq n} \left(\frac{m_e p_{(i)}}{m_{e,i}} \right), \tag{3.5}$$

where m_e stands for the effective number of independent p-values of all n windows and $m_{e,i}$ represents the effective number of independent p-values of the top i p-values for $i = 1, \ldots, n$. Following the original GATES idea, the calculations of m_e and $m_{e,i}$ include eigenvalues from the correlation of p-values, which has been calculated in Section 3.2. We modify the sixth order polynomial in GATES for our setting. To obtain $\widehat{\Omega}_p$, we first calculate the estimated correlation matrix of test statistics $\widehat{\Omega}_z$, and then fit $\widehat{\Omega}_z$ with the sixth order polynomial to estimate $\widehat{\Omega}_p$. In this process, we consider an AR(1) model with data length $T = 100$, AR(1) coefficients $\delta_1 = 0, 0.1, \ldots, 0.9$, and window sizes $m = 10, 11, \ldots, 50$ are considered. The length of the p-value vector is $n = T - m + 1$ and the distance between two windows $h = 0, 1, \ldots, T - m$.

The steps to approximate $\widehat{\Omega}_p$ are listed below.

1. Fix δ_1 and m. Randomly generate $10,000$ AR(1) data with AR coefficient δ_1 and data length $T = 100$. Compute the test statistics and their p-values on rolling windows.

2. Calculate the n by n sample correlation matrix Λ_p of p-values, based on the $10,000$ replications.

3. Using (3.4), compute the theoretical correlation of test statistics Λ_z using the true AR(1) structure and the true AR(1) coefficient.

4. For each h, compute the average of the correlations of p-values from Λ_p and the average of correlation of test statistics from Λ_z.

5. Repeat Steps 1-4 for all pairs of δ_1 and m. We have two sets of data for all the AR(1) coefficient δ_1 and window size m combinations: one is the averages of the sample correlations of p-values with the same h for all the pairs of δ_1 and m, referred as Avg_p. The other is the averages of the theoretical correlations of test statistics with the same h for all the pairs of δ_1 and m, denoted as Avg_z.

6. Regress Avg_p on Avg_z, setting Avg_p as the response variable. The sixth order polynomial fitted in our simulation is $\widehat{Avg}_p = -0.000076 - 0.068106Avg_z + 1.572703Avg_z^2 - 5.706136Avg_z^3 + 14.000826Avg_z^4 - 15.196462Avg_z^5 + 6.379301Avg_z^6$ with the coefficient of determination $R^2 = 0.9901$. See Appendix A for the fitted curve plot.

7. Swap the row and columns of $\widehat{\Omega}_z$, which is obtained from (3.4), according to

the rank order of p-values. Denote the ordered matrix as $\widehat{\Omega}_{zo}$, and its (i,j)-th element as $\widehat{\Omega}_{zo,i,j}$. Define $\widehat{\Omega}_{po}$ for $\widehat{\Omega}_{p}$, similarly to $\widehat{\Omega}_{zo}$. The (i,j)-th element of $\widehat{\Omega}_{po}$ can be obtained as $-0.000076 - 0.068106\widehat{\Omega}_{zo,i,j} + 1.572703\widehat{\Omega}^2_{zo,i,j} - 5.706136\widehat{\Omega}^3_{zo,i,j} + 14.000826\widehat{\Omega}^4_{zo,i,j} - 15.196462\widehat{\Omega}^5_{zo,i,j} + 6.379301\widehat{\Omega}^6_{zo,i,j}$.

We also consider the sixth order polynomial using the AR(1) coefficient from -0.9 to 0.9 by 0.1. While the two sets of sixth order polynomials based on nonnegative AR coefficients $\delta_1 = 0, \ldots, 0.9$ and the new set of coefficients including the negative correlation result in very similar sizes and statistical powers in the same setting in our unreported simulation. For the rest of this thesis, we only report the results with the sixth order polynomials from the non-negative AR(1) coefficients.

Thereafter, recalling the calculations of m_e and $m_{e,i}$, we compute the eigenvalues λ_b from the top $i \times i$ submatrix of $\widehat{\Omega}_{po}$ and calculate $m_{e,i}$ using $m_{e,i} = i - \sum_{b=1}^{i}[I(\lambda_b > 1)(\lambda_b - 1)] > 0$ for $i = 1, \ldots, n$. When $i = n$, λ_b is determined from the full matrix $\widehat{\Omega}_{po}$ and $m_e = m_{e,n}$. Now, we can use (3.5) to get the overall P_{GATES}. Comparing with the significance level α, if P_{GATES} is larger than α, we cannot reject the null, suggesting that there is no mean change in the original time series data.

3.4 Adjusted Minimum p-value Method

The theoretical covariance matrix of the test statistics is taken into consideration by OB and MinP. Let $\boldsymbol{Z} = (Z_1, Z_2, \cdots, Z_n)'$ be the vector of zero mean test statistics. Since Z_i are sample means, $\boldsymbol{Z}$ approximately follows a multivariate normal distribution with mean zero and a covariance matrix $\widehat{\Sigma}_z$. Each element of $\widehat{\Sigma}_z$ can be estimated by equation (3.3). We write the estimated

MinP obtains the maximum absolute value of test statics among all windows, representing as

$$T_{minP} = \max_{1 \le i \le n} |Z_i|,$$

The combined test statistic T_{minP} follows multivariate normal distribution with mean zero and covariance $\widehat{\Sigma}_z$ [12]. The MinP p-value can be found from this multivariate normal distribution. If the calculated overall p-value is greater than the significance level, we cannot reject the null, indicating there is no mean change in the original data. MinP is a very popular approach since it has a simple calculation and it is expected more powerful when the rolling windows are correlated [54].

3.5 Adjusted O'Brien's method

OB can be adjusted to a linear combination of the individual rolling window test statistics with equal weights.

$$T_{OB} = \boldsymbol{e}'\widehat{\Sigma}_z^{-1}\boldsymbol{Z},$$

where $\boldsymbol{e} = (1, 1, \cdots, 1)'$ is the equal weight vector and the overall test statistics T_{OB} follows a normal distribution with mean zero and the variance $\boldsymbol{e}'\widehat{\Sigma}_z^{-1}\boldsymbol{e}$ under the null. The OB p-value can be found using its distribution. If the overall p-value is less than α, then there is enough evidence that the mean changes not to zero in the original time series data. OB is easy to be implemented and handles the independent and dependent test statistics, which is suitable for our rolling window structure in time series. However, it is less powerful when the mean of data is heterogeneous [52].

The methods involved with the theoretical correlation or covariance matrix of test statistics have been discussed. The remaining portion of this chapter will provide introductions to the approaches that only require original test statistics or p-values

3.6 Benjamini-Hochberg procedure

The Benjamini-Hochberg procedure is widely used in multiple testing and it controls FDR at the significance level. FDR is the the number of false positive discoveries divided by the number of total discoveries. Benjamini and Hochberg [4] proposed this idea to estimate the proportion of incorrect rejections which is less conservative. Let's assume the null hypothesis are $H_1, H_2, \ldots, H_n$, and the corresponding p-values in an ascending order $p_{(1)}, \ldots, p_{(n)}$ for n rolling windows. BH finds the largest i at the given significance level α and works as follows, $i_0 = \max\left\{i : p_{(i)} \leq \alpha\frac{i}{n}\right\}$. Then if i_0 exist, we reject $H_{(1)}, \ldots, H_{(i_0)}$, otherwise, we cannot reject the null hypotheses.

3.7 Harmonic Mean p-value

The harmonic mean p-value method is combining p-values when the dependency exists. Harmonic mean p-value controls FWER. FWER is the probability of having at least one false positives. Controlling FWER are more stringent than controlling FDR. Harmonic mean p-value can be expressed as $\mathring{p} = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i p_i^{-1}}$, where w_i is the weight of the ith hypothesis. The weights can be the equal weights $w_i = \frac{1}{n}$ in our setting. With the assumption of generalized central limit theorem, Wilson [51] proves

that an asymptotically exact p-value $p_{\mathring{p}}$ can be calculated as follows:

$$p_{\mathring{p}} = \int_{\frac{1}{\mathring{p}}}^{\infty} f_{Landau}(x|\log n + 0.874, \frac{\pi}{2})dx$$

where the Landau distribution probability density function is $f_{Landau}(x|\mu,\sigma) = \frac{1}{\pi\sigma}\int_0^\infty e^{-s\frac{(x-\mu)}{\sigma}-\frac{2}{\pi}s\log(s)}\sin(2s)ds$. HMP proposed by Wilson [51] can be used to control the size when the p-values are dependent with each other. We adopt this method to manage the dependency among the rolling windows. In addition, we compare the harmonic mean p-value and the harmonic mean p-value proposed by Wilson. From the unreported simulation results, the size is better controlled by Wilson's harmonic mean p-value at the expense of power. The simulation results in our study show that when the dependency between rolling windows is very strong, we have high inflated type I error and low power for many methods. In order to better control the size, we continue the study using Wilson's harmonic mean p-value approach.

3.8 Residual Bootstrap Method

MinpBt is a method bootstrapping the residuals (3.1) from the fitted AR model. We follow the procedures from [41] and break down the procedure in six steps to implement the method.

1. We apply rolling window on the original time series data X_t and store the maximum absolute mean Z_o among all n windows.

2. We fit the data to an AR($\widehat{p}$) as mentioned earlier this chapter, and then save the residuals $\widehat{\epsilon}_t$ and coefficients $\widehat{\rho}_i$ where i is from 1 to $\widehat{p}$ from the AR sieve approximation model. The fitted residuals under the null can be expressed as

$$\widehat{e_t} = (X_t - \overline{X}) - \sum_{i=1}^{\widehat{p}} \widehat{\rho}_i (X_{t-i} - \overline{X}) \quad \text{for} \quad t = \widehat{p}+1, \ldots, T. \tag{3.1}$$

3. Let T be the sample size of the bootstrapped time series data, then we generate the bootstrap data as

$$X_t^* = \widehat{\rho}_1 X_{t-1}^* + \widehat{\rho}_2 X_{t-2}^* + \cdots + \widehat{\rho}_{\widehat{p}} X_{t-\widehat{p}}^* + \epsilon_t^*, \quad \text{for} \quad t = \widehat{p}+1, \ldots, n,$$

where $\epsilon*_t$ is randomly drawn from the saved residuals $\widehat{\epsilon_t}$ with replacement.

4. After having the bootstrap time series X_t^*, we again apply rolling window and store the maximum absolute mean Z_1^* among all n windows.

5. Repeat steps 3 to 4 B times. We can obtain a vector of bootstrapped test statistics $\boldsymbol{Z^*} = \{Z_1^*, Z_2^*, \ldots, Z_B^*\}$.

6. The critical value is determined as $95th$ percentiles of $\boldsymbol{Z^*}$ and then compare the critical value with Z_o.

This bootstrap method does not require the covariance matrix of test statistics among rolling windows. MinpBt can be effective in enhancing power and control the size close to the nominal level. Even though it can be simply implemented, the computation time is long.

Chapter 4

Simulation

In this section, we explore the size, the power and the size adjusted power for all methods discussed in Chapter 3. We begin by introducing the settings throughout the simulation. After that, we look at how sensitive the AR sieve technique is to the actual data generation process. We then compare the size, the power, and the size adjusted power for all methods to determine the optimal window size ranges. After determining the optimal window size, we compare how the strength of dependence affect the size, the power and the size adjusted power. Lastly, we compare the performance of rejection rates for all the p-value combination methods.

We consider $1,000$ Monte Carlo replications, the length T of each time series is 100 and the number of bootstrap replications B is 500 for the MinpBt. The model in the

simulation is

$$X_t = \mu_t + u_t \quad \text{for} \quad t = 1, \ldots, T \tag{4.1}$$

where μ_t is the mean and u_t is the error process. We consider the error process in two cases: AR(1) and ARMA$(1, 1)$.

1. AR(1) model is $u_t = \rho u_{t-1} + \varepsilon_t$, where the AR(1) coefficient ρ ranges from -0.8 to 0.8 by 0.2 and ε_t is i.i.d from the standard normal distribution.

2. ARMA$(1, 1)$ model is $u_t - \rho u_{t-1} = \varepsilon_t + \theta \varepsilon_{t-1}$. We use two sets of ρ and θ: (0.2, 0.1) and (0.7, -0.3). ε_t is i.i.d from the standard normal distribution.

The variance of ε_t could be other options and here we choose 1 in our simulation. Under the null, the mean μ_t is set to 0 for all $t = 1, \ldots, T = 100$. Under the alternative, we consider two cases:

$$\text{Case } 1 : \mu_t = 0.5 \ \{\tfrac{t}{T} \leq \tfrac{1}{3}\} \ \text{ and } \ \mu_t = 0 \ \{\tfrac{t}{T} > \tfrac{1}{3}\};$$
$$\text{Case } 2 : \mu_t = 1 \ \{\tfrac{t}{T} \leq \tfrac{1}{3}\} \ \text{ and } \ \mu_t = 0 \ \{\tfrac{t}{T} > \tfrac{1}{3}\}.$$

The window size m is from 10 to 60 by 5, and the significance level is 5%. We use two `R` packages in our simulation. To fit an AR$(\widehat{p})$ model and obtain the estimated AR coefficients, we use the `Arima` function in the `forecast` package in `R`, setting the method to `ML` in the function. Another package we use is `harmonicmeanp` in `R`. We

use the function `p.hmp` to calculate the overall HMP p-value.

4.1 The Effect of AR Sieve Approach for Individual p-values

As mentioned in Chapter 3, HAC estimator could be another option for us to find the p-values for each rolling window. We evaluate the distributions of p-values to see if the HAC estimation in our rolling window setting works. The findings are detailed in Appendix B.5. Given the distributions of p-values, it appears that the HAC estimation does not yield good p-values. In order to acquire good p-values, we employ the AR sieve technique. We also examine whether fitting the error process using AR sieve approach can produce good p-values.

we generate data on (4.1) where AR(1) and ARMA$(1,1)$ are on the error process discussed above. We demonstrate the original p-values distribution with a fixed window size 25 when the error process is AR(1). We also provide p-values distributions for window sizes 10, 45 and 60 in Appendix B.1 – B.3 and p-value distribution for ARMA$(1,1)$ in Appendix B.4.

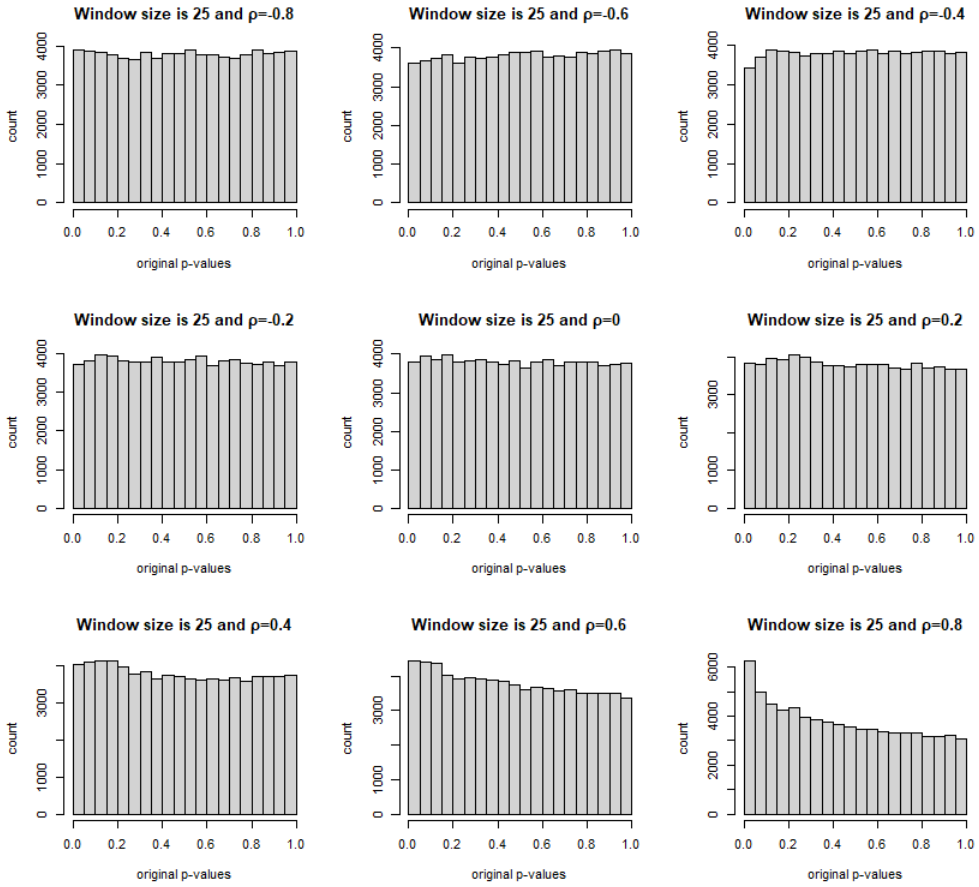

Figure 4.1: This figure shows when window size is 25, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

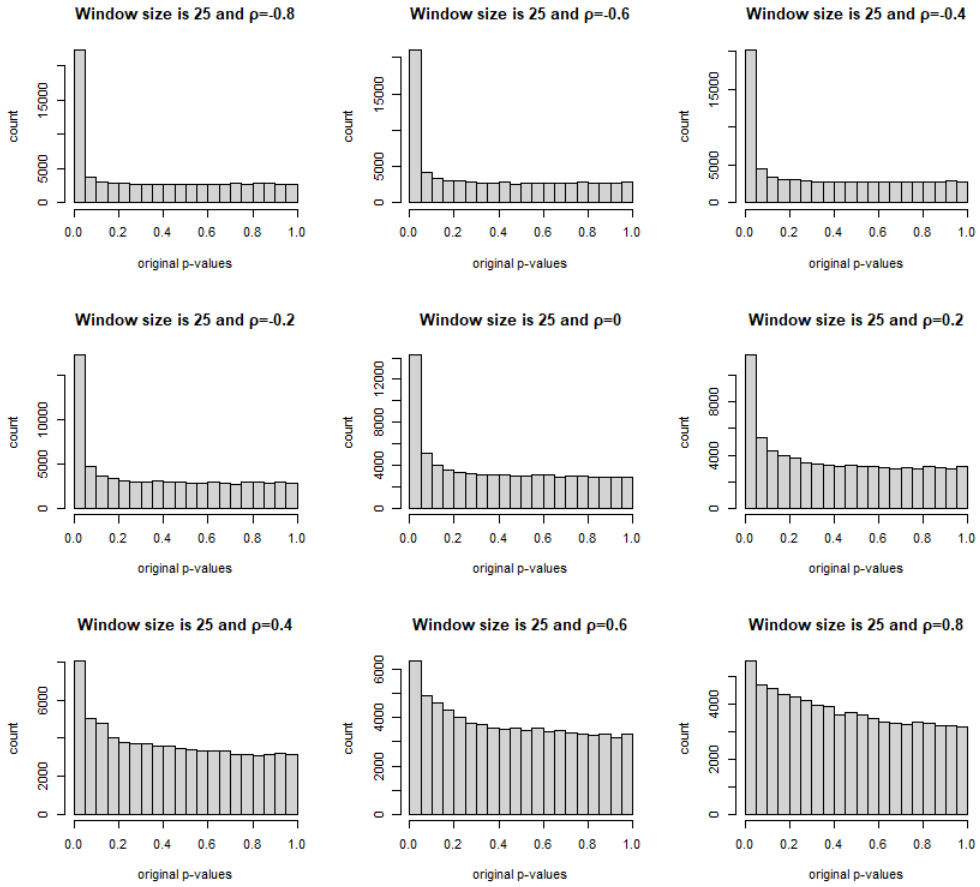

Figure 4.2: This figure shows when window size is 25, the p-values behaviors, under the alternative $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

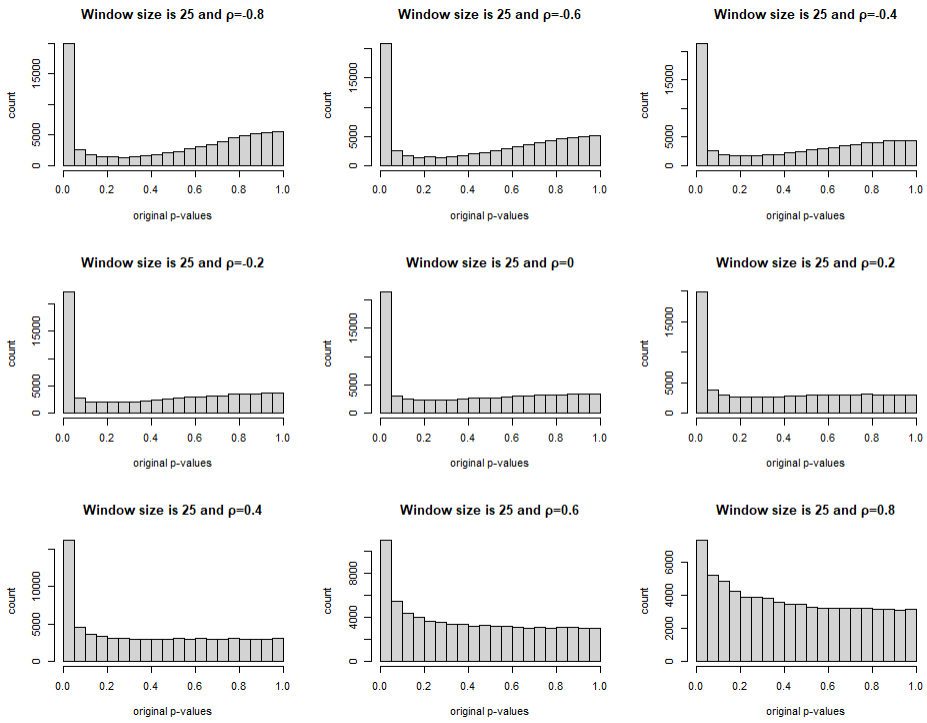

Figure 4.3: This figure shows when window size is 25, the p-values behaviors, under the alternative $\mu_t = 1\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

Figure 4.1 illustrates that, when the strength of dependency in the data generating process is positively strong, adopting AR sieve approach cannot produce uniformly distribution p-values under the null. Figure 4.2 is the p-value distribution under the alternative Case 1. It is seems that more p-values are under the nominal rejection rate 5% and ,as the dependency strength positively increases, it is less powerful. Figure 4.3 is the p-value distribution under the alternative Case 2. The behavior of p-values are slightly different than the behavior in Case 1. Figure 4.3 displays that some p-values are near 1 and most of the p-values are near 0, which could result in less

power under the alternative when performing p-value combination methods. Based on the distributions of p-values with different fixed window sizes, we conclude that, regardless of the window sizes, estimating the unknown data structure using the AR sieve approach has its limitation when the unknown data are highly correlated. but better than hac at least.

