A Unified Cross-Validatory Approach to HAC Standard Error Estimation

by

Zhihao (Tony) Xu

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Arts

Business and Economics Honors Program

NYU Shanghai

May 2021

Professor Marti G. Subrahmanyam Professor Christina Wang Professor Wendy Jin	Professor Clifford Hurvich
Faculty Advisers	Thesis Adviser

Acknowledgements

I want to express my deepest gratitude towards my advisor Professor Clifford Hurvich, whose insightful advice and patient guidance have made my thesis possible. Our weekly meetings and his close reading of several drafts are precious. Also, I would like to thank him for showing me the beauty of time series, statistics, and, most importantly, research. Also, I would like to thank my faculty advisors, Professor Marti Subrahmanyam, Professor Christina Wang, and Professor Wendy Jin, for organizing the seminar. Thanks to my roommate Yuhao Ding for the support of debugging. Finally, I would extend my gratitude towards Zhichen Liu for her patience in improving my thesis presentation and encouraging me when I was down.

I dedicate this thesis to my grandma, Zhuang Guorui.

Abstract

We propose to use a unified frequency domain cross-validatory (FDCV) estimator for the HAC standard error estimator. Our proposed method allows for model/tuning parameter selection across parametric and nonparametric estimators at zero frequency simultaneously. Our candidate class **C** consists of REML-based autoregressive spectrum estimators and lag-weights estimators with Parzen kernel. In the Monte Carlo study, I demonstrate the reliability of our FDCV compared with the popular HAC estimators of Andrews-Monahan and Newey-West.

Contents

Chapter 1

Introduction

> Essentially, all models are wrong, but some are useful.
>
> George E. P. Box, 1987

In many regression problems involving economic and financial time series, we tend to have autocorrelated errors. Under such circumstances, the OLS coefficient estimator's consistency will be preserved, but the standard error estimator derived under the uncorrelated error framework will no longer be consistent. Hence, the confidence interval and test statistics that are based on the usual standard error estimator of the OLS coefficients can be distorted and the statistical inference is not credible. Such issues remain even when the length of our data is large. In the past 30 years, several heteroskedasticity and autocorrelation consistent (HAC) standard error estimators such as the one proposed by Newey and West (1986, 1994), Andrews (1991), and Andrews and Monahan (1992) have been introduced. They are implemented in various statistical packages. We will focus on the special case of estimating the standard error of the sample mean of a stationary univariate time series. In this case, our regression will only include an intercept

$$X_t = \mu + \varepsilon_t$$

and the OLS estimator of μ is just the sample mean of time series $\{X_t\}_{t=0}^{n-1}$

$$\hat{\mu} = \overline{X_n} = \frac{1}{n}\sum_{t=0}^{n-1} X_t$$

Under the assumption that $\{X_t\}$ is stationary and has short memory (the spectral density of $\{X_t\}$ at zero frequency $f(0)$ is finite and positive), we want to conduct valid inference for μ when the X_t are autocorrelated. Let $c_j = Cov(X_t, X_{t-j})$ be the lag-j autocovariance function, then we have

$$\begin{aligned}
Var(\hat{\mu}) &= Var(\overline{X_n}) \\
&= Var(\frac{1}{n}\sum_{t=0}^{n-1} X_t) \\
&= \frac{1}{n^2}\Big[Var(X_0) + Cov(X_0, X_1) + ... + Cov(X_0, X_{n-1}) \\
&+ Cov(X_1, X_0) + Var(X_1) + ... + Var(X_{n-1})\Big] \\
&= \frac{1}{n^2}\Big[nVar(X_t) + 2(n-1)Cov(X_t, X_{t-1}) \\
&+ 2(n-2)Cov(X_t, X_{t-2}) + ... + 2Cov(X_t, X_{t-n+1})\Big] \\
&= \frac{1}{n}\Big[c_0 + 2\sum_{j=1}^{n-1}(\frac{n-j}{n}c_j)\Big]
\end{aligned}$$

Note that one can show that asymptotically,

$$c_0 + 2\sum_{j=1}^{n-1}(\frac{n-j}{n}c_j) \xrightarrow{n\to\infty} \sum_{j=-\infty}^{\infty} c_j$$

We call $S^2 = \sum_{j=-\infty}^{\infty} c_j$ the **long-run variance**. The long-run variance is closely connected with the concept of spectral density. The spectral density f of a time series $\{X_t\}$ at frequency ω is defined as

$$f(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty} c_j \exp(i\omega j) \qquad \omega \in [-\pi, \pi]$$