4.2 The Effect of Window Size in Rolling Window Analysis

In Section 4.1, we discuss that the distribution of p-values from the estimated AR($\widehat{p}$) process, and in this section, we explore how the window size selection affects the p-value combination methods. When applying p-value combination methods on rolling windows, the window size is important because it affects the dependency structure of test statistics. The calculation of the correlation and covariance matrix of tests statistics on rolling windows can be found in equations (3.3) and (3.4). We conduct simulation and generate data from (4.1) where the error processes are AR(1) and ARMA(1,1) discussed above. We present selected results in this section. The full detailed results can be found in Appendix C.

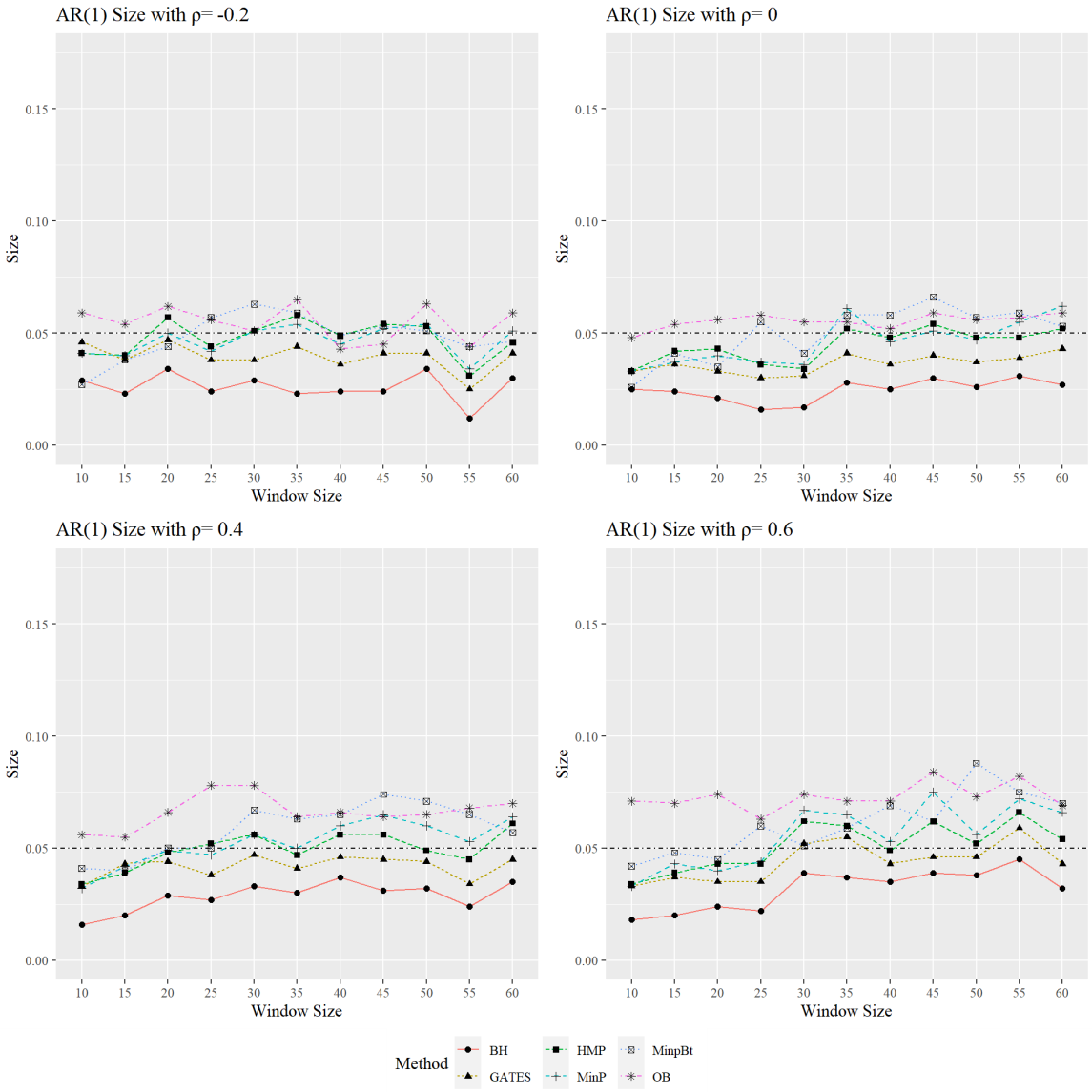

Figure 4.4: This figure shows the size comparison for the six methods when AR(1) coefficient ρ is -0.2, 0, 0.4 and 0.6.

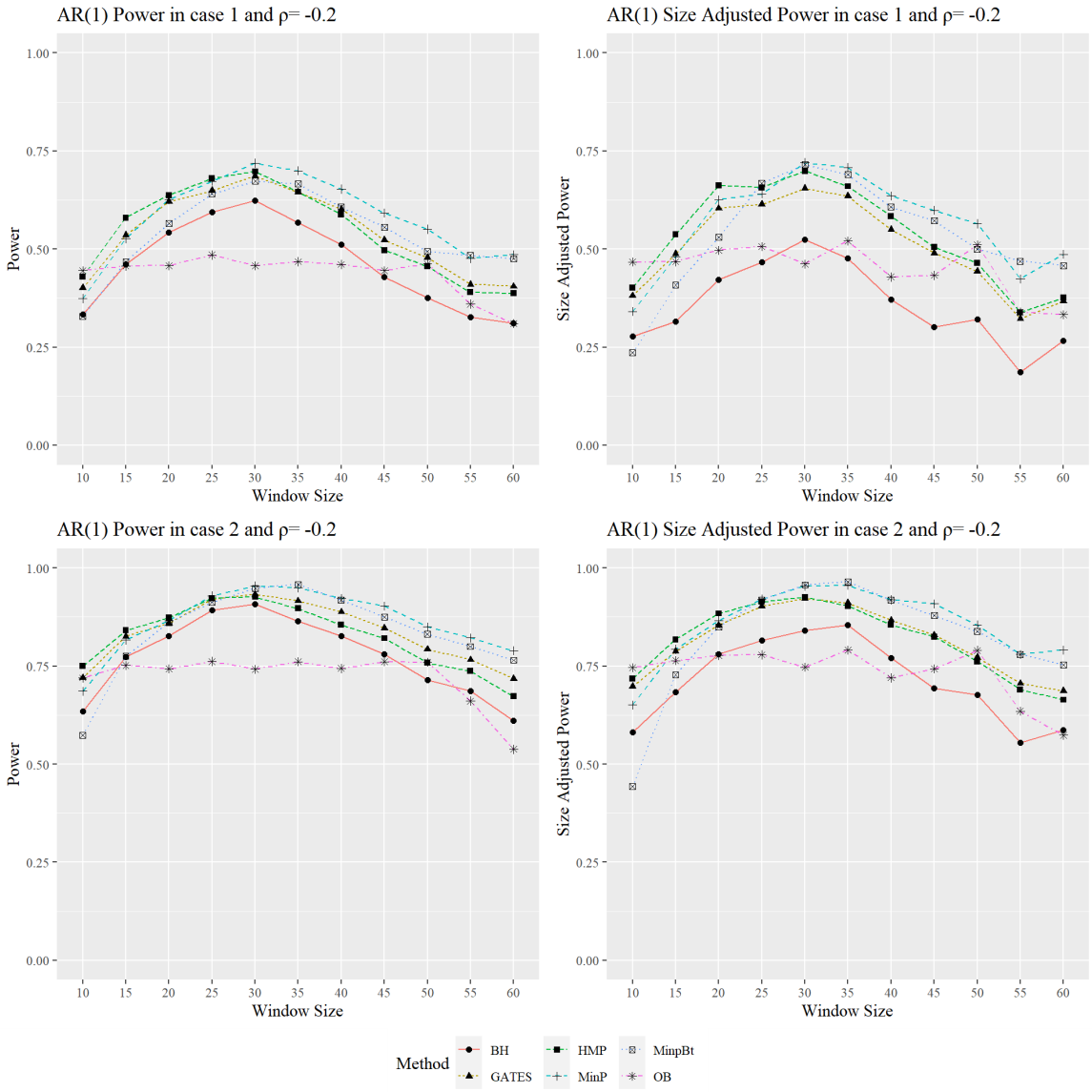

Figure 4.5: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is -0.2. Referring the simulation setting, Case 1 is $\mu_t = 0.5\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$; Case 2 is $\mu_t = 1\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$

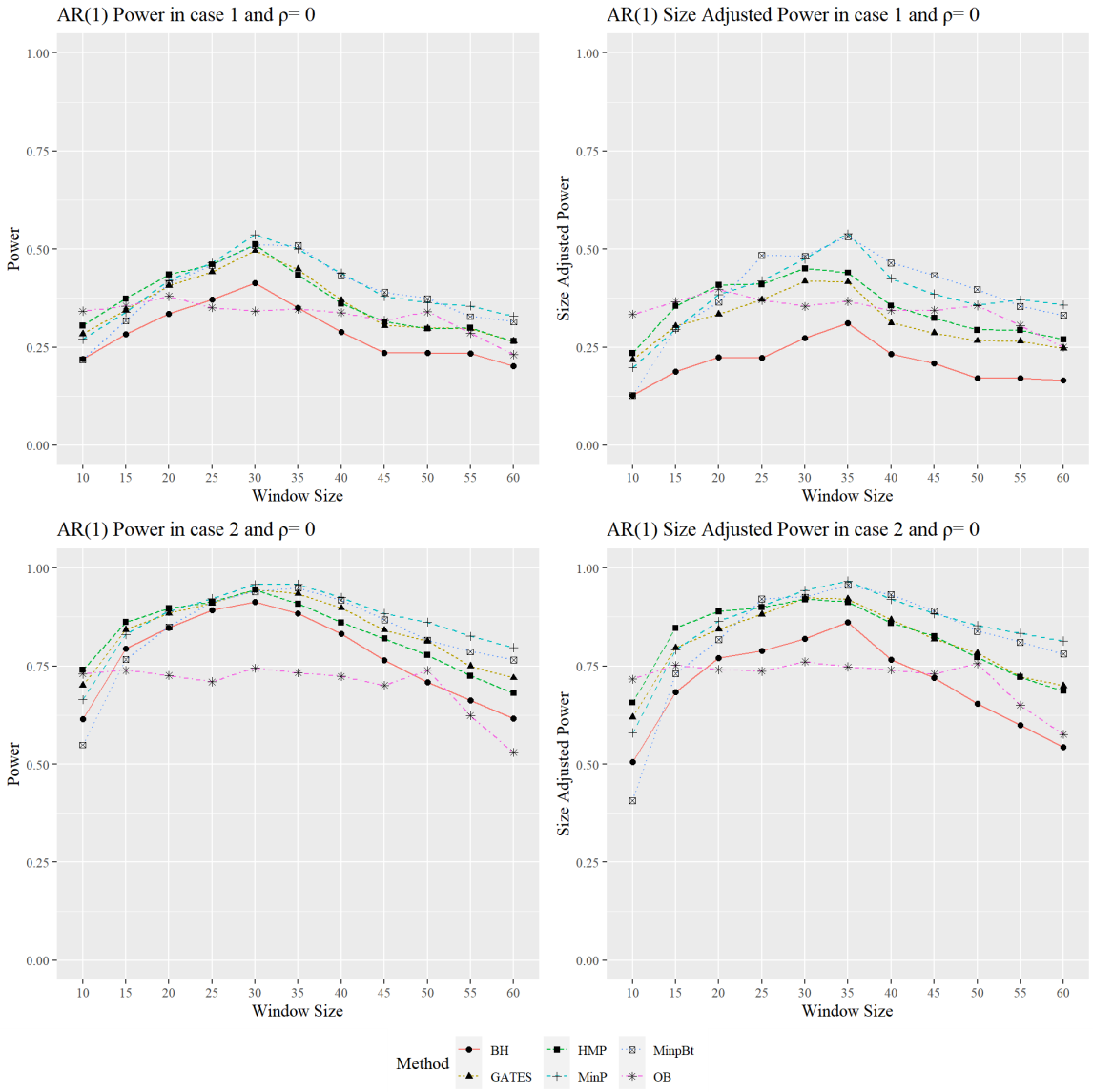

Figure 4.6: This figure shows the size, power and size adjusted power comparisons for the six methods when AR(1) coefficient ρ is 0. Referring the simulation setting, Case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; Case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

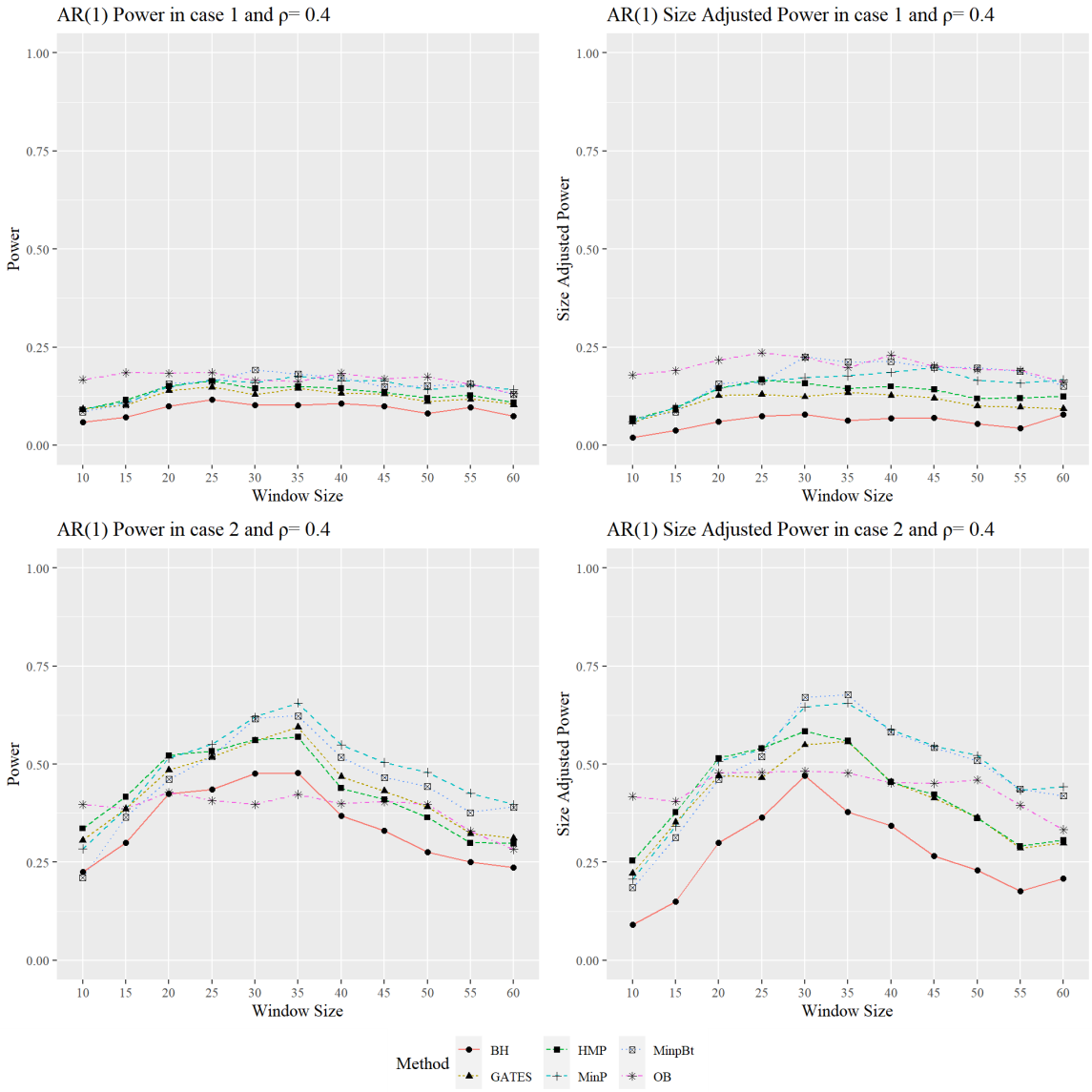

Figure 4.7: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is 0.4. Referring the simulation setting, Case 1 is $\mu_t = 0.5\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$; Case 2 is $\mu_t = 1\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$

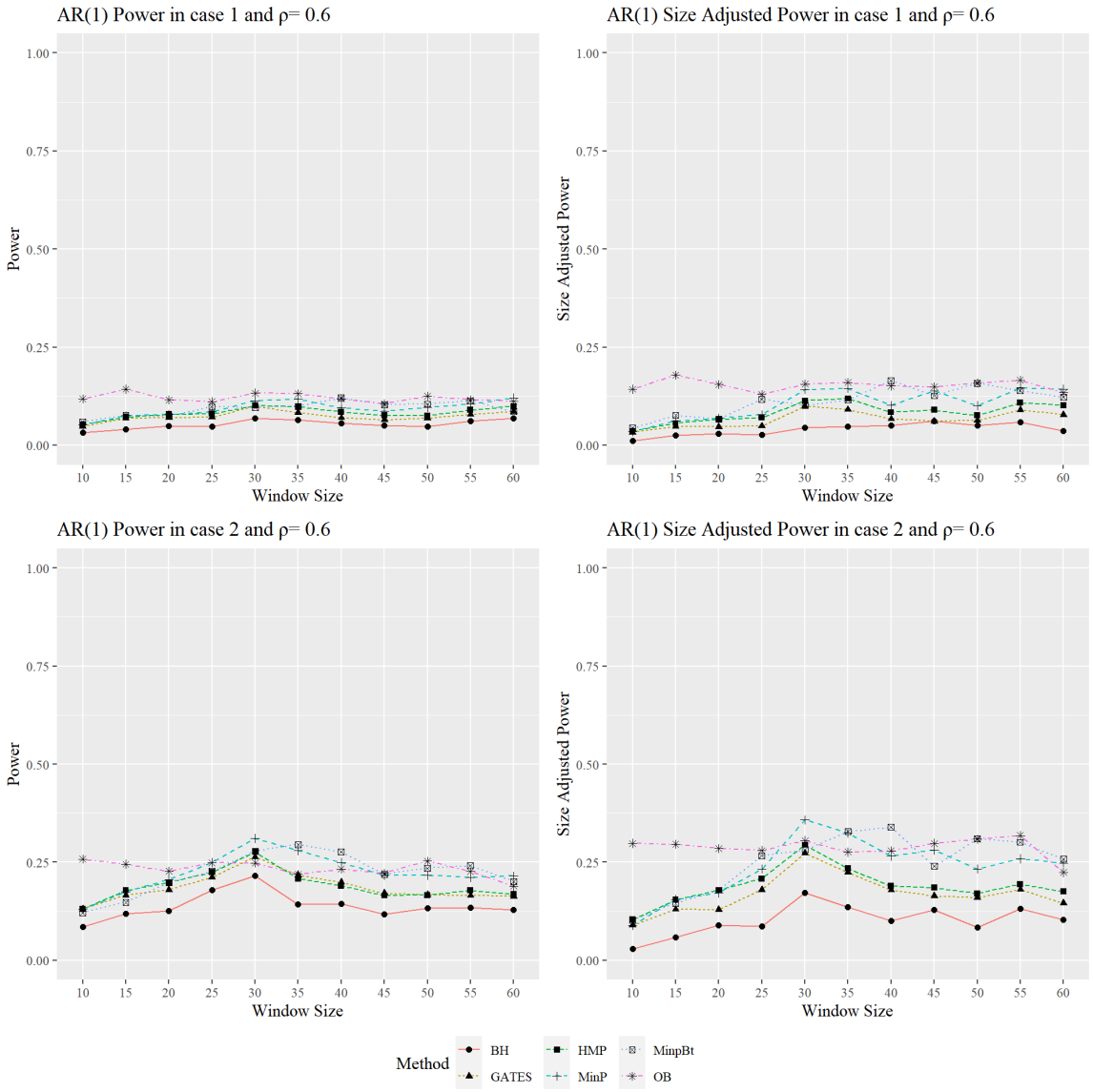

Figure 4.8: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is 0.6. Referring the simulation setting, Case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; Case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

Figure 4.4 is the size comparison for the six methods when AR(1) coefficient ρ is -0.2, 0, 0.4 and 0.6 respectively. Figure 4.5 – 4.8 are the power and the size adjust power

comparisons for the six methods under the alternative for Case 1 and Case 2. We notice that no matter the strength of dependency in the data generating process, when window size around 25% to 35% of the total data length, the statistical power remains at a higher level. The results in Appendix C.2 for ARMA(1,1) also demonstrate that the optimal window size is around 25% to 35% of the total data length. However, when the ARMA(1,1) coefficients are $\rho = 0.7$ and $\theta = -0.3$, all of the methods are not effective. This is because, under strong dependency, the error process approximation using AR sieve approach does not have a concise AR model to be presented. While, another case, when $\rho = 0.2$ and $\theta = 0.1$, it has a concise AR model approximation so that all the p-value combination methods are effective.

Our proposed window size choice is consistent with Shi et al. [41]'s minimum fixed window size suggestion, 24% of the total data length. In the next section, we set the window size to be 25 and 30 and explore how all the methods are affected by the strength and direction of the dependencies on the rolling window.

4.3 The Effect of Strength and Direction of Dependencies on the Rolling Window Analysis

As mentioned in Chapter 3, the estimated correlation and the theoretical covariance matrices of tests are calculated using the estimated AR coefficients from the estimated

AR($\widehat{p}$). Therefore, the strength of theoretical correlation of tests can be influenced by the data generating procedure. We investigate this relationship with errors u_t generated from our AR (1) models.

Figures 4.9 and 4.10 exhibit the size, the power and the size adjusted power comparisons with window size 25 and 30 respectively. The X-axis for each subfigure is the AR(1) coefficients ρ of the error process, from -0.8 to 0.8 by 0.2. The Y-axis represents the power for the left top two subfigures, the size adjusted power for the right top two subfigures and the size for the bottom subfigure.

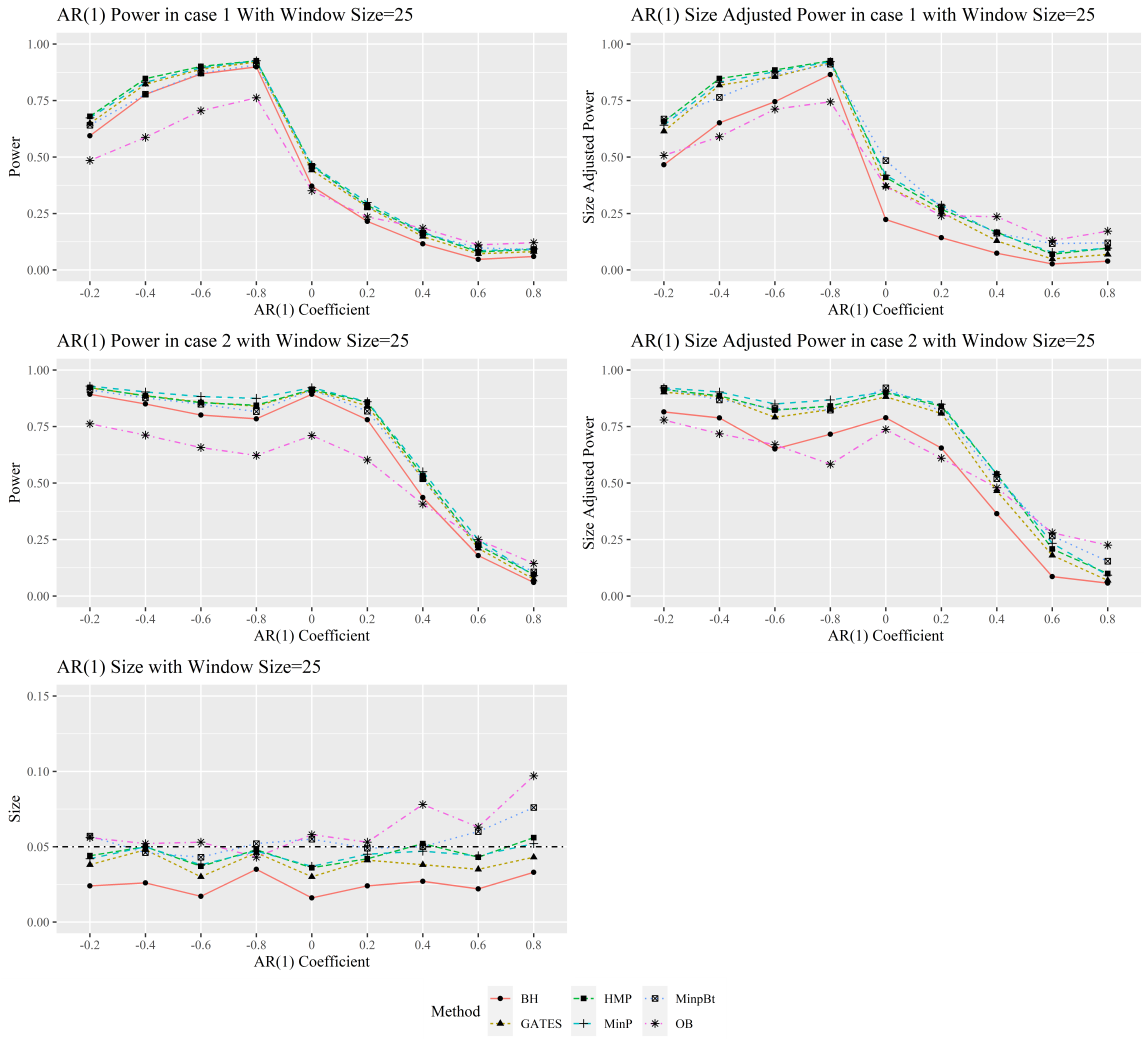

Figure 4.9: This figure shows the size, power and size adjusted power comparisons for the six methods with fixed window size 25. Referring the simulation setting, Case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; Case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

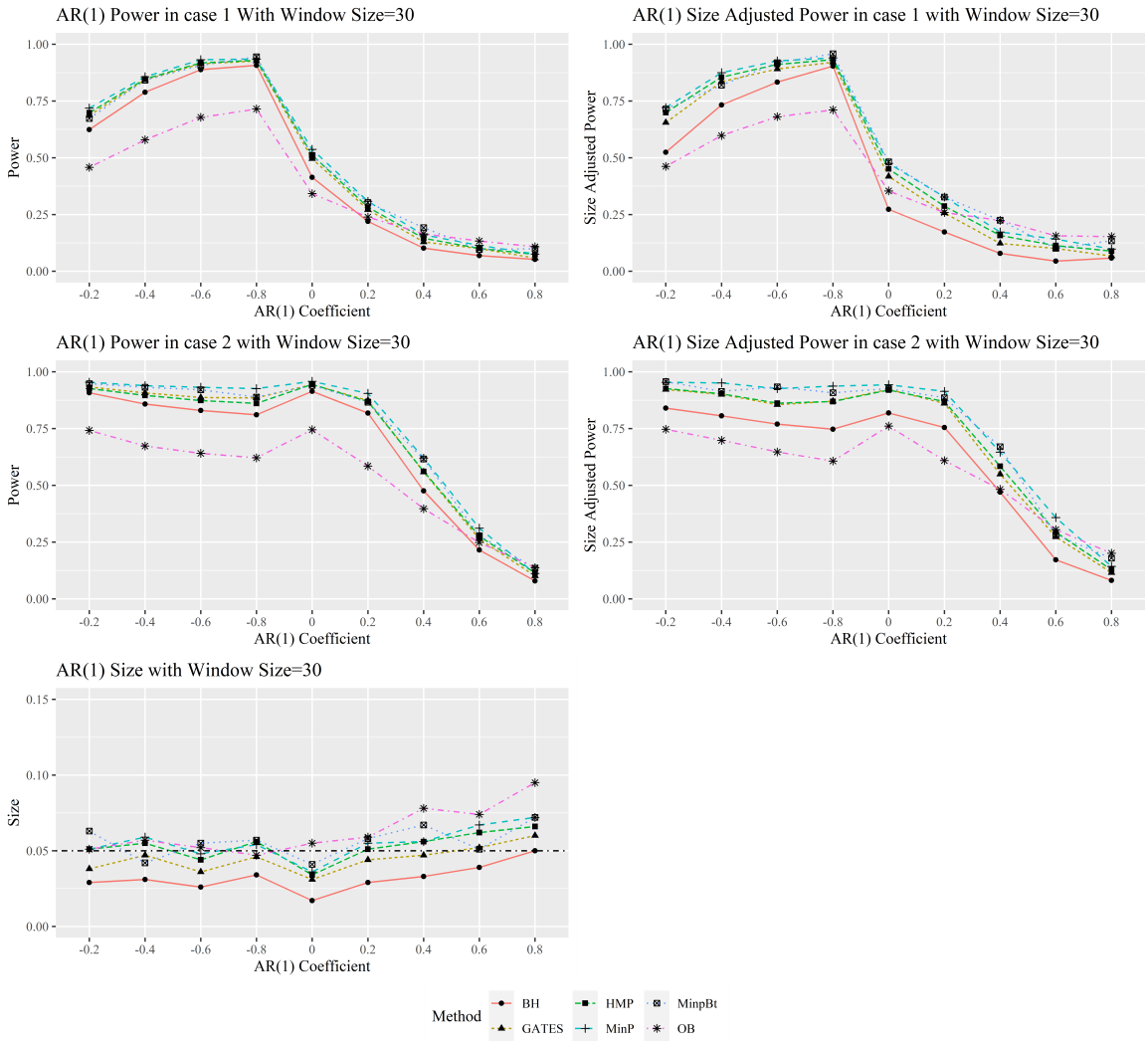

Figure 4.10: This figure shows the size, power and size adjusted power comparisons for the six methods with fixed window size 30. Referring the simulation setting, Case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; Case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

From Figure 4.9 and 4.10, it seems that the size and the power are significantly influenced by the direction and strength of dependency represented in ρ. When ρ is

positive and large, we lose the control of the size, the power and the size adjusted power. BH, HMP and MinpBt have inflated sizes and low power, resulting from the inaccurate original p-values from the AR sieve approximation. The distribution of original p-values has already been discussed in Section 4.1. GATES, MinP and OB are affected by the estimated correlation and covariance matrices of tests based on the estimated AR coefficients from the AR($\widehat{p}$) model.

However, when ρ is negative, the size can be controlled and the power and the size adjusted power are robust. In general, negative correlation generally leads to estimators with smaller variance [13]. This is because when negatively correlated, the time series data tend to oscillate back and forth across the mean, the estimated mean tends to be more accurate.

4.4 The Effect of p-value Combination Methods in Rolling Window Analysis

We compare the six methods by investigating the size, the power and the size-adjusted power. The size-adjusted power is the power adjusted by the $\mu_t = 0$ Case. We investigate all the methods with errors u_t generated from our AR (1) models.

We compare the methods with a fixed window size 25. In AppendixD, we also present

the methods comparisons with window size 30.

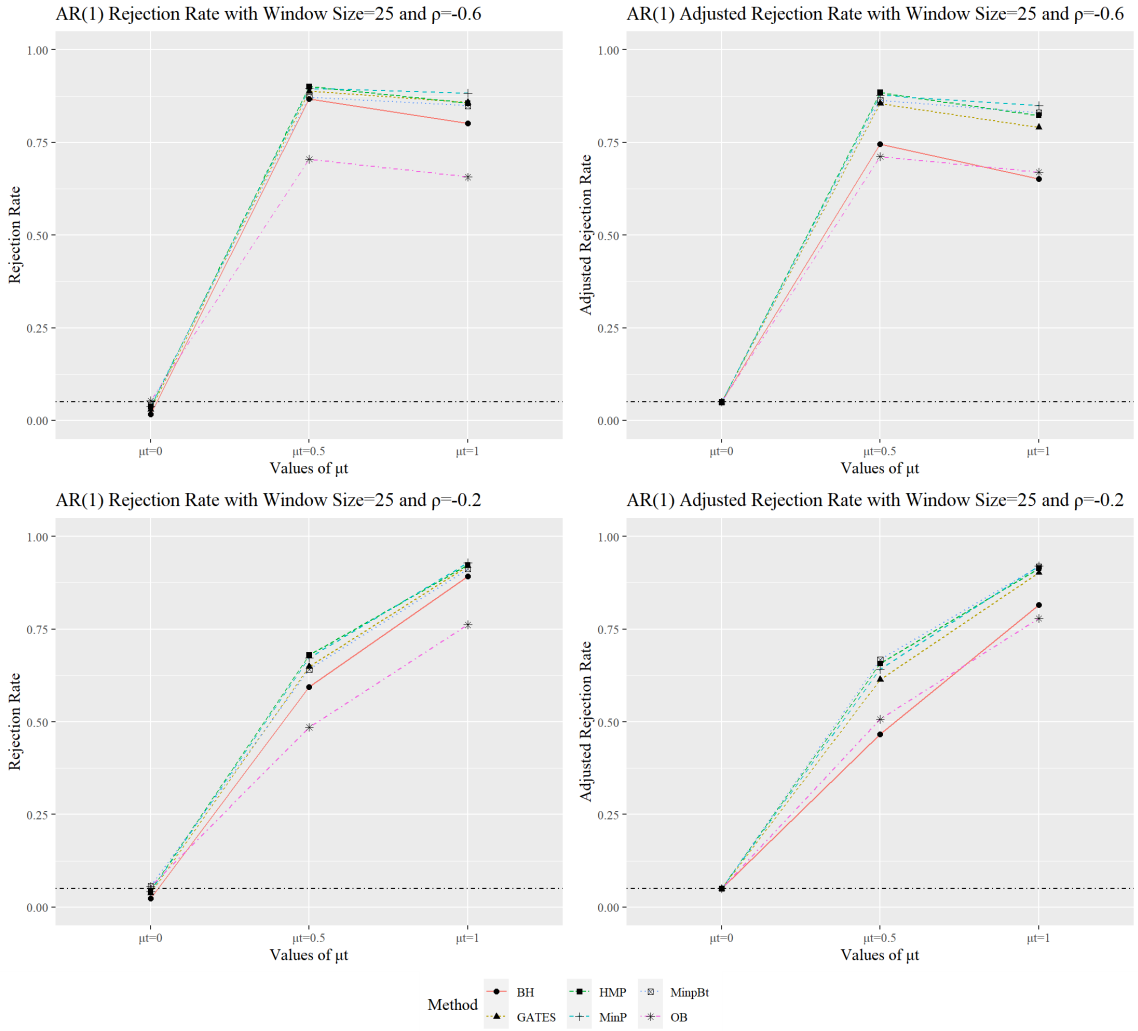

Figure 4.11: This figure shows the rejection rate comparisons for the six methods with fixed window size 25 and six AR(1) coefficients: -0.6, -0.2

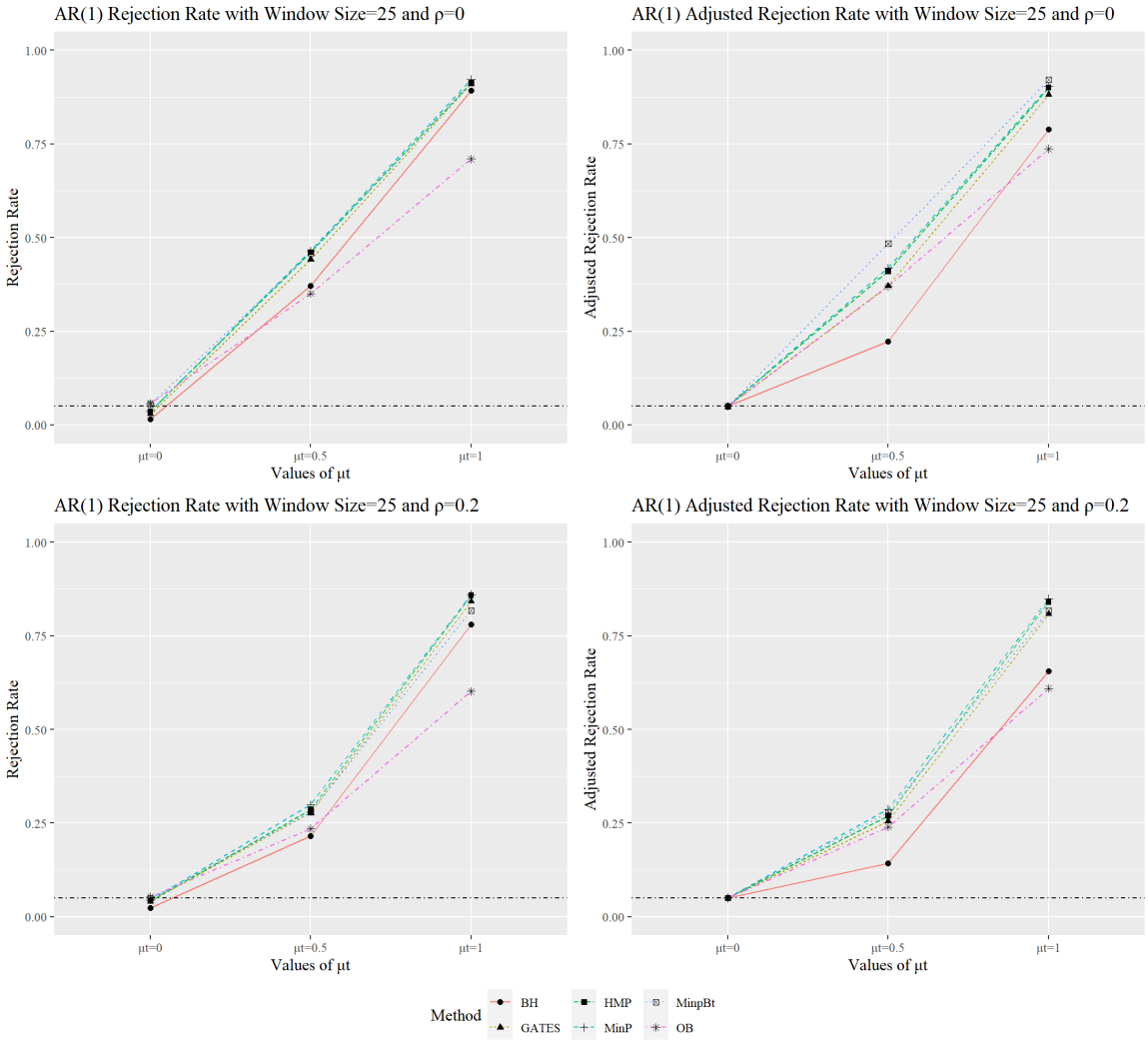

Figure 4.12: This figure shows the rejection rate comparisons for the six methods with fixed window size 25 and six AR(1) coefficients: 0, 0.2

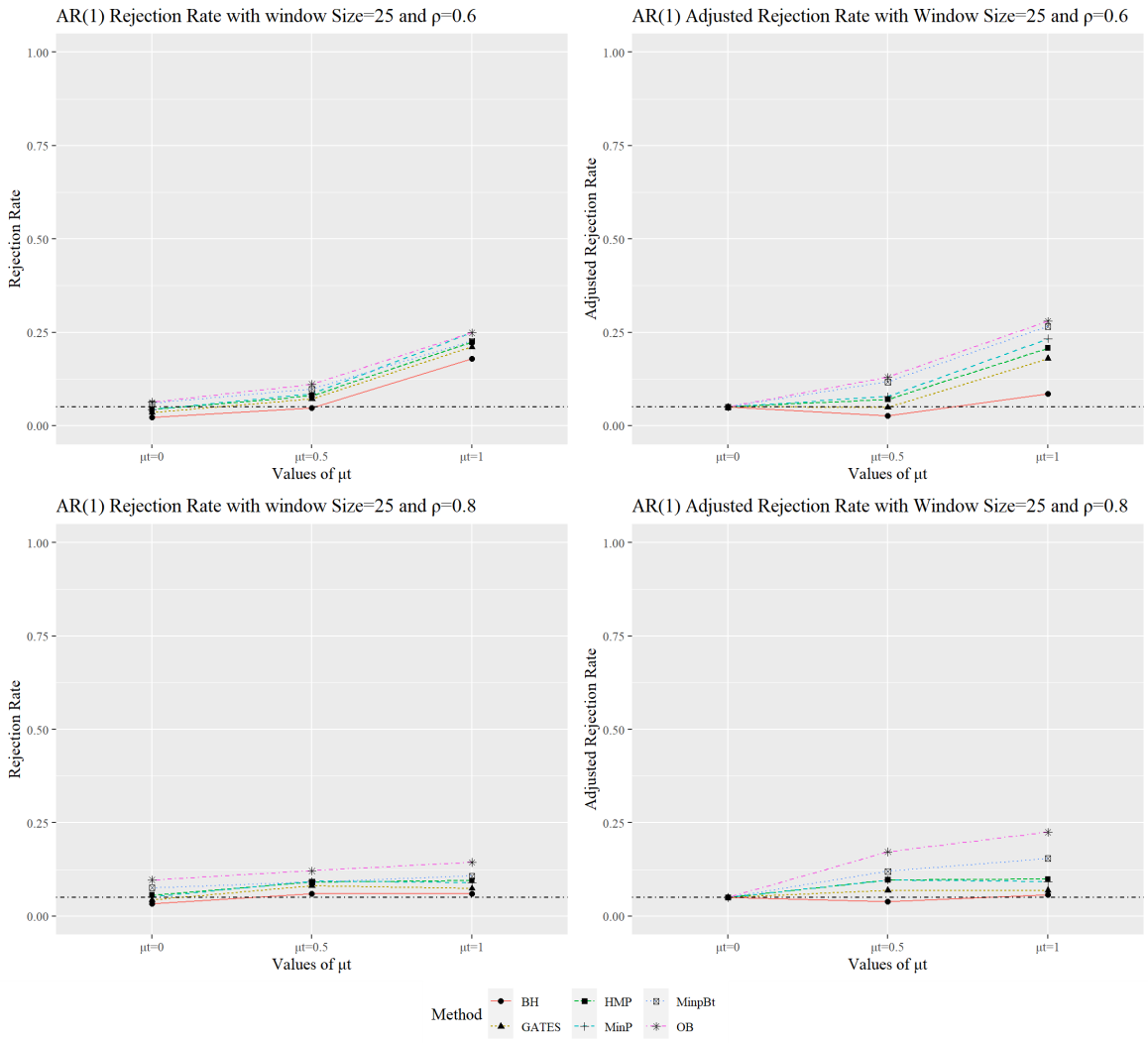

Figure 4.13: This figure shows the rejection rate comparisons for the six methods with fixed window size 25 and six AR(1) coefficients: 0.6, 0.8

Figures 4.11 – 4.13 exhibit the method comparisons. In general, the power tends to decrease as the absolute value of ρ increases. In particular, when $\rho < 0$, the power is better than the power generated with positive correlation. In addition, as

μ increases, the power is inclined to increases, except for one case-relatively strong negative correlation in original data. When the AR coefficient ρ is -0.6, the raw and adjusted rejection rates decrease as μ_t increases from 0.5 to 1. This is because when $\mu = 1$, the original p-value are not ideally near the peak in histogram, which has been discussed in Section 4.1. and give figure name... decreasing power may be due p-values, but not necessarily due to the p-value combination methods. It seems that AR sieve approximation might be problematic under the alternative when the strong negative correlation exists. When generating data with negative AR(1) coefficients, the rejection rates increase slowly when μ is from 0.5 to 1 but sharply when when μ is from 0 to 0.5. In contrast, when generating data with positive AR(1) coefficients, it has opposite behaviors that rejection rates increase sharply when μ is from 0.5 to 1.