In particular, when $\omega = 0$, we can obtain a relationship between the spectral

density at zero frequency $f(0)$ and the long-run variance S^2

$$nVar(\overline{X_n}) \to S^2 = 2\pi f(0)$$

Estimation of the spectral density at zero frequency has been the heart of the HAC problem of the past 30 years. It is important to note that spectral density is a frequency domain concept, we will provide a more thorough overview of concepts and techniques of frequency domain time series analysis in Chapter 2. One can estimate $f(0)$ parametrically or nonparametrically. To estimate $f(0)$ parametrically, we can estimate a time series model and construct the value of $f(0)$ that this model implies. One of the most popular parametric estimates is the autoregressive spectrum estimate. In practice, the user needs to choose the order of the autoregressive model. On the other hand, one can estimate $f(0)$ nonparametrically. A well-known class of nonparametric estimators in the literature are the lag-weights estimators. The tuning-parameter selection problem in this case is the choice of the truncation point or bandwidth. Note that the choice of the kernel function for nonparametric estimation is also a problem that users must face.

The first-generation answers to the HAC problem are based on minimizing some local criteria such as mean squared error (MSE) of $\hat{f}(0)$

$$MSE(\hat{f}(0)) = E\Big[\Big(\hat{f}(0) - f(0)\Big)^2\Big]$$

using a lag-weights estimator (see Andrews (1991)). The choice of the kernel can be motivated by the asymptotic properties of the corresponding estimate, but the choice of the truncation point is much more challenging since the optimal choice in terms of criteria like MSE depends on the actual unknown spectral density. In practice, Andrews (1991) and Andrews and Monahan (1992) propose plug-in approaches to estimate the optimal truncation point or bandwidth. Unfortunately, the first-generation HAC estimator based on nonparametric estimation has a substantial mean square error under certain data-generating mechanisms. In particular, the nonparametric estimator will have desirable performance if the spectral density function is relatively flat around zero, such as the case of white

noise but will have a substantial bias if the spectrum has a peak at zero frequency.

The second-generation answer to the HAC problem uses the idea of prewhitening to address the bias issue. Andrews and Monahan (1992) propose using a fixed-order autoregressive filter to transform the data such that the spectrum of the transformed data will be flatter in a neighborhood of zero frequency and therefore the nonparametric estimator will be less biased. The idea of prewhitening was subsequently implemented by Andrews and Monahan (1992) and Christiano and Den Haan (1996) as a part of their HAC estimators. The HAC literature measures performance by the coverage rates of the confidence intervals for the regression parameters. In the simulation study of Andrews and Monahan (1992), such fixed-order prewhitening can improve the performance of the coverage probability in many cases. In Andrews and Monahan's simulation study, the filter is an AR(1) filter based on the least-squares estimator of the autoregressive model.

What we view as the second-generation answer to the HAC problem is an attempt to combine the parametric approach and nonparametric approach. The fixed-order autoregressive filter serves as a parametric component and the nonparametric estimator allows for the non-flatness in the spectral density of the prewhintened data. However, as these approaches are currently implemented, there is no model selection of the order of the prewhitening filter. Den Haan and Levin (1997) perform a simulation study that shows the drawbacks of fixed-order prewhitening. They show that if the first-order autocorrelation of the prewhitened series is small, but higher-order autocorrelation coefficients are substantial, the confidence interval of prewhitening-based HAC methods tend to significantly overcover or undercover μ. Den Haan and Levin (1997) point out that such poor performance is due to fixed-order prewhitening. Furthermore, using least-squares to estimate the autoregressive filter may not be desirable. If our data generating process has a strong peak in the spectral density at or near zero frequency, using an AR(1) filter based on the least-squares estimator fails in prewhitening the data. In this case, the spectral density of the transformed data still has a substantial peak around zero frequency. But if the underlying data generating process is really an autoregression, then a good parameter estimator, such as the restricted likelihood estimator (REML), will lead to a good prewhitening (See Cheang and Reinsel

(2000), Harville (1974) and Chen and Deo (2012)). If we knew that actual data generating mechanism was truly an autoregression, then we should use the parametric autoregressive spectral density estimator rather than the nonparametric approach. The central issue, however, is that we do not know the actual data generating mechanism. Therefore, we propose to use a method that can allow for unified model selection across both parametric and nonparametric estimators. Our proposed unified model selection for HAC standard error estimation is based on the idea of frequency domain cross-validation (FDCV).