Comparing the methods, as we expected, BH is the most conservative and has the lowest raw and size-adjusted powers for all settings under considerations. The performance of BH gets worse, compared to other methods, as dependence gets stronger in either direction. GATES is less conservative compared to BH but not providing enough power when dependence are strong in the data generating process. HMP controls the size best among all methods at the nominal level 5% at the expense of sacrificing the power. OB has a slightly stronger rejection strength than other methods when the data generating process using positively strong coefficients. MinP and MinpBt are more powerful compared to other methods. MinP can control size and

has the highest power in most cases.

In general, we recommend methods based on MinP, which has already been used much in literature. However, it might be cumbersome to compute the estimated covariance matrix for test statistics. In particular, MinpBt is has accurate size an high powe, but takes longer time to run and may be complicated to construct it. If the research focuses strong dependency and more concerned about size than the power, then we recommend HMP. HMP also has the advantage of simpler computations, serving as an attractive alternative to practitioners. GATES is not ideally to be used in practice since GATES needs the six-order polynomials, which might be difficult to be obtained.

Chapter 5

Conclusion and Discussion

We study the multiple testing correction methods in GWAS and adapt them to the time series rolling window analysis. The dependency structure, in GWAS, can be obtained from an external source, which does not exist in time series. We propose to approximate the unknown time series data structure to an autoregressive (AR) model by adopting the idea of AR sieve. The AR sieve idea was used for two purposes. One is to obtain better p-values, and the other is to approximate the dependence structure among the test statistics from rolling windows. The AR sieve idea works well in general, producing well-behaving p-values both under the null and the alternative. When the true data cannot be approximated by a simple AR model, the AR sieve method is not as effective but is still competitive to its competitor based on the heteroskedasticity and autocorrelation consistent estimators.

Based on our simulations, we suggest using 25% to 30% of the total length of time series as the window size. This choice is consistent with Shi et al. [41]'s minimum fixed rolling window size suggestion. With a fixed window size, we analyze how the strength and direction of dependencies influence the size, the power and the size adjust power. For negative AR coefficients in the data generating process, the time series data tends to oscillate back and forth across the mean, then the estimated mean is more accurate. However, when the AR coefficients are positively strong, we need to be cautious. The methods we adopted from GWAS can be applied to the rolling window setting but use with caution. We recommend to use MinP if the correlation structure is easy to be obtained, to implement HMP if the research focuses on the strong dependency and concerned about the size more than power, and to adopt MinpBt if program running time is not important.

There are lots of work need to be discussed in the future work. First is the data structure approximation problem. In this article, we assume the error is the AR process but the assumption constraints other non AR distribution processes. From the simulation, AR sieve approach seems hard to provide good p-values before combing them when the strong dependency exists in data. This problem also need more future work to be addressed Another one is computing the theoretical covariance and correlation matrices of the test statistics. Since our test is simple, the matrix is not difficult to calculate. Other complex test statistics might be difficult to building up the theoretical matrices of test statistics.

References

[1] Alonso, A. M., D. Peña, and J. Romo (2003). On sieve bootstrap prediction intervals. *Statistics & Probability Letters 65*(1), 13–20.

[2] Andre'es, M. A., D. Pena, and J. Romo (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference 100*(1), 1–11.

[3] Basu, S., Y. Zhang, D. Ray, M. B. Miller, W. G. Iacono, and M. McGue (2013). A rapid gene-based genome-wide association test with multivariate traits. *Human heredity 76*(2), 53–63.

[4] Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57*(1), 289–300.

[5] Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

[6] Bland, J. M. and D. G. Altman (1995). Multiple significance tests: the bonferroni method. *Bmj 310*(6973), 170.

[7] Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, 123–148.

[8] Bühlmann, P. (1998). Sieve bootstrap for smoothing in nonstationary time series. *The Annals of Statistics 26*(1), 48–83.

[9] Bühlmann, P. (2002). Bootstraps for time series. *Statistical science*, 52–72.

[10] Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics 136*(1), 163–188.

[11] Cole, D. A., S. E. Maxwell, R. Arvey, and E. Salas (1994). How the power of manova can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological bulletin 115*(3), 465.

[12] Conneely, K. N. and M. Boehnke (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics 81*(6), 1158–1168.

[13] Cryer, J. D. and K.-S. Chan (2008). *Time series analysis: with applications in R*, Volume 2. Springer.

[14] Dangl, T. and M. Halling (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics 106*(1), 157–181.

[15] Efron, B. (1997). The length heuristic for simultaneous hypothesis tests. *Biometrika 84*(1), 143–157.

[16] Ferreira, M. A. and S. M. Purcell (2009). A multivariate test of association. *Bioinformatics 25*(1), 132–133.

[17] Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics*, pp. 66–70. Springer.

[18] Galesloot, T. E., K. Van Steen, L. A. Kiemeney, L. L. Janss, and S. H. Vermeulen (2014). A comparison of multivariate genome-wide association methods. *PloS one 9*(4), e95923.

[19] Good, I. J. (1958). Significance tests in parallel and in series. *Journal of the American Statistical Association 53*(284), 799–813.

[20] Guidi, F. and M. Ugur (2014). An analysis of south-eastern european stock markets: Evidence on cointegration and portfolio diversification benefits. *Journal of International Financial Markets, Institutions and Money 30*, 119–136.

[21] Härdle, W., J. Horowitz, and J.-P. Kreiss (2003). Bootstrap methods for time series. *International Statistical Review 71*(2), 435–459.

[22] He, Q., C. L. Avery, and D.-Y. Lin (2013). A general framework for association tests with multivariate traits in large-scale genomics studies. *Genetic epidemiology 37*(8), 759–767.

[23] Huang, P., B. C. Tilley, R. F. Woolson, and S. Lipsitz (2005). Adjusting o'brien's test to control type i error for the generalized nonparametric behrens–fisher problem. *Biometrics 61*(2), 532–539.

[24] Kim, J., Y. Bai, and W. Pan (2015). An adaptive association test for multiple phenotypes with gwas summary statistics. *Genetic epidemiology 39*(8), 651–663.

[25] Kreiss, J.-P., E. Paparoditis, and D. N. Politis (2011). On the range of validity of the autoregressive sieve bootstrap. *The Annals of Statistics 39*(4), 2103–2130.

[26] Li, M.-X., H.-S. Gui, J. S. Kwan, and P. C. Sham (2011). Gates: a rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics 88*(3), 283–293.

[27] Li, Q., G. Zheng, Z. Li, and K. Yu (2008). Efficient approximation of p-value of the maximum of correlated tests, with applications to genome-wide association studies. *Annals of human genetics 72*(3), 397–406.

[28] Liu, H. and H. Song (2018). New evidence of dynamic links between tourism and economic growth based on mixed-frequency granger causality tests. *Journal of Travel Research 57*(7), 899–907.

[29] Ma, L., A. J. Grant, and G. Sofronov (2020). Multiple change point detection and validation in autoregressive time series data. *Statistical Papers 61*(4), 1507–1528.

[30] McCaw, Z. R., T. Colthurst, T. Yun, N. A. Furlotte, A. Carroll, B. Alipanahi, C. Y. McLean, and F. Hormozdiari (2022). Deepnull models non-linear covariate effects to improve phenotypic prediction and association power. *Nature communications 13*(1), 1–10.

[31] Minlah, M. K. and X. Zhang (2021). Testing for the existence of the environmental kuznets curve (ekc) for co2 emissions in ghana: evidence from the bootstrap rolling window granger causality test. *Environmental Science and Pollution Research 28*(2), 2119–2131.

[32] Moran, M. D. (2003). Arguments for rejecting the sequential bonferroni in ecological studies. *Oikos 100*(2), 403–405.

[33] Müller, U. K. (2014). Hac corrections for strongly autocorrelated time series. *Journal of Business & Economic Statistics 32*(3), 311–322.

[34] O'Reilly, P. F., C. J. Hoggart, Y. Pomyen, F. C. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin (2012). Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PloS one 7*(5), e34861.

[35] Papież, M. and S. Śmiech (2015). Dynamic steam coal market integration: Evidence from rolling cointegration analysis. *Energy Economics 51*, 510–520.

[36] Phillips, P. C. and D. Sul (2003). Dynamic panel estimation and homogeneity testing under cross section dependence. *The econometrics journal 6*(1), 217–259.

[37] Ray, D. and M. Boehnke (2018). Methods for meta-analysis of multiple traits using gwas summary statistics. *Genetic epidemiology 42*(2), 134–145.

[38] Rémillard, B. (2017). Goodness-of-fit tests for copulas of multivariate time series. *Econometrics 5*(1), 13.

[39] Romano, J. P., A. M. Shaikh, M. Wolf, et al. (2010). Multiple testing. *The New Palgrave Dictionary of Economics*.

[40] Sarkar, S. K. and C. Y. Tang (2021). Adjusting the benjamini-hochberg method for controlling the false discovery rate in knockoff assisted variable selection. *arXiv preprint arXiv:2102.09080*.

[41] Shi, S., S. Hurn, and P. C. Phillips (2020). Causal change detection in possibly integrated systems: Revisiting the money–income relationship. *Journal of Financial Econometrics 18*(1), 158–180.

[42] Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), 626–633.

[43] Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika 73*, 751–754.

[44] Su, C., Y. Xu, H. L. Chang, O.-R. Lobont, and Z. Liu (2020). Dynamic causalities

between defense expenditure and economic growth in china: evidence from rolling granger causality test. *Defence and Peace Economics 31*(5), 565–582.

[45] Swanson, N. R. (1998). Money and output viewed through a rolling window. *Journal of monetary Economics 41*(3), 455–474.

[46] Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. *International journal of forecasting 16*(4), 437–450.

[47] Teslovich, T. M., K. Musunuru, A. V. Smith, A. C. Edmondson, I. M. Stylianou, M. Koseki, J. P. Pirruccello, S. Ripatti, D. I. Chasman, C. J. Willer, et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*(7307), 707–713.

[48] Van der Sluis, S., D. Posthuma, and C. V. Dolan (2013). Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS genetics 9*(1), e1003235.

[49] Wang, T. and R. C. Elston (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *The american journal of human genetics 80*(2), 353–360.

[50] Westman, V. (2021). A small sample study of some sandwich estimators to handle heteroscedasticity.

[51] Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences 116*(4), 1195–1200.

[52] Yang, Q., H. Wu, C.-Y. Guo, and C. S. Fox (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genetic epidemiology 34*(5), 444–454.

[53] Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators.

[54] Zhang, Y., Z. Xu, X. Shen, W. Pan, A. D. N. Initiative, et al. (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. *NeuroImage 96*, 309–325.

[55] Zivot, E. and J. Wang (2003). Rolling analysis of time series. In *Modeling Financial Time Series with S-Plus®*, pp. 299–346. Springer.

Appendix A

Sixth Order Polynomial Fitting

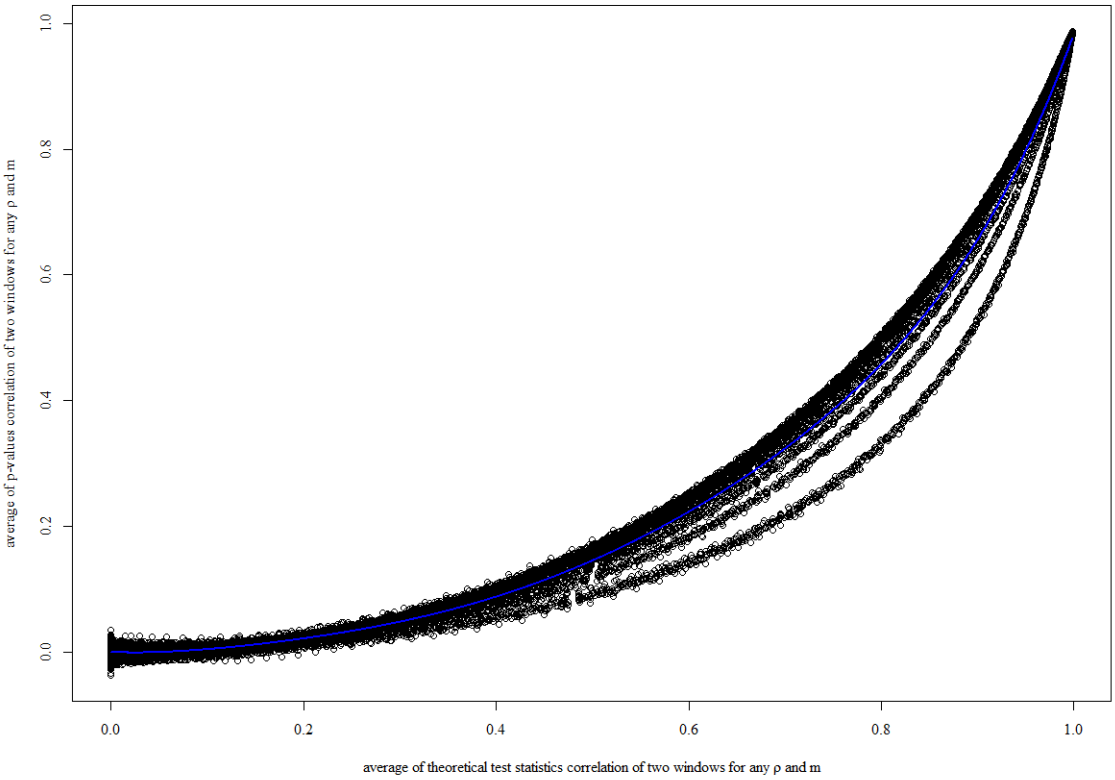

Figure A.1: This figure shows the sixth order polynomial for Avg_p and Avg_z where Avg_p is the response variable. The blue line the fitting line and the coefficient of determination $R^2 = 0.9901$

Appendix B

p-value Distributions

B.1 Data Generated Under the Null on the Error Process AR(1)

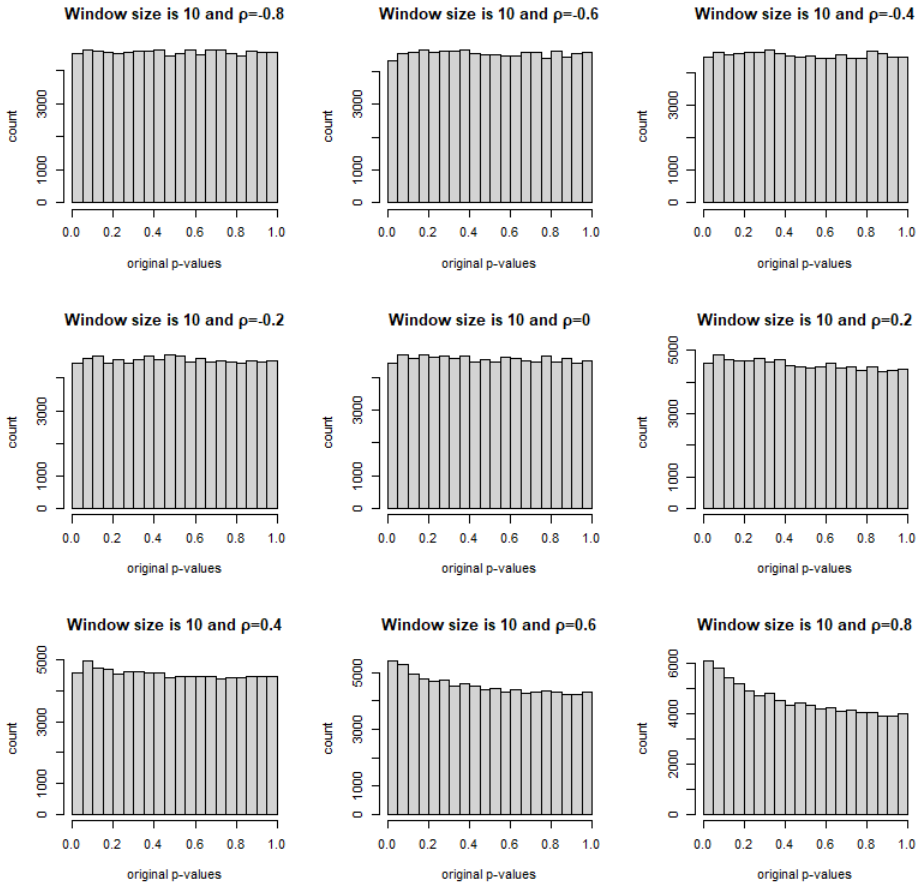

Figure B.1: This figure shows when window size is 10, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

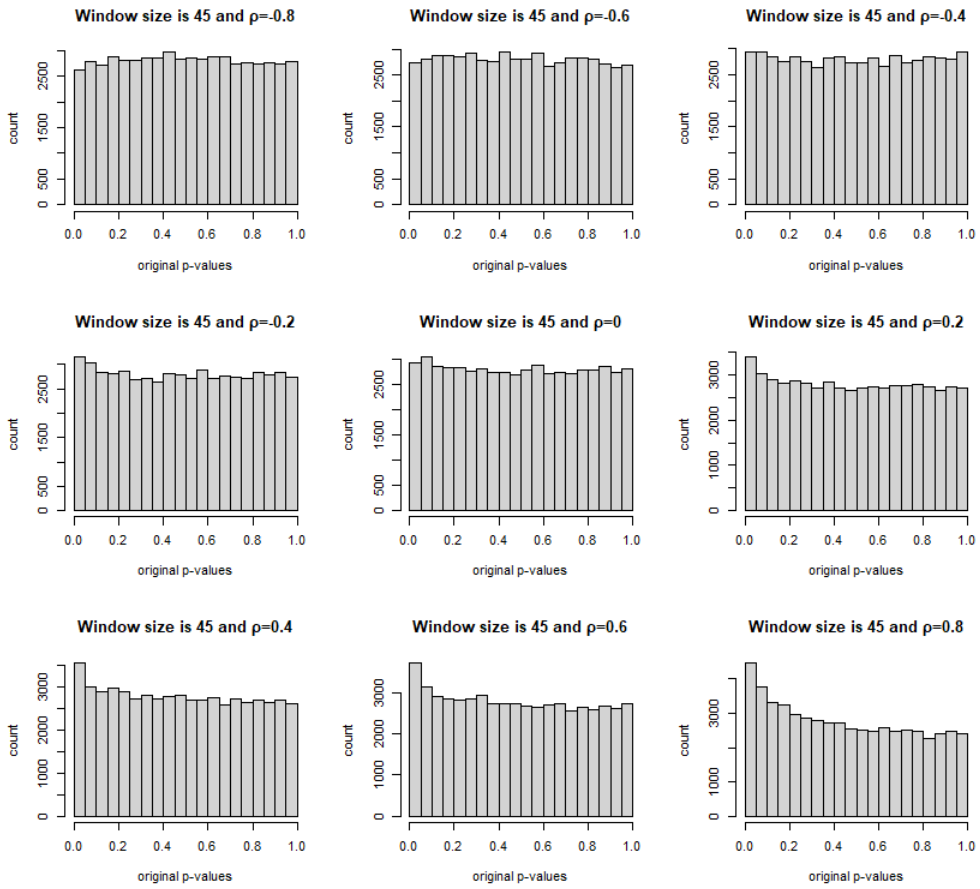

Figure B.2: This figure shows when window size is 45, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

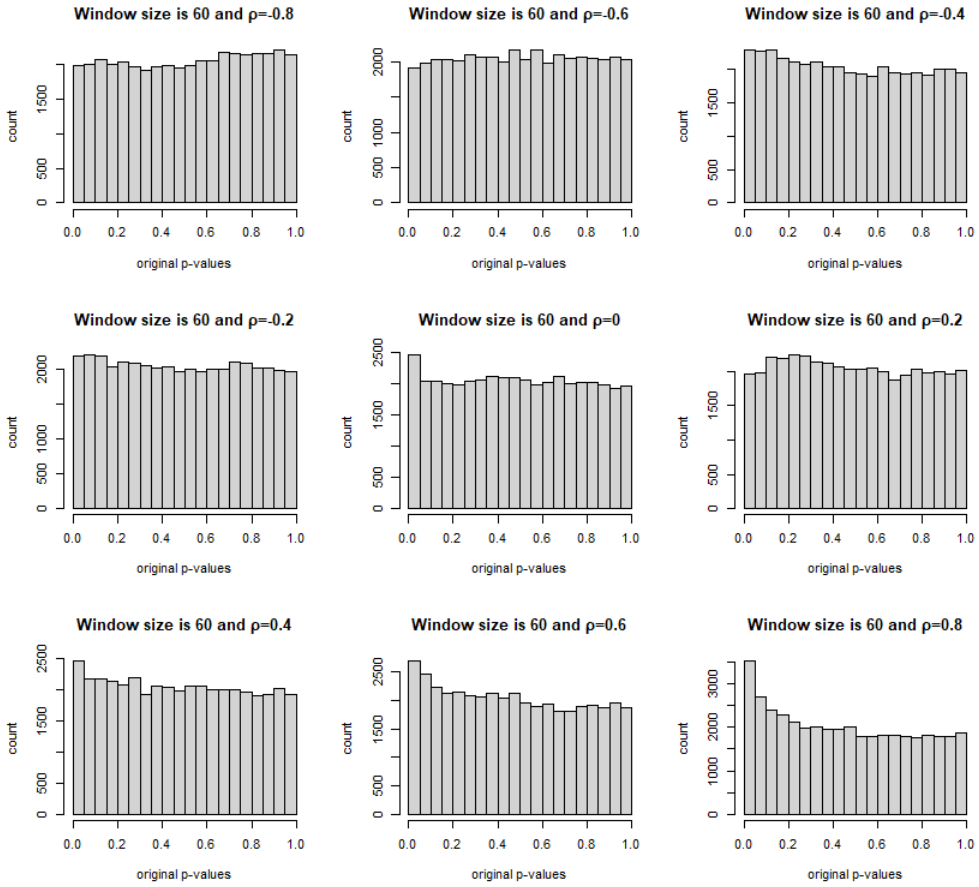

Figure B.3: This figure shows when window size is 60, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

B.2 Data Generated Under the Alternative Case 1 on the Error Process AR(1)

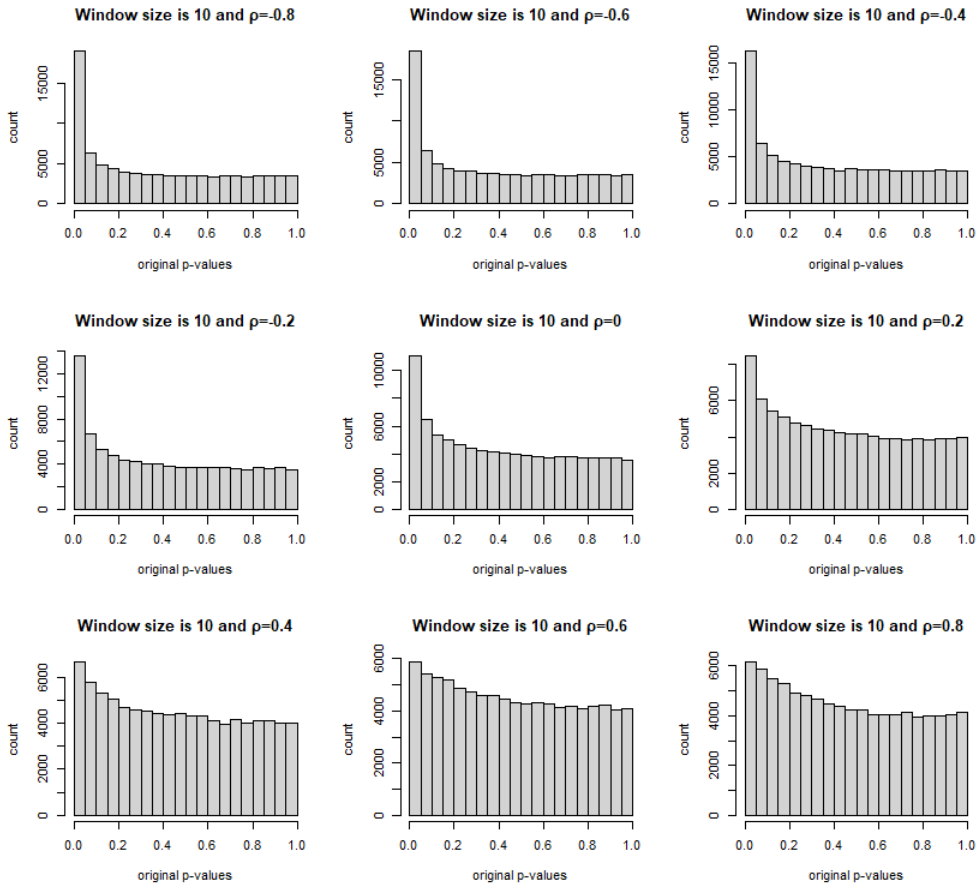

Figure B.4: This figure shows when window size is 10, the p-values behaviors, under the alternative $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

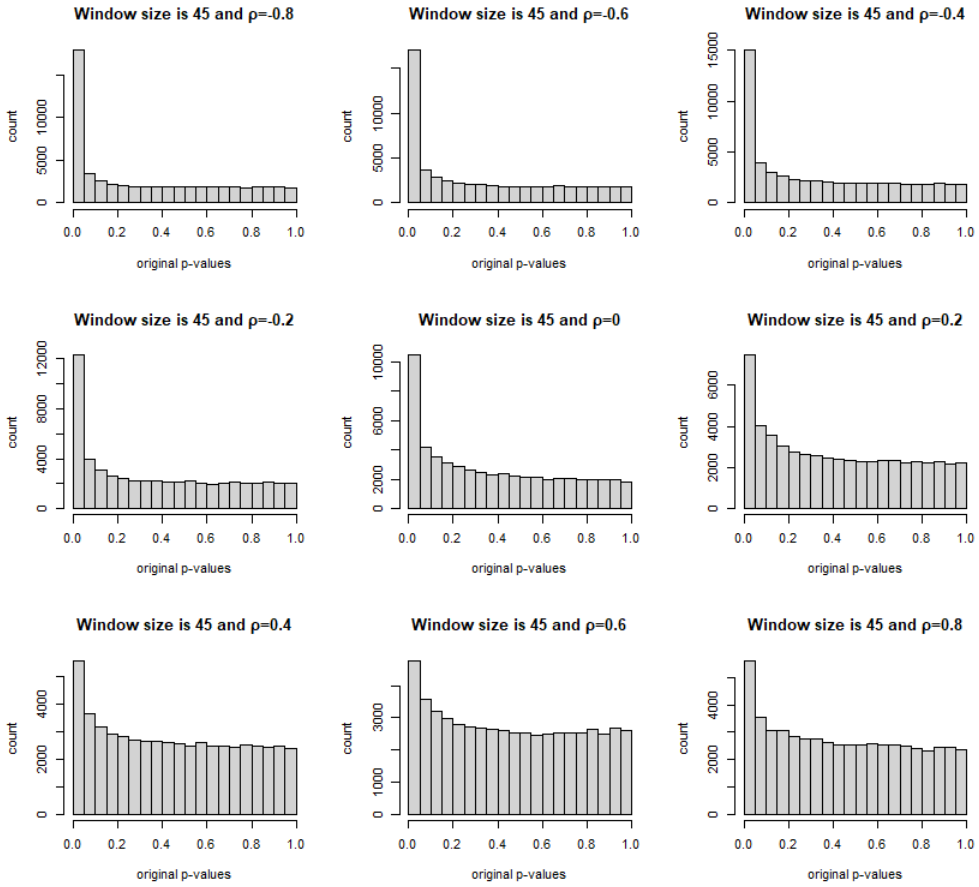

Figure B.5: This figure shows when window size is 45, the p-values behaviors, under the alternative $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

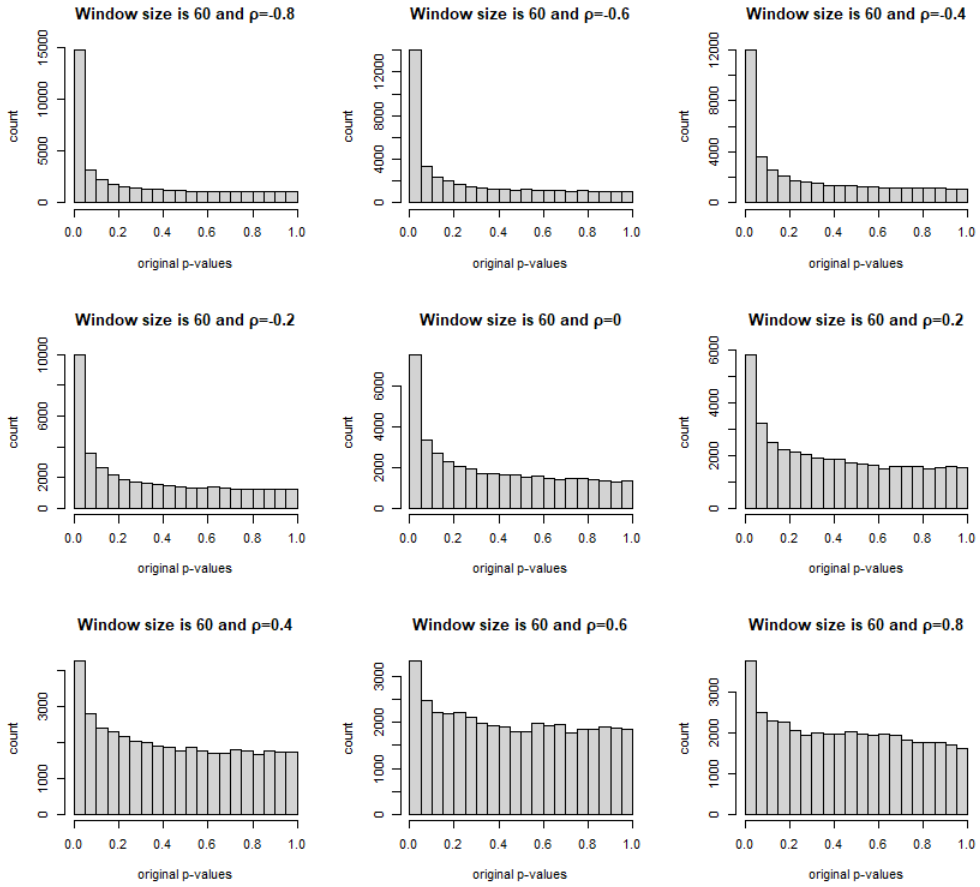

Figure B.6: This figure shows when window size is 60, the p-values behaviors, under the alternative $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

B.3 Data Generated Under the Alternative Case 2 on the Error Process AR(1)

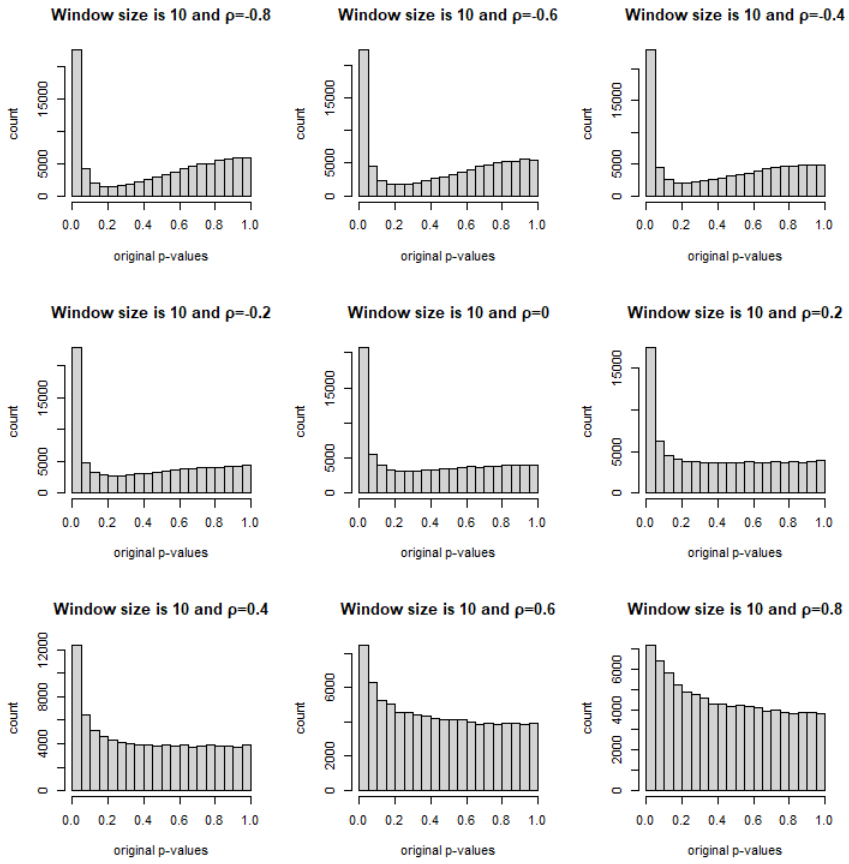

Figure B.7: This figure shows when window size is 10, the p-values behaviors, under the alternative $\mu_t = 1\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

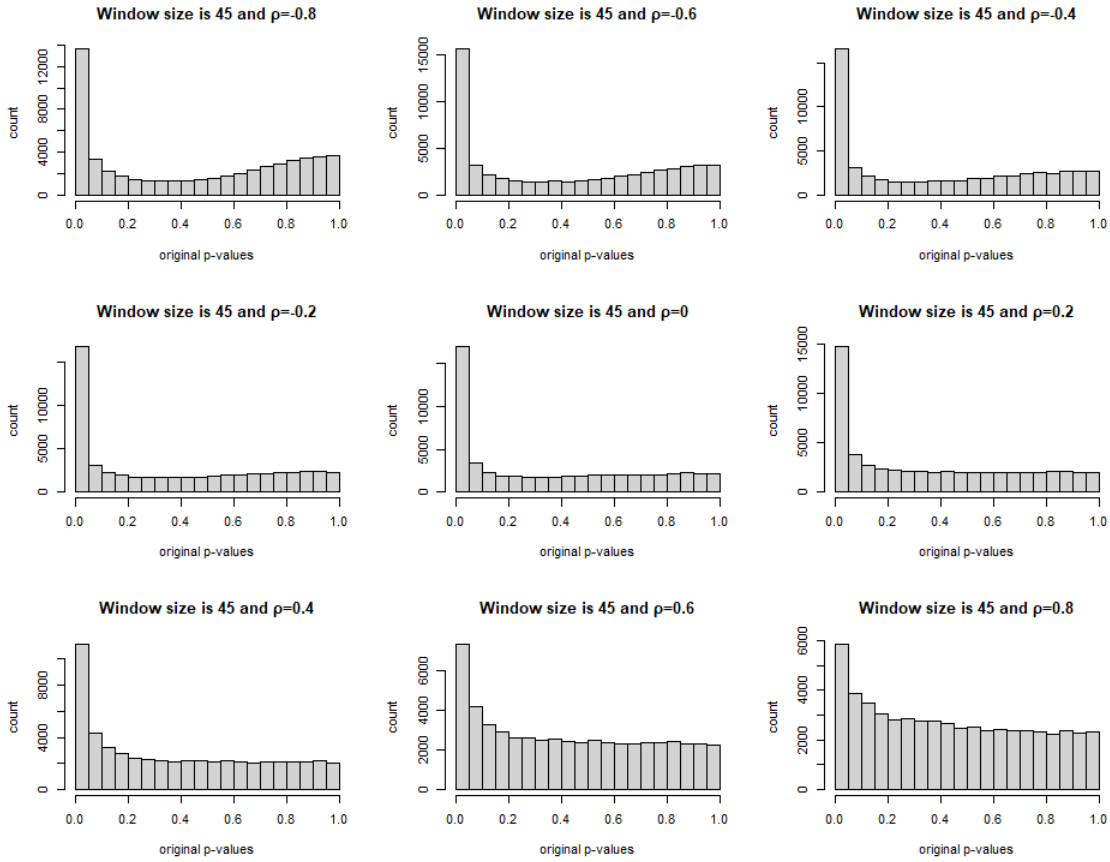

Figure B.8: This figure shows when window size is 45, the p-values behaviors, under the alternative $\mu_t = 1\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

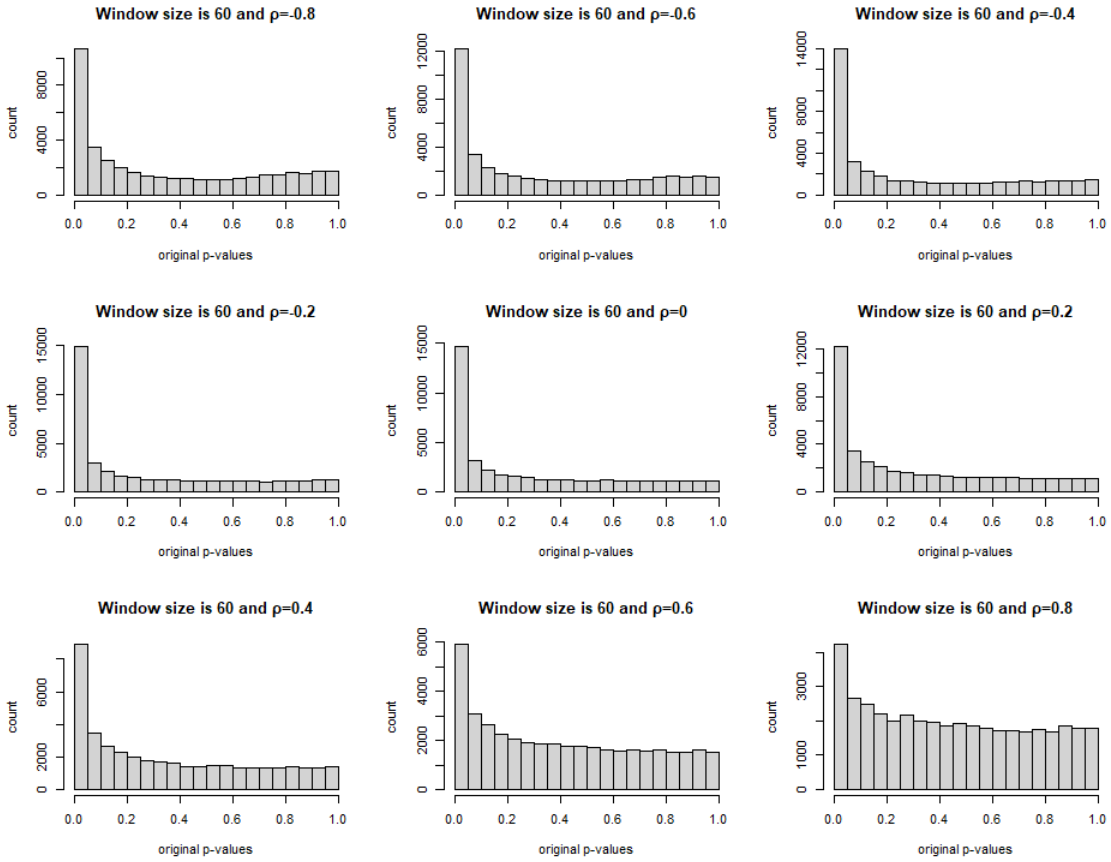

Figure B.9: This figure shows when window size is 60, the p-values behaviors, under the alternative $\mu_t = 1\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$, for each AR(1) coefficient individually in the subfigures

B.4 Data Generated Under the Null on the Error Process ARMA(1,1)

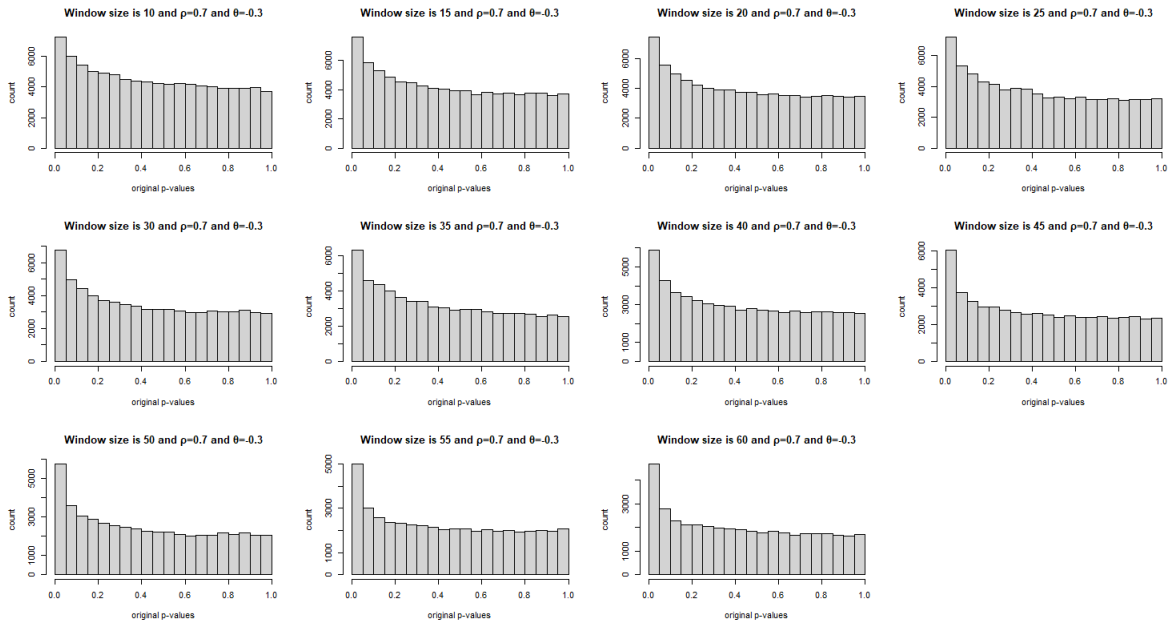

Figure B.10: This figure shows that the p-values behaviors when the error process ARMA(1,1) with $\rho = 0.7$ and $\theta = -0.3$, window size ranges from 10 to 60 by 5, under the null

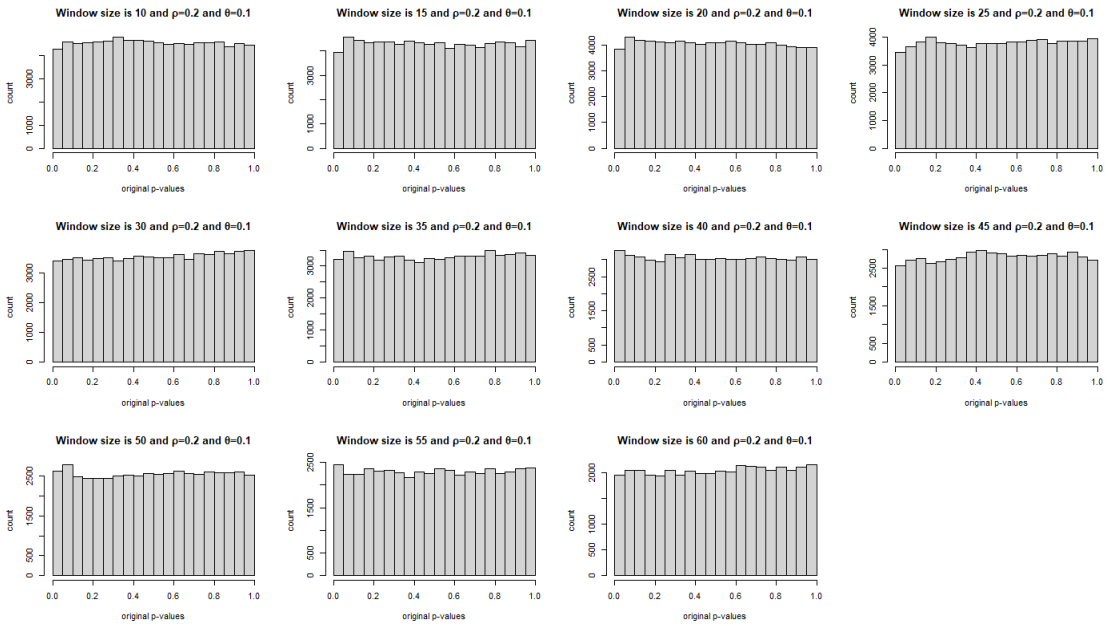

Figure B.11: This figure shows that the p-values behaviors when the error process ARMA(1,1) with $\rho = 0.2$ and $\theta = 0.1$, window size ranges from 10 to 60 by 5, under the null

B.5 Data Generated Under the Null on the Error Process AR(1) Using HAC Estimation