FDCV was originally purposed by Wahba and Wold (1975) to select the tuning parameter of spline-based spectrum estimates. Beltrao and Bloomfield (1987) propose a cross-validated log likelihood ($CVLL$) for cross-validation in the frequency domain to select the bandwidth of average periodogram estimates. Hurvich (1985) uses the cross-validation function of Wahba and Wold (1975), but instead of restricting attention to splines, he allows for an arbitrary estimator of the spectral density. Hurvich defines a frequency domain leave-out-one version of any spectrum estimate, opening up the possibility for unified selection among several types of estimators, simultaneously including nonparametric estimators and parametric estimators. All of the frequency domain cross-validation methods described above originally focused on the entire frequency range, $[0, \pi]$. The use of this global frequency band makes such methods apparently incompatible with the problem of HAC as HAC focuses on the spectrum at zero frequency.

In this thesis, we will propose a localized version of FDCV for the HAC problem, based on a class of candidates that includes both autoregressive (REML-based) estimates and nonparametric estimates, and we will examine in simulations the coverage rates of the resulting confidence interval for μ in comparison with the coverage rates corresponding to the Newey-West and Andrews-Monahan methods. Chapter 2 describes essential concepts and techniques of frequency domain time series analysis. It will also include the foundation and early works on FDCV in the spectrum estimation literature by Beltrao and Bloomfield (1987) and Hurvich (1985). Chapter 3 introduces a unified approach based upon FDCV for HAC standard error estimation. Chapter 4 reports Monte-Carlo results for several kernel-based HAC methods and FDCV (A detailed description of the kernel-based

HAC approaches will also be provided). Chapter 5 provides some concluding remarks.

Chapter 2

Informal Overview of Frequency Domain Cross-Validation

We start with an overview of some critical concepts in time series analysis, especially techniques and results in the frequency domain. After that, we will review frequency domain cross-validation (FDCV) methods in the spectrum estimation literature.

2.1 Basic Concepts and Techniques of Frequency Domain Time Series Analysis

2.1.1 Discrete Fourier Transform, periodogram

Let $\{x_t\}_{t=0}^{n-1}$ to be a real-valued data set. We define the **discrete Fourier transform (DFT)** of $\{x_t\}_{t=0}^{n-1}$ to be the sequence of complex numbers

$$J_j = \frac{1}{n}\sum_{t=0}^{n-1} x_t \exp(-i\omega_j t) \quad j = 0, ..., n-1$$

where ω_j is the j**-th Fourier frequency** defined as $\omega_j = \frac{2\pi j}{n}$
Given a sequence of complex numbers $\{z_j\}_{j=0}^{n-1}$, we define the **inverse Fourier**

transform of $\{z_j\}_{j=0}^{n-1}$ to be

$$\sum_{j=0}^{n-1} z_j \exp(i\omega_j t) \qquad t = 0, ..., n-1$$

It can be shown that $\{x_t\}$ is exactly the inverse Fourier transform of $\{J_j\}$

$$x_t = \sum_{j=0}^{n-1} J_j \exp(i\omega_j t) \qquad t = 0, ..., n-1$$

As a function of the Fourier frequency ω_j, the **periodogram** $I(\omega_j)$ at Fourier frequency ω_j is defined as

$$I(\omega_j) = \frac{n}{2\pi}|J_j|^2$$

2.1.2 The Spectrum

Now we treat our time series as a sequence of random variables. A time series $\{X_t\}$ is said to be **weakly stationary** if it satisfies:

(i) $E(X_t) = \mu$, a constant

(ii) $Cov(X_t, X_s)$ depends only on $t - s$

If we assume that our process has zero mean, then the **autocovariance** sequence $\{c_r\}$ is

$$c_r = Cov(X_t, X_{t+r}) = E(X_t X_{t+r}) = E(X_t X_{t-r}) = c_{-r}$$

An important estimate of c_r is the **sample autocovarinance**

$$\hat{c}_r = \frac{1}{n} \sum_{t=|r|}^{n-1} x_t x_{t-|r|}$$

One can show an essential relationship between sample autocovariance $\hat{c}_r$ and the periodogram $I(\omega)$:

$$I(\omega) = \frac{1}{2\pi} \sum_{|r|<n} \hat{c}_r \exp(-ir\omega)$$

and

$$\hat{c}_r = \int_{-\pi}^{\pi} I(\omega) \exp(ir\omega) d\omega$$

We define **the spectral density** $f(\omega)$ of a time series at frequency ω to be

$$f(\omega) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} c_r \exp(ir\omega) \qquad \omega \in [-\pi, \pi]$$

An important concept that will later appear in many important works in the HAC literature is the concept of the q**-th generalized derivative of a spectral density** $f(\omega)$ given by

$$f^{(q)}(\omega) = \frac{1}{2\pi} \sum_{r=-\infty}^{\infty} |r|^q c_r \exp(ir\omega) \qquad \omega \in [-\pi, \pi]$$

2.1.3 Spectrum Estimation

Estimation of the spectral density is a fundamental problem in time series analysis. There are two popular types of spectrum estimates: **classical spectrum estimates** and **autoregressive spectrum estimates**.

The classical spectrum estimate is based on the asymptotic theory of the periodogram. A widely-used approximation that is exact for Gaussian white noise is that the periodogram ordinates $I_1, .., I_{\tilde{n}}$, where $\tilde{n}$ is the largest integer less than or equal to $\frac{n-1}{2}$, are independently distributed as $f(\omega_j)\frac{1}{2}\chi^2_2$. Indeed, under regularity conditions that include short memory, it can be shown that

$$\lim_{n\to\infty} E\Big(\frac{I_j}{f(\omega_j)}\Big) = 1 \quad \text{and} \quad \lim_{n\to\infty} Var\Big(\frac{I_j}{f(\omega_j)}\Big) = 1$$

Therefore, even though periodogram ordinate I_j is an asymptotically unbiased estimate of $f(\omega_j)$, $\lim_{n\to\infty} Var(I(\omega_j))$ does not approach zero as n goes to infinity. Hence, I_j is not a consistent estimator of $f(\omega_j)$. However, one can obtain an asymptotically unbiased and consistent estimate by averaging the periodogram ordinates. This type of spectrum estimate is known as a **discrete periodogram average estimate** and has the form

$$\hat{f}(\omega_j) = \sum_{k=-m}^{m} g_k I_{j-k}$$

with $g_k \geq 0$ and $g_k = g_{-k}$ for all k, $\sum_{|k|\leq m} g_k = 1$ and $\lim_{n\to\infty} \sum_{|k|\leq m} (g_k)^2 = 0$.

Here, m is tuning constant that determines the number of periodogram ordinates to include in estimating $f(\omega_j)$. In the nonparametric spectrum estimation literature, such a tuning constant is called the **bandwidth**. From Brockwell, Davis, et al. (1991), under some weak assumptions on $\{X_t\}$, if the bandwidth m is a function of sample size n satisfying $m \to \infty$ and $\frac{m}{n} \to 0$ as $n \to \infty$, we will have

$$\lim_{n\to\infty} E(\hat{f}(\omega)) = f(\omega) \qquad \forall \omega \in [0, \pi]$$

$$Var(\hat{f}(\omega)) \sim \begin{cases} 2f^2(\omega)\sum_{|k|\leq m} g_k^2 & \text{if } \omega = 0 \text{ or } \omega = \pi. \\ f^2(\omega)\sum_{|k|\leq m} g_k^2 & \text{otherwise} \end{cases}$$

Having $\lim_{n\to\infty} \sum_{|k|\leq m}(g_k)^2 = 0$ ensures consistency of the discrete periodogram average estimate.

Another widely-used estimator in both the spectral density estimation literature and the HAC literature is the **lag-weights** (also called **Blackman-Tukey**) estimate, defined as

$$\hat{f}(\omega) = \sum_{|r|\leq h} w(\frac{r}{h})\hat{c}_r \exp(ir\omega)$$

where h is a non-negative integer, called the **truncation point**. The function $w(x)$ is even and satisfying $w(0) = 1$, $|w(x)| \leq 1$ for all x and $w(x) = 0$ for $|x| > 1$. The function $w(x)$ is called the **lag window** or **kernel**. From the relationship between sample autocovarinace $\hat{c}_r$ and periodogram $I(\omega_j)$, we can also express the lag-weights estimate as an integral average of the periodogram

$$\hat{f}(\omega) = \int_{-\pi}^{\pi} W(\omega - \lambda)I(\lambda)d\lambda$$

where $W(\lambda)$ is called the **spectral window**, and is defined as
$W(\lambda) = \frac{1}{2\pi}\sum_{|r|\leq h} w(\frac{r}{h})\exp(ir\lambda)$. Thus, the lag-weights estimate is an integral average of the periodogram.