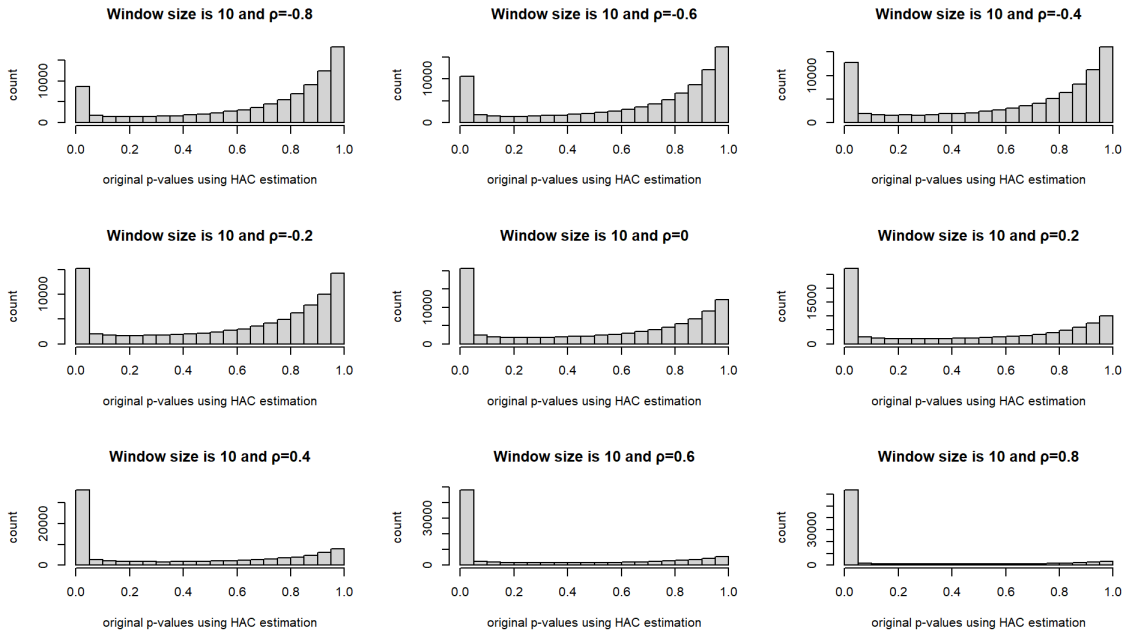

Figure B.12: This figure shows using HAC estimation when window size is 10, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

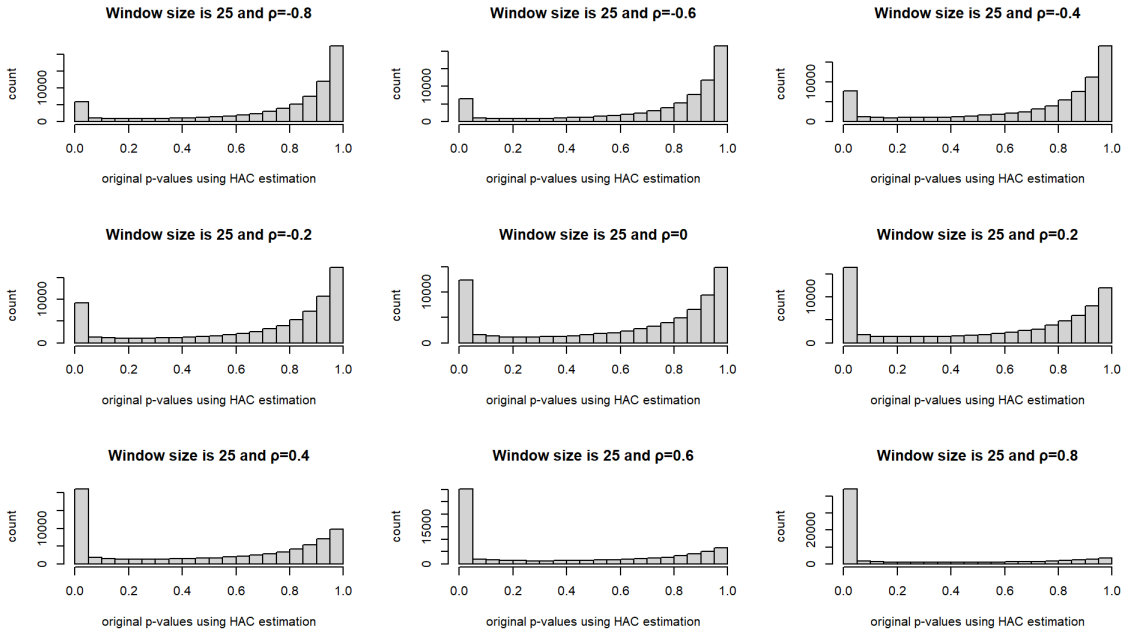

Figure B.13: This figure shows using HAC estimation when window size is 25, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

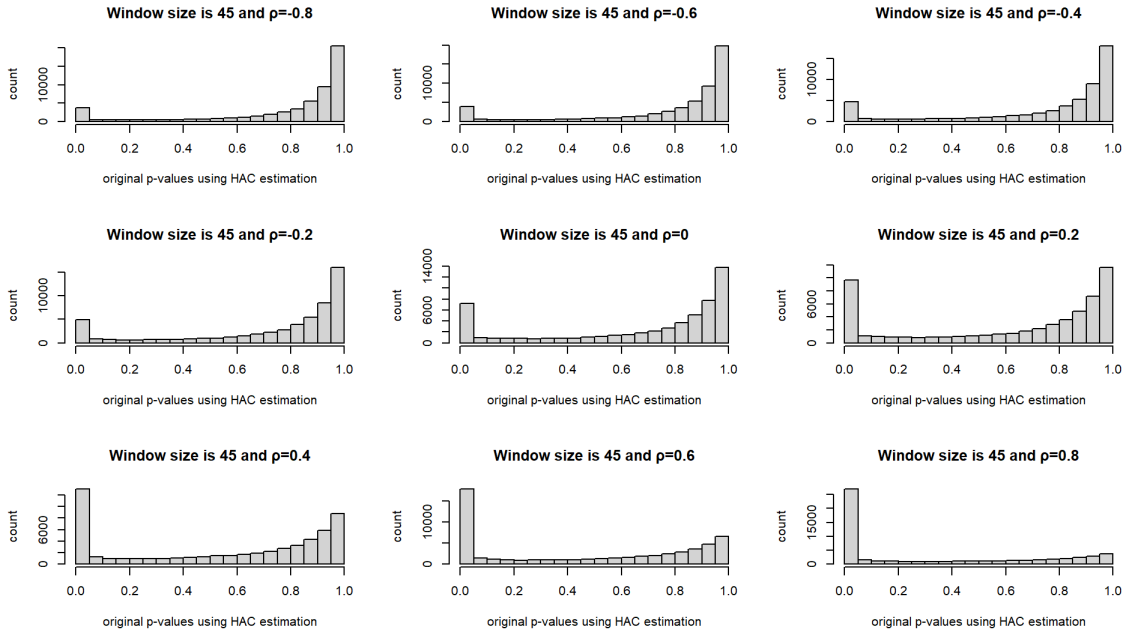

Figure B.14: This figure shows using HAC estimation when window size is 45, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

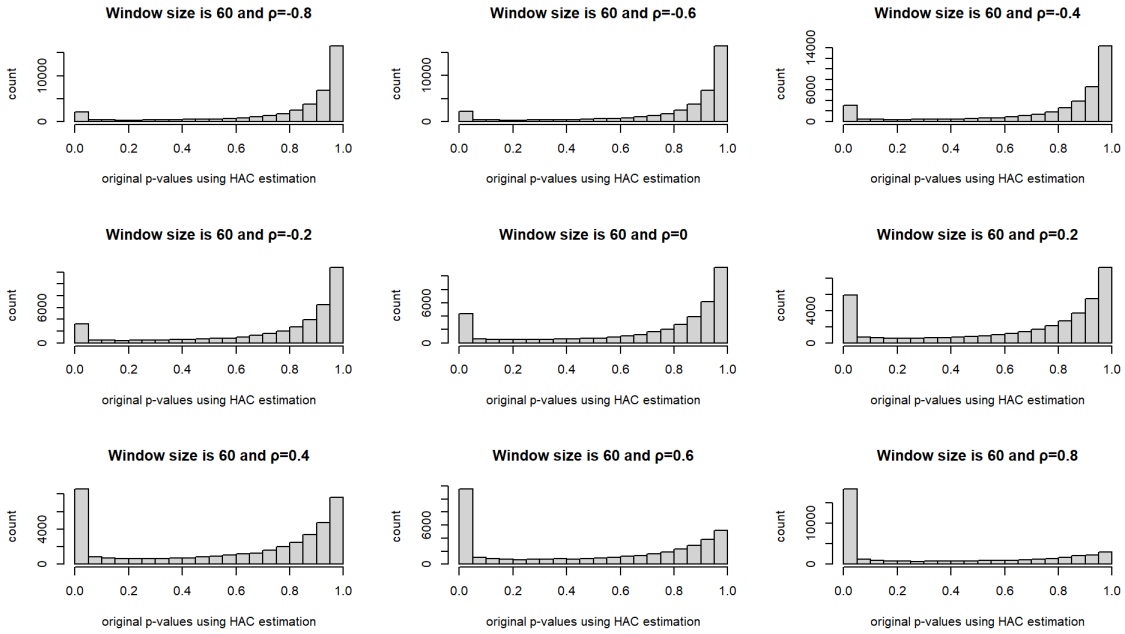

Figure B.15: This figure shows using HAC estimation when window size is 60, the p-values behaviors, under the null, for each AR(1) coefficient individually in the subfigures

Appendix C

Supplementary Simulation Results in 4.2

C.1 Data Generated on the Error Process AR(1)

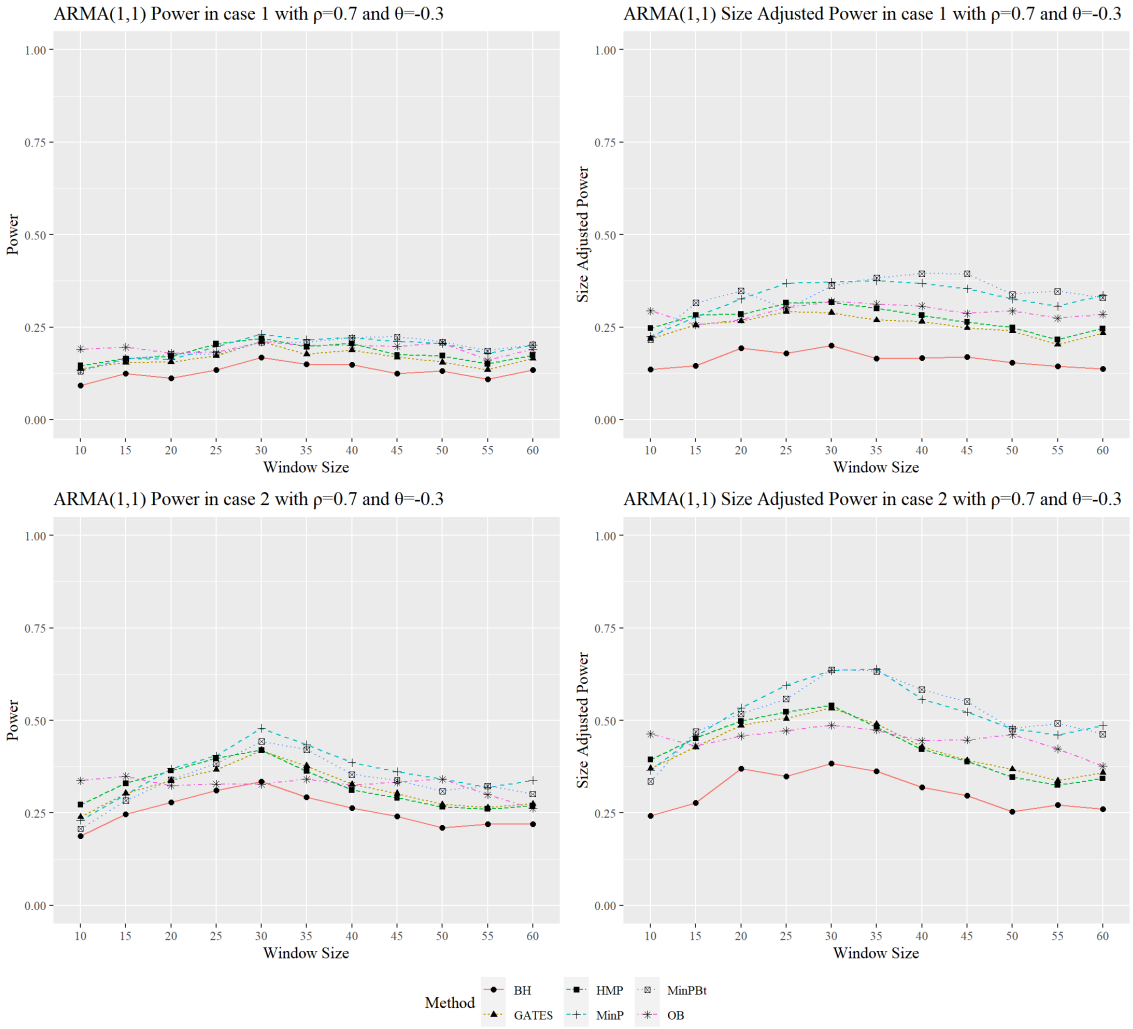

Figure C.1: This figure shows the power and the size adjusted power comparisons for the six methods when ARMA$(1,1)$ coefficient ρ is 0.7 and θ is -0.3. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

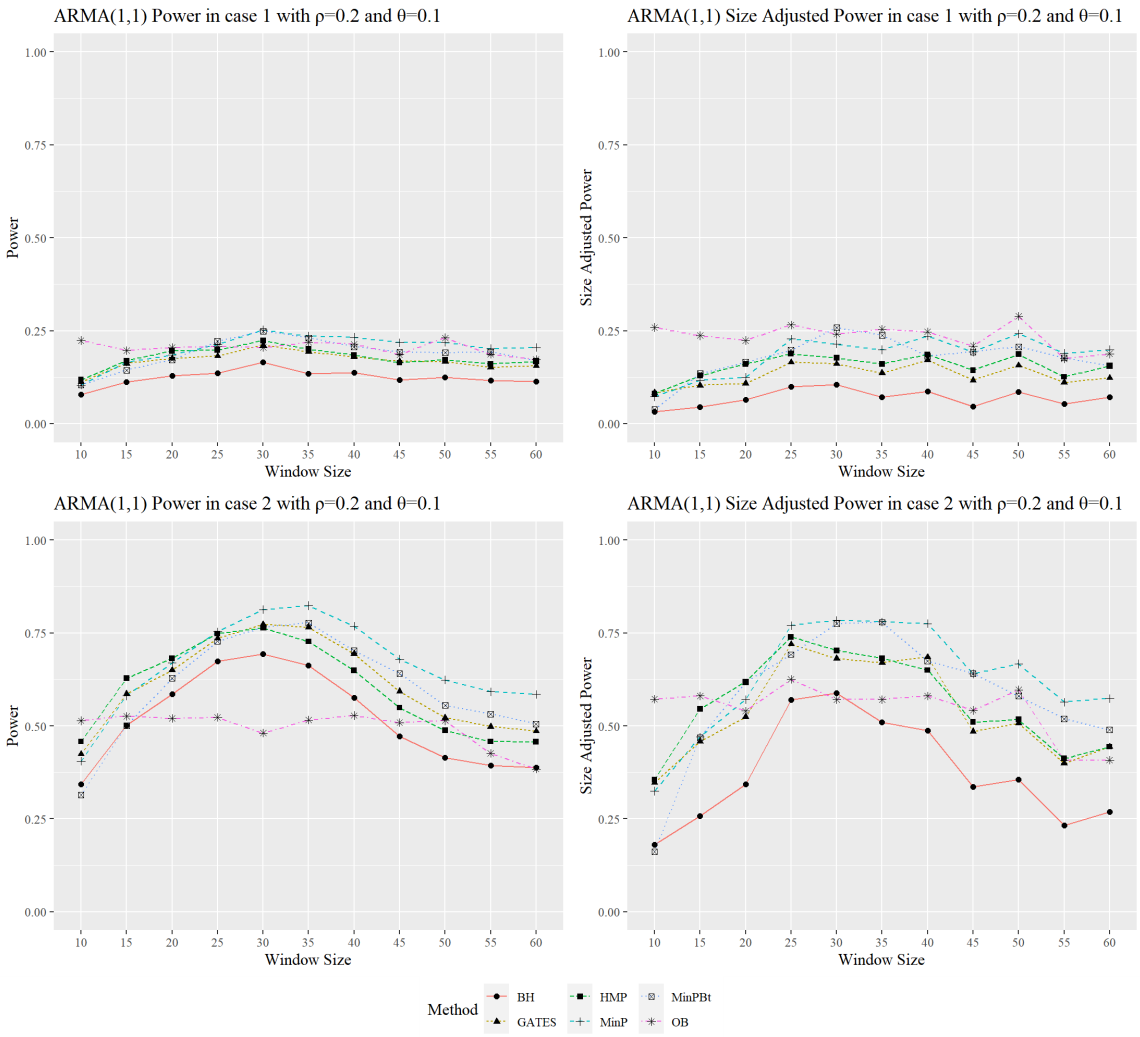

Figure C.2: This figure shows the power and the size adjusted power comparisons for the six methods when ARMA$(1,1)$ coefficient ρ is 0.2 and θ is 0.1. Referring the simulation setting, case 1 is $\mu_t = 0.5\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1\ \{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0\ \{\frac{t}{T} > \frac{1}{3}\}$

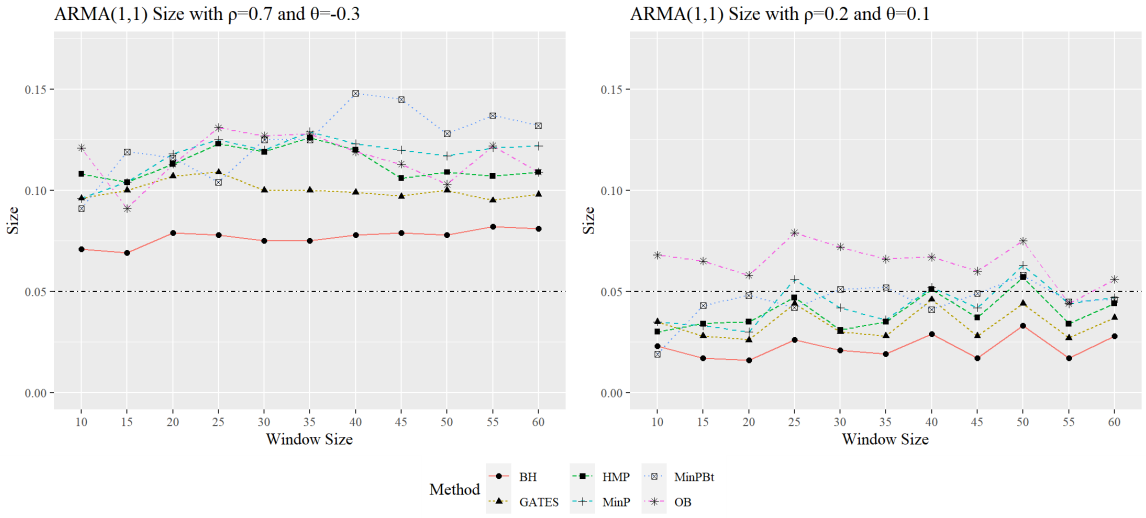

Figure C.3: This figure shows the size comparison for the six methods when ARMA(1,1) coefficients are $(0.7, -0.3)$ and $(0.2, 0.1)$.

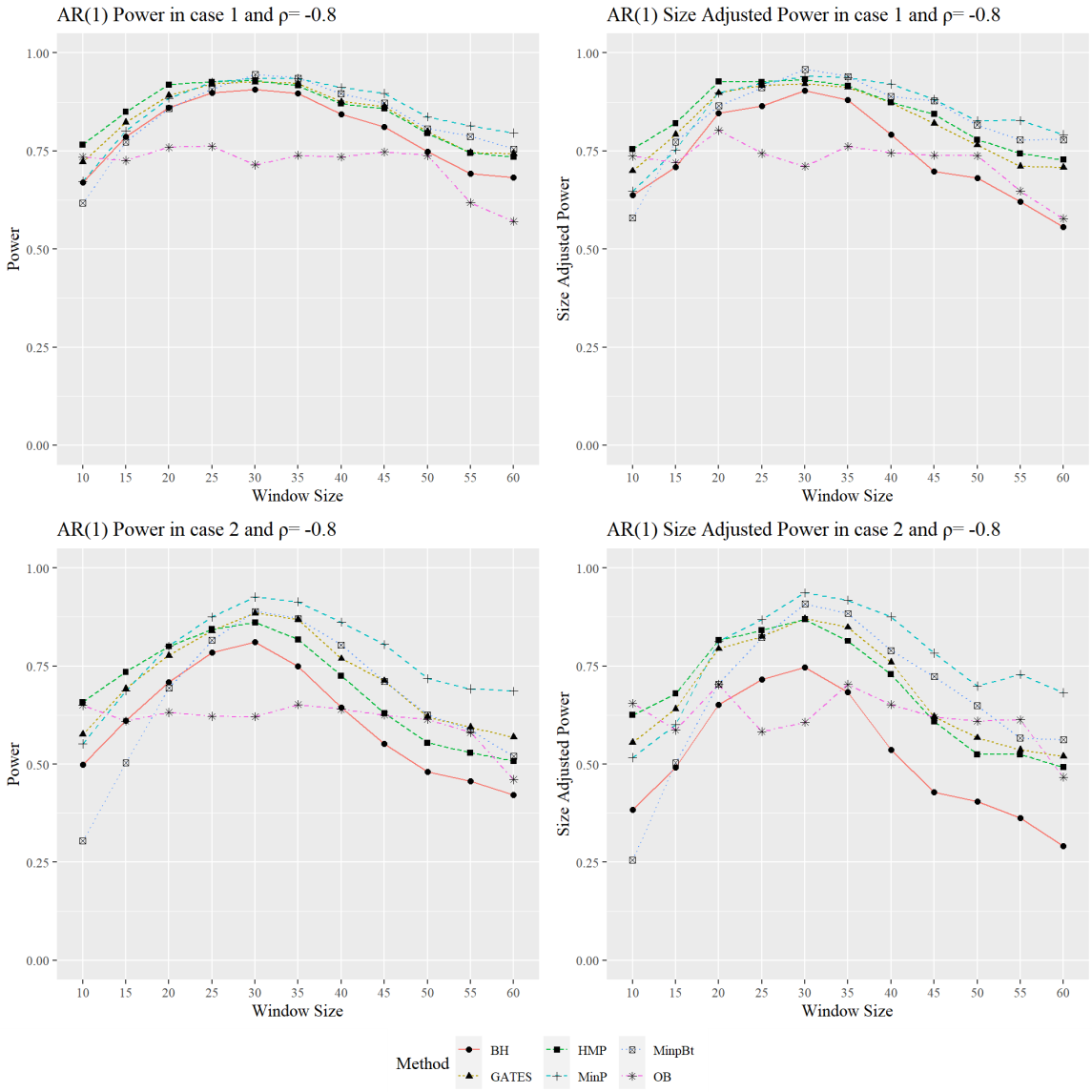

Figure C.4: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is -0.8. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

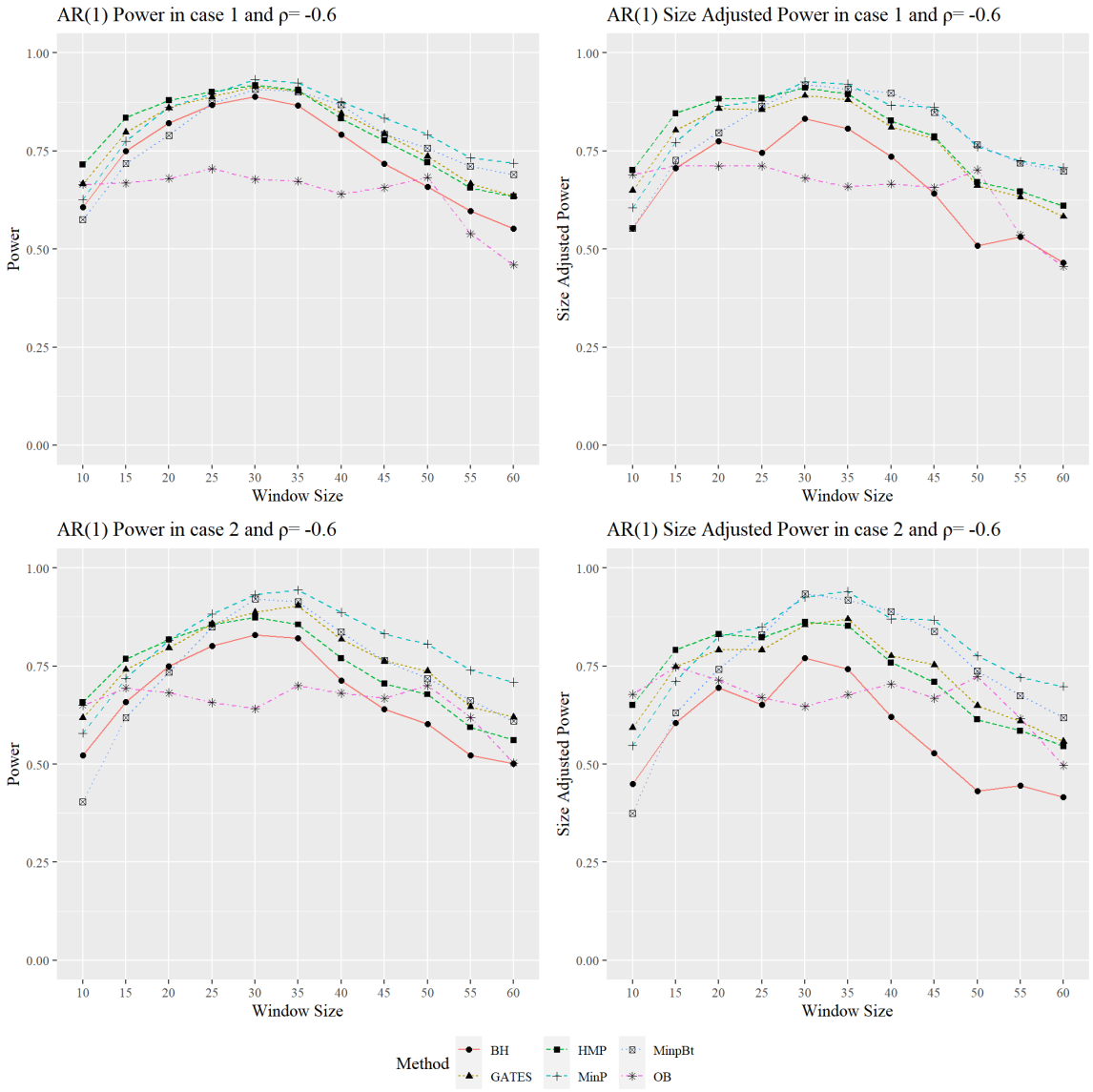

Figure C.5: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is -0.6. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

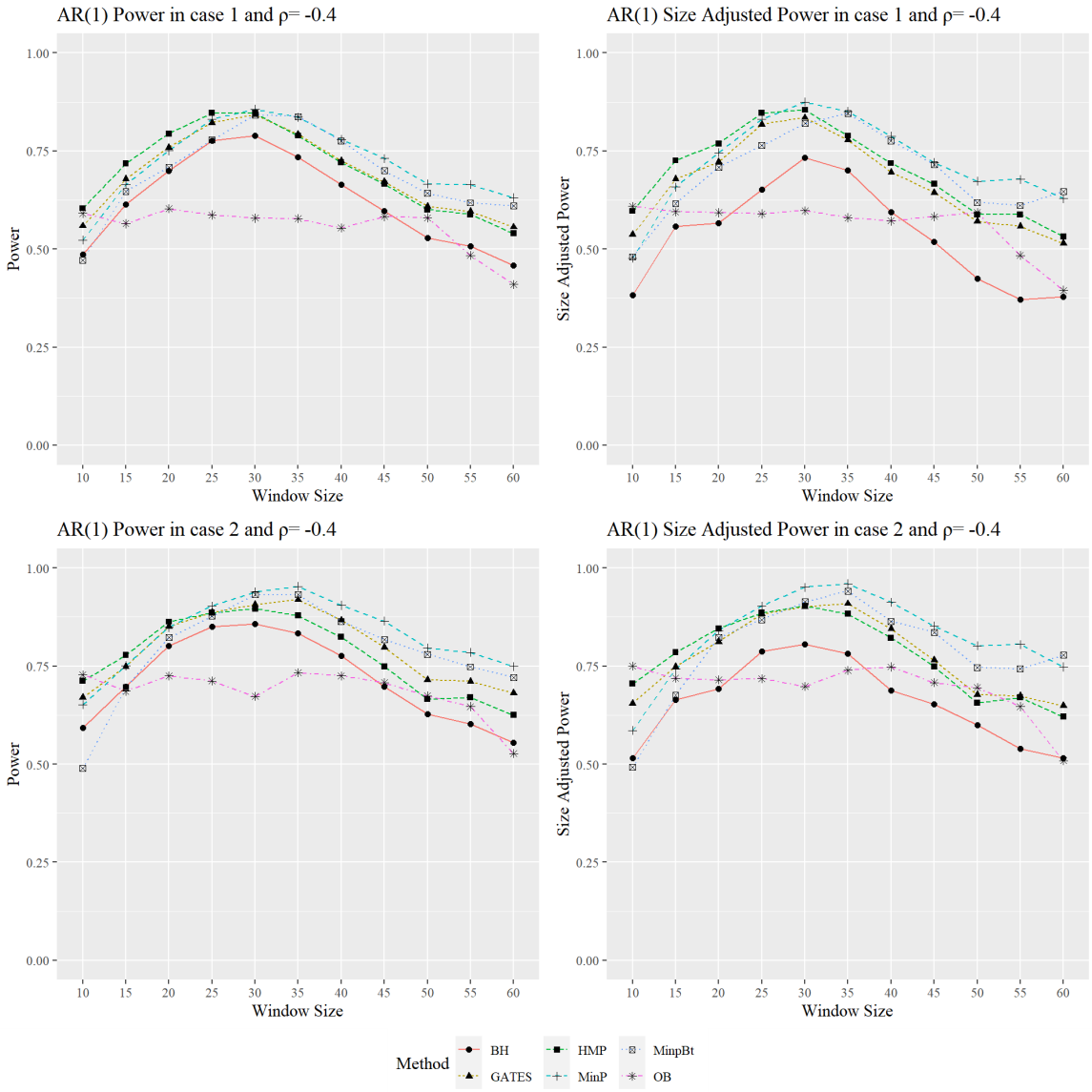

Figure C.6: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is -0.4. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

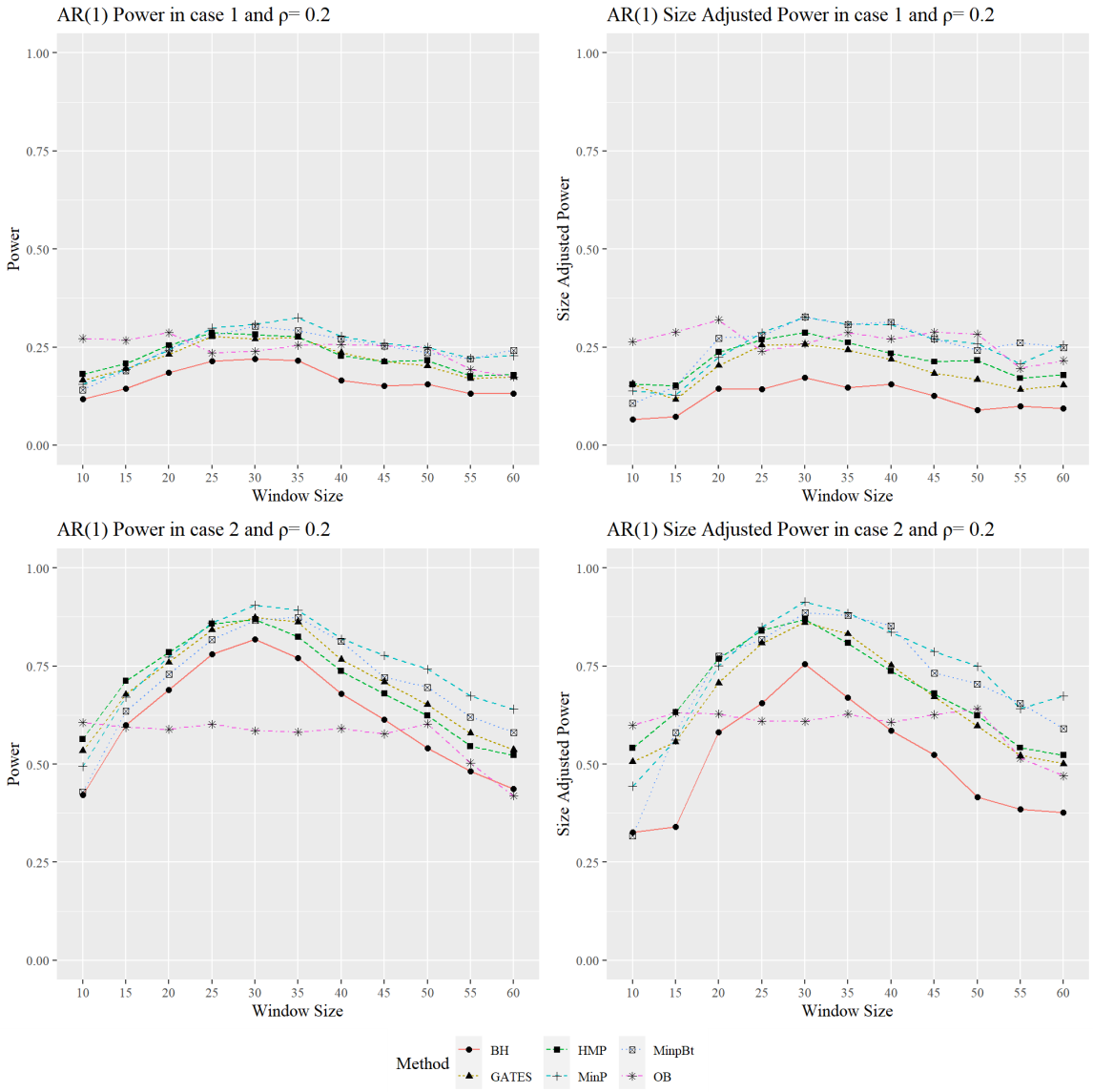

Figure C.7: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is 0.2. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

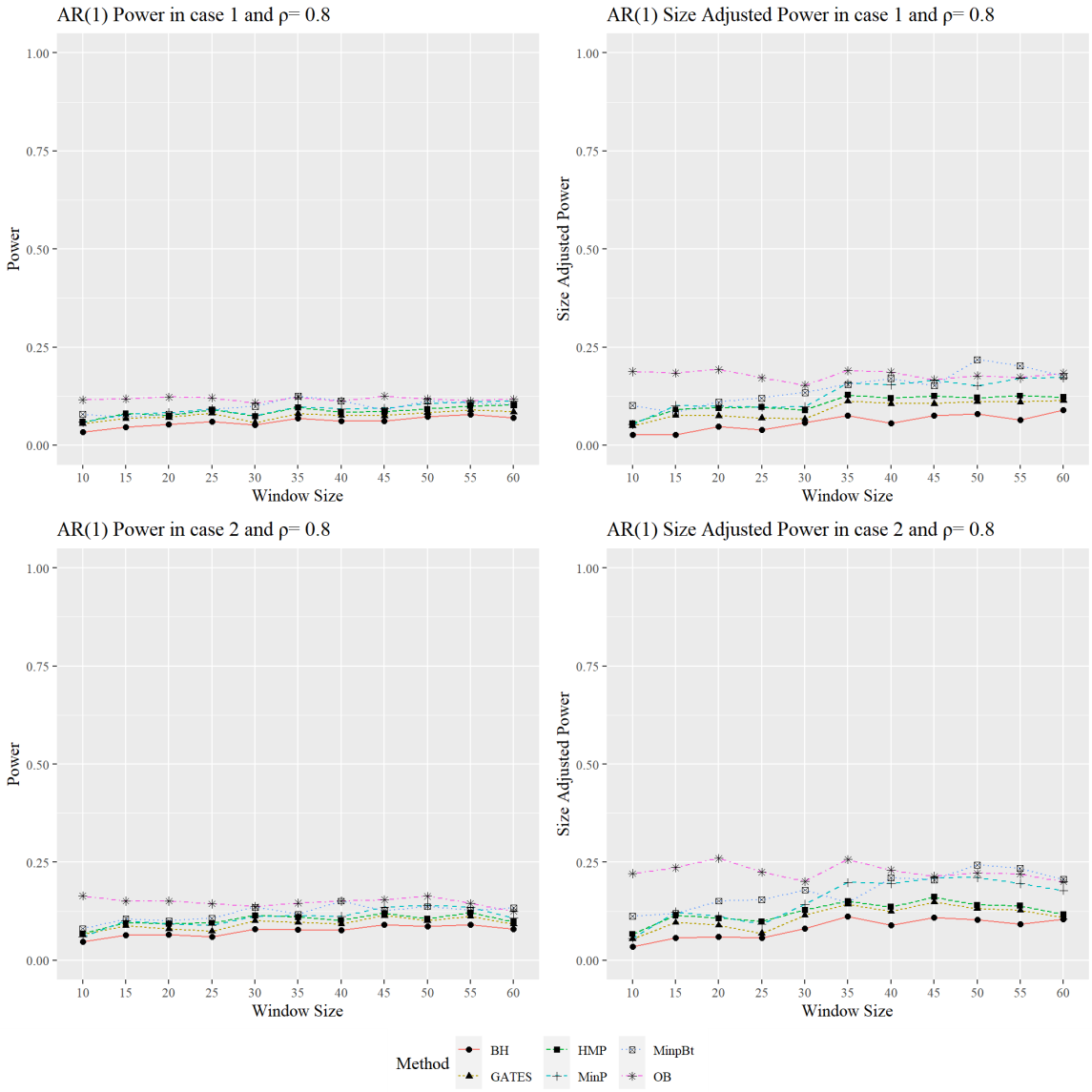

Figure C.8: This figure shows the power and the size adjusted power comparisons for the six methods when AR(1) coefficient ρ is 0.8. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

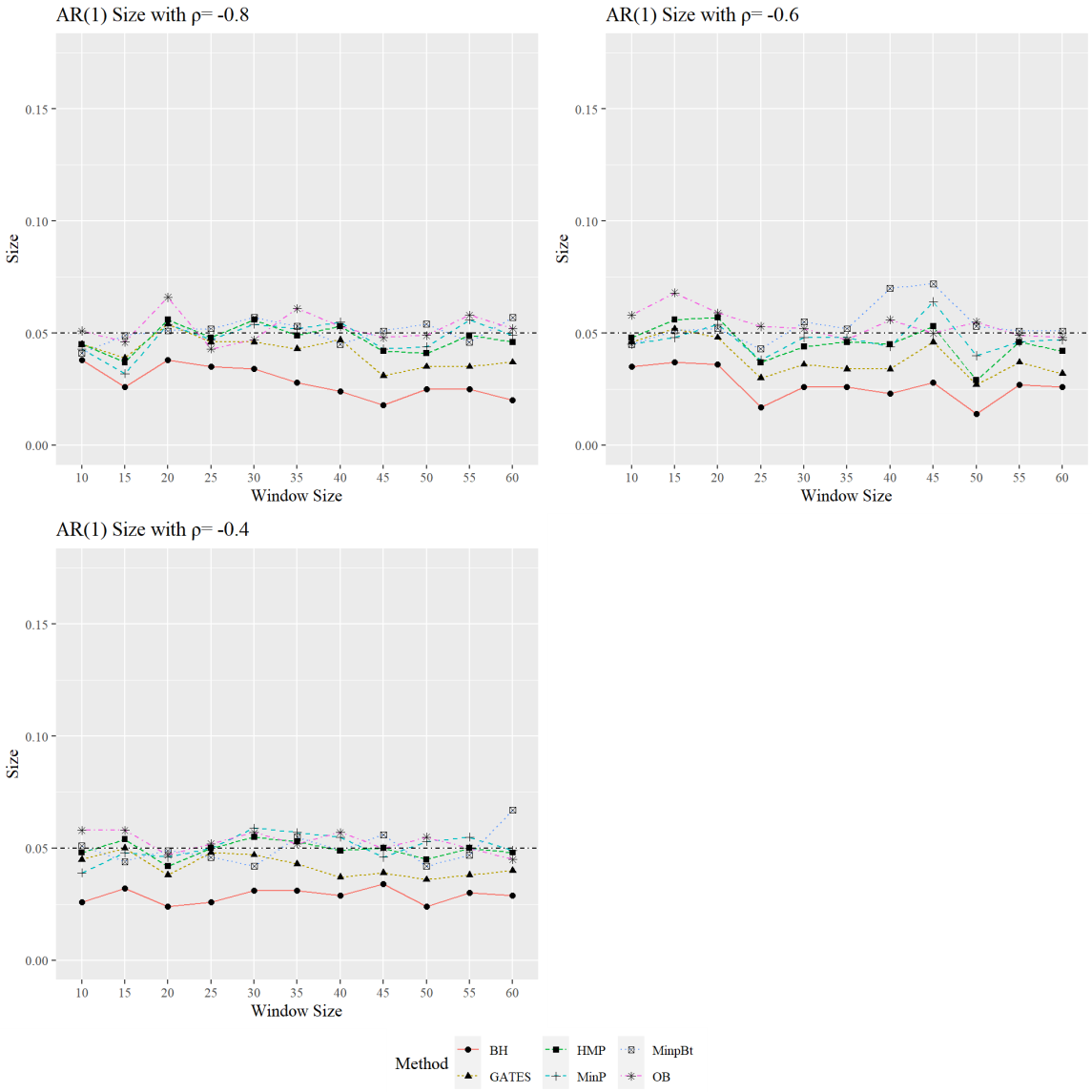

Figure C.9: This figure shows the size comparison for the six methods when AR(1) coefficient ρ is -0.8, -0.6 and -0.4.

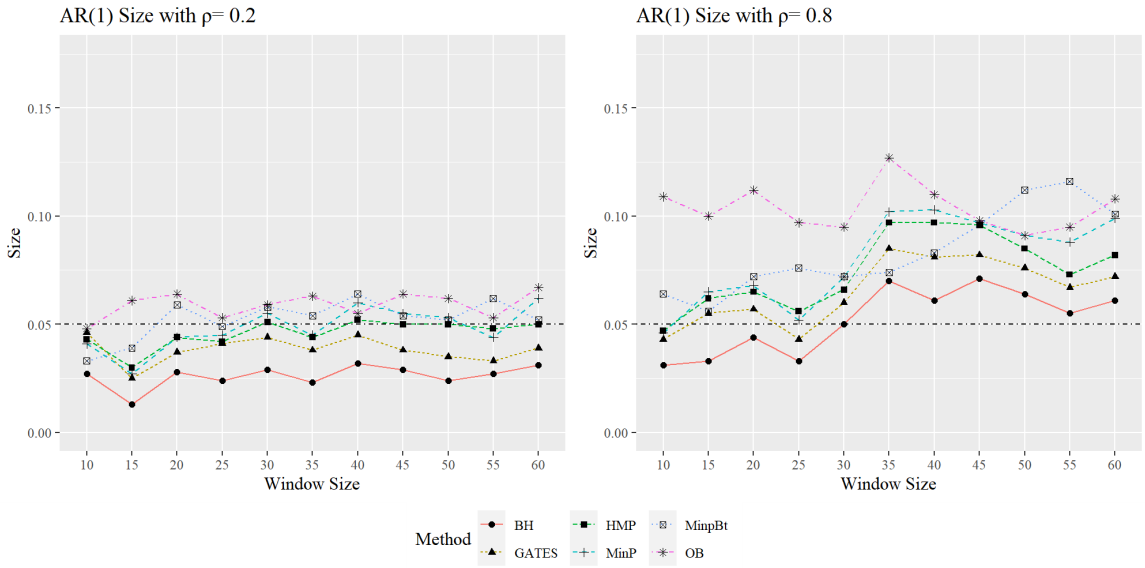

Figure C.10: This figure shows the size comparison for the six methods when AR(1) coefficient ρ is 0.2 and 0.8.