When $n \to \infty$, for lag-weights estimate

$$Var(\hat{f}(\omega)) \sim \begin{cases} 2f^2(\omega)\frac{h}{n}\int_{-1}^{1} w^2(x)dx & \text{if } \omega = 0 \text{ or } \omega = \pi. \\ f^2(\omega)\frac{h}{n}\int_{-1}^{1} w^2(x)dx & \text{otherwise} \end{cases}$$

It is clear that $Var(\hat{f}(\omega)) \to 0$ as $n \to \infty$. Hence, we have that the lag-weights estimate is consistent. Note that the asymptotic unbiasedness also follows the result from the discrete periodogram average estimate and only requires that $h \to \infty$ and $\frac{h}{n} \to 0$ as $n \to \infty$.

There are two critical questions regarding the lag-weights estimate: **the choice of kernel** and **the choice of truncation parameter** h. We proceed to analyze properties of several popular window functions both in the spectrum estimation and in the HAC literature. Those windows are Bartlett window, Parzen window, Tukey-Haning window, and Quadratic Spectral (QS) window.

$$\text{Bartlett window:} \quad w(x) = \begin{cases} 1 - |x| & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Parzen window:} \quad w(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & \text{if } |x| \leq \frac{1}{2} \\ 2(1 - |x|)^3 & \text{if } \frac{1}{2} \leq |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Tukey-Hanning window:} \quad w(x) = \begin{cases} \frac{1+cos(\pi x)}{2} & \text{if } |x| \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{QS window:} \quad w(x) = \frac{25}{12\pi^2 x^2}\left(\frac{sin(\frac{6\pi x}{5})}{\frac{6\pi x}{5}} - cos(\frac{6\pi x}{5})\right)$$

Note that for QS window, we cannot write it as a lag-weights estimator of the form we described above. The spectral density that QS window associated with is

$$\hat{f}(\omega) = \sum_{j=-n+1}^{n+1} w(\frac{j}{h})\hat{c}_r \exp(ir\omega)$$

However, for the coherence of discussion, we would present the results of QS under the section of the lag-weights estimator.

With the formula of the asymptotic variance of the lag-weights estimate, we can obtain the following results for $\omega \neq 0$ and $\omega \neq \pi$:

Asymptotic Variance	
Bartlett	$0.667\frac{h}{n}f^2(\omega)$
Parzen	$0.539\frac{h}{n}f^2(\omega)$
Tukey-Hanning	$0.750\frac{h}{n}f^2(\omega)$
QS	$0.990\frac{h}{n}f^2(\omega)$

Under suitable regularity conditions it can be shown that for $\omega \neq 0$ and $\omega \neq \pi$,

$$\sqrt{\frac{n}{h}}\Big(\hat{f}(\omega) - f(\omega)\Big) \to N(0, f^2(\omega)\int_{-1}^{1} w^2(x)dx)$$

Another important question is what is the optimal choice of the truncation parameter h of the lag-weights estimate and the related properties of such truncation parameter choice. And it is very clear that such optimal choice of h will determine the speed of convergence of the distribution above.

But first, let us introduce an important concept called **characteristic exponent**. A characteristic exponent c is the largest positive integer such that $\lim_{x\to\infty} \frac{1-w(x)}{|x|^c}$ exists and is non-zero. For Bartlett window, the characteristic exponent is 1 and for Parzen, Tukey-Hanning and QS window, the characteristics exponent is 2. The characteristic exponent is an essential quantity for determing the asyptotically optimal bandwidth choice. We define the mean squared percentage error ($MSPE$) at frequency ω with bandwidth h as

$$MSPE(\omega; h) = E\Big(\Big(\frac{\hat{f}(\omega; h) - f(\omega)}{f(\omega)}\Big)^2\Big)$$

When the objective function is $\max_{0<\omega<\pi} MSPE(\omega)$, Priestley (1981) shows that if

$$h^* = argmin_h(\max_{0<\omega<\pi} MSPE(\omega; h))$$

then

$$\max_{0<\omega<\pi} MPSE(\hat{f}(\omega; h^*)) = O(n^{-\frac{2c}{c+1}})$$

This result suggests that if the spectral density is sufficiently smooth it is asymptotically preferable to choose the characteristic exponent c as large as possible. Since the Bartlett window has $c = 1$, it is asymptotically inferior in terms of $MSPE(\hat{f}(\omega); h^*)$ to the Parzen, Tukey-Hanning, and QS windows.