C.2 Data Generated on the Error Process ARMA(1,1)

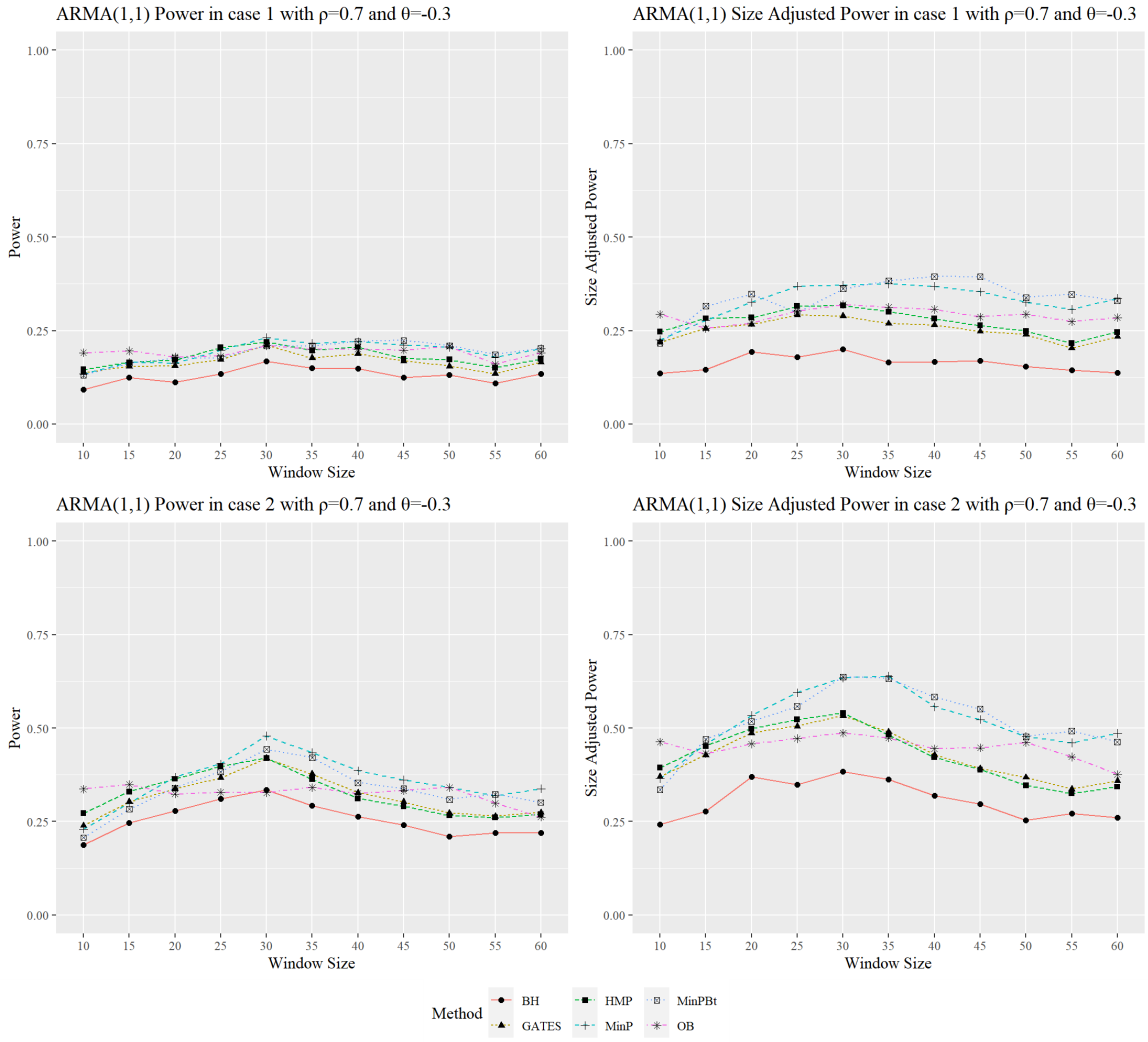

Figure C.11: This figure shows the power and the size adjusted power comparisons for the six methods when ARMA$(1,1)$ coefficient ρ is 0.7 and θ is -0.3. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

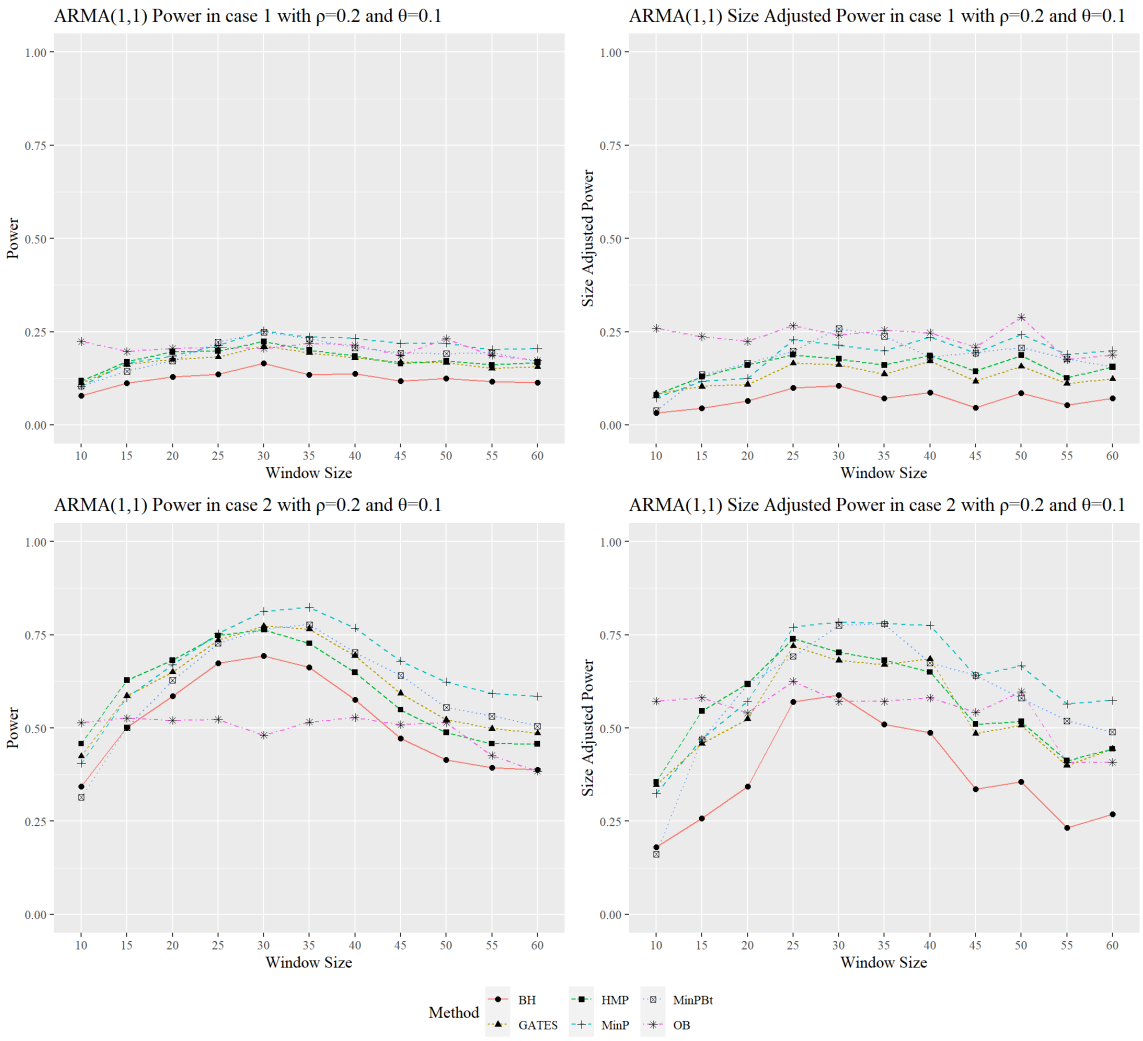

Figure C.12: This figure shows the power and the size adjusted power comparisons for the six methods when ARMA$(1,1)$ coefficient ρ is 0.2 and θ is 0.1. Referring the simulation setting, case 1 is $\mu_t = 0.5$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$; case 2 is $\mu_t = 1$ $\{\frac{t}{T} \leq \frac{1}{3}\}$ and $\mu_t = 0$ $\{\frac{t}{T} > \frac{1}{3}\}$

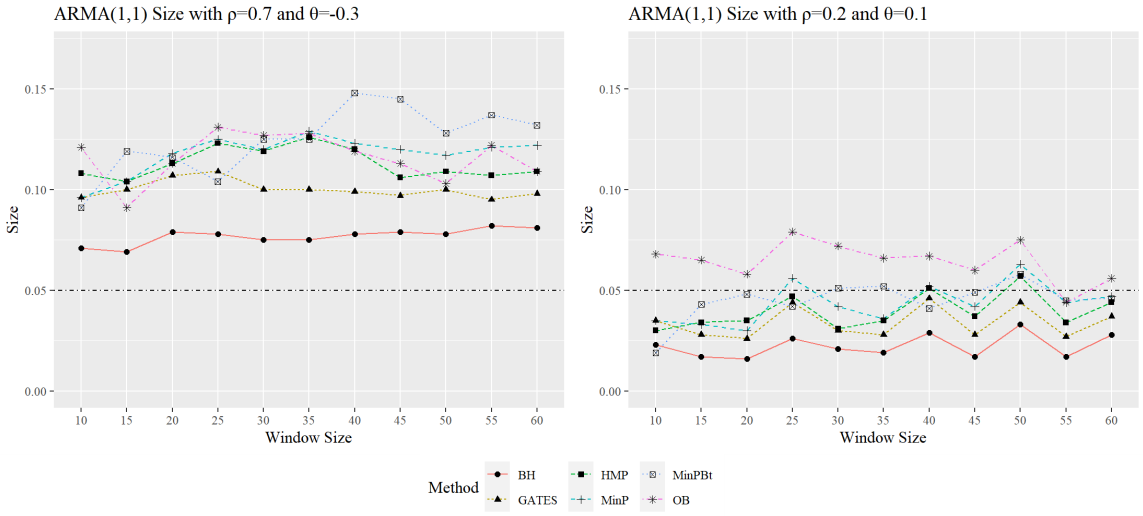

Figure C.13: This figure shows the size comparison for the six methods when ARMA(1,1) coefficients are $(0.7, -0.3)$ and $(0.2, 0.1)$.

Appendix D

Supplementary Simulation Results in 4.4

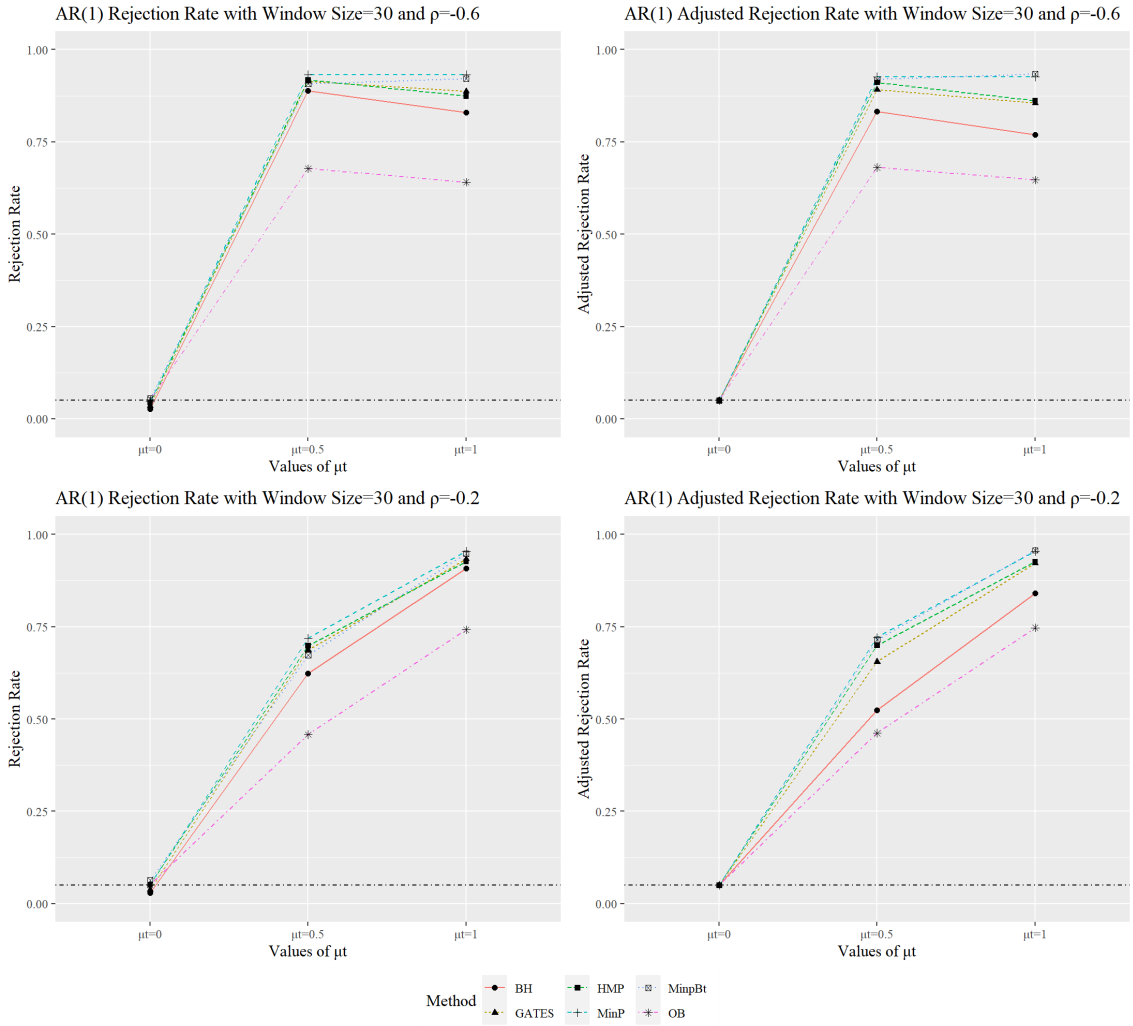

Figure D.1: This figure shows the rejection rate comparisons for the six methods with fixed window size 30 and six AR(1) coefficients: -0.6, -0.2

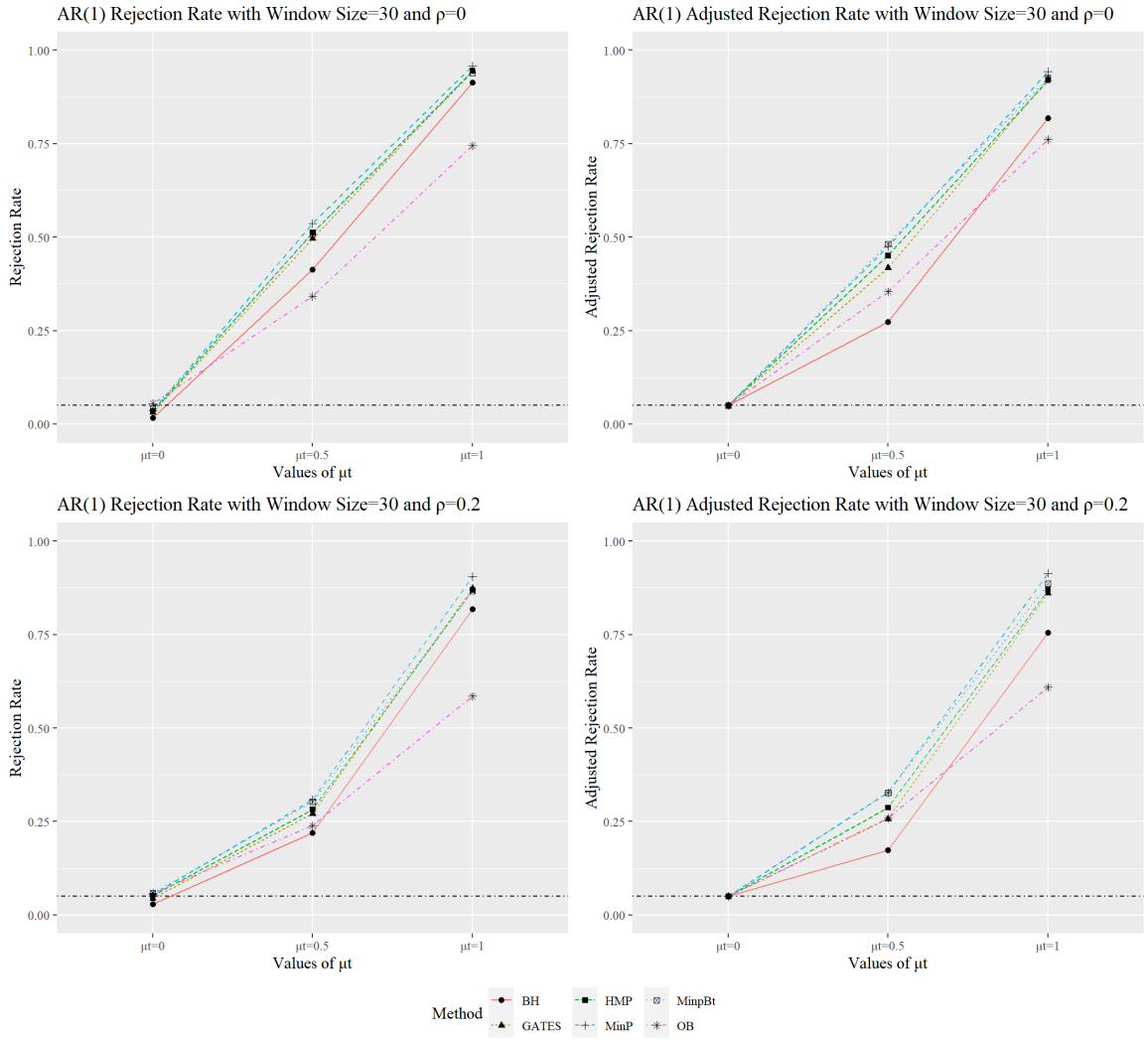

Figure D.2: This figure shows the rejection rate comparisons for the six methods with fixed window size 30 and six AR(1) coefficients: 0, 0.2

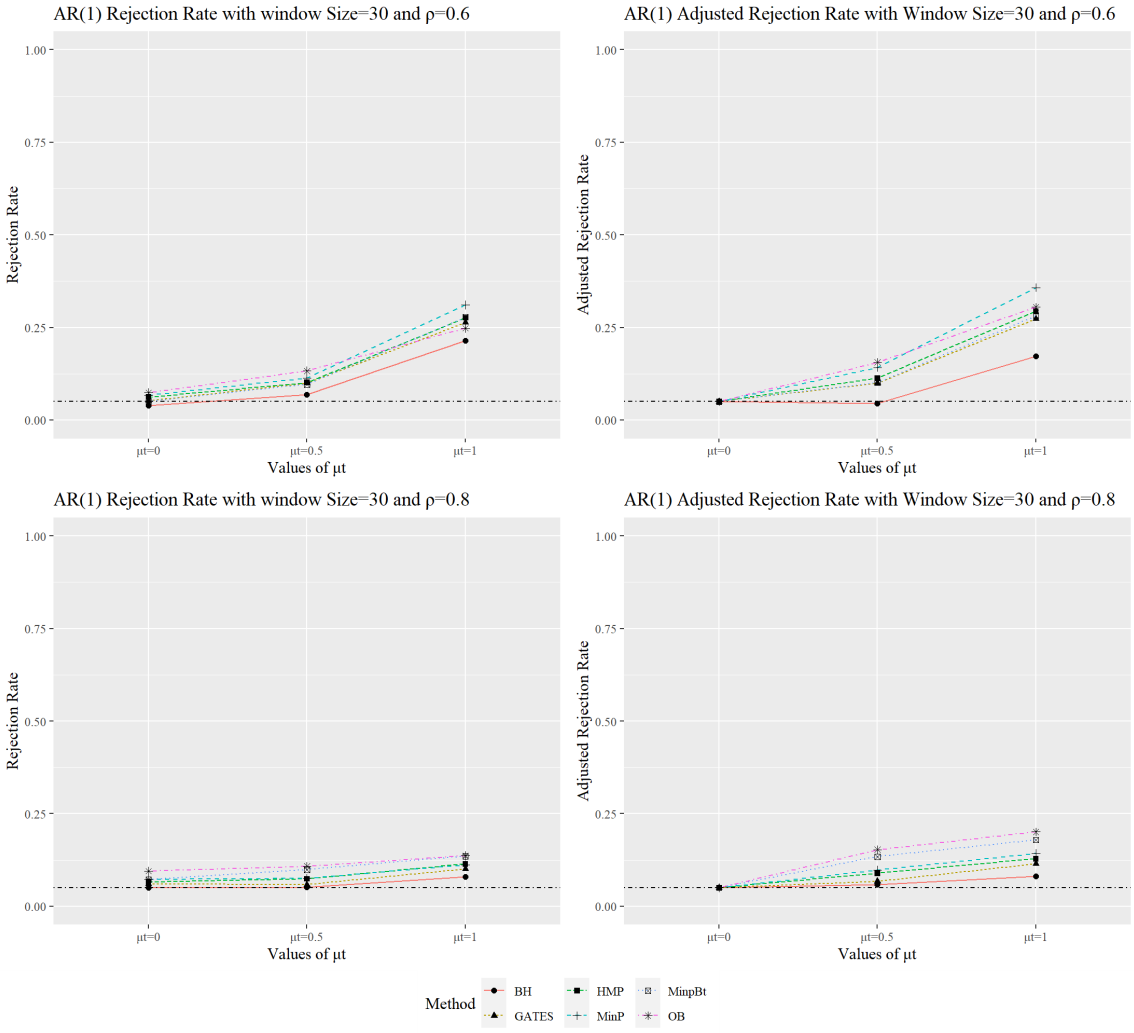

Figure D.3: This figure shows the rejection rate comparisons for the six methods with fixed window size 30 and six AR(1) coefficients: 0.6, 0.8