In particular, the Bartlett window has

$$h^* = O(n^{\frac{1}{3}}) \qquad \max_{0<\omega<\pi} MSPE(\hat{f}(\omega; h^*)) = O(n^{-\frac{2}{3}})$$

and for Parzen, Tukey-Hanning and QS window

$$h^* = O(n^{\frac{1}{5}}) \qquad \max_{0<\omega<\pi} MSPE(\hat{f}(\omega; h^*)) = O(n^{-\frac{4}{5}})$$

Another practical consideration is that we do not want the spectral density estimate to generate a negative result. Note that Bartlett and Parzen will always give nonnegative results, but Tukey-Hanning and QS might not. For the HAC problem, a negative estimator of $f(0)$ is completely useless for inference as it implies that the estimator of the variance of the sample mean is negative.

All the estimators we described above are nonparametric as they do not assume that our data was generated by any prescribed model. However, if we assume that our data follows exactly a particular data generating mechanism or can be well described by a particular model, we can introduce another type of spectrum estimate: **parametric spectrum estimate**. The one that is commonly used is the **autoregressive spectrum estimates**. Berk (1974) proves that under some weak conditions on $\{X_t\}$, if the order of the autoregressive model is asymptotically sufficient to overcome bias, the autoregressive spectrum estimate can yield a consistent estimator of the spectral density of X_t. The asymptotic variance of the autoregressive spectrum estimator is equivalent to that of the nonparametric smoothed periodogram estimator. Thus, autoregressive estimates can be used in a nonparametric context. We will see later in the simulation section that the parametric estimator can be useful in the nonparametric problem of HAC.

A weakly stationary process $\{X_t\}$ is said to be an **autoregressive process of order p**, denoted as **AR(p)**, if there exist constants $\phi_1, ..., \phi_p$ such that

$$X_t = \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \varepsilon_t \qquad \text{for all} \quad t$$

$$\varepsilon_t \overset{i.i.d}{\sim} N(0, \sigma^2)$$

$$\varepsilon_t \quad \text{independent of} \quad X_{t-s} \quad \text{for all} \quad s > 0$$

It can be shown that if the process follows an AR(p), then the spectral density is of the form

$$f(\omega) = \frac{\sigma^2}{2\pi|1 - \sum_{k=1}^{p} \phi_k \exp(i\omega k)|^2}$$

It then follows that if we can estimate the model parameters, $\phi_1, ..., \phi_p, \sigma^2$ with $\hat{\phi}_1, ..., \hat{\phi}_p, \hat{\sigma}^2$, we can estimate the spectral density $f(\omega)$ with

$$\hat{f}(\omega) = \frac{\hat{\sigma}^2}{2\pi|1 - \sum_{k=1}^{p} \hat{\phi}_k \exp(i\omega k)|^2}$$

There are two critical problems regarding autoregressive spectrum estimation. First, how many lagged observations to include in the model. In other words, what value of p should we use? Note that if all candidate models are autoregressive models, we could use classical model selection criteria such as Akaike's information criterion (AIC) (See Akaike (1974)), Bayesian information criterion (BIC) (See Schwarz et al. (1978)), etc. The second problem is, given p, how to estimate the model coefficients. There are many existing methods, including Yule-Walker, Burg, least-squares, maximum likelihood, and restricted maximum likelihood. We will discuss those methods more in Chapter 3 and the Appendix.

2.2 Frequency Domain Cross-Validation

As we have discussed above, for a nonparametric estimate (discrete periodogram average estimate or lag-weights estimate), the user must select a bandwidth or truncation parameter, and for a parametric estimate, one needs to determine the order of the model. To determine which bandwidth or order is optimal, one might attempt to minimize some criterion that measures the discrepancy between the actual spectral density function and the spectrum estimate. However, in terms of such criteria, the optimal bandwidth of a nonparametric estimate (discrete periodogram average estimate or lag-weights estimate) or the optimal order for a parametric estimate depends on the actual spectral density function, which is unknown in practice. Frequency domain cross-validation methods can be used to give a data-driven selection of a spectrum estimate without restricting the form, eg., parametric or nonparametric. Wahba and Wold (1975) were the first to use

FDCV for selection of tuning constants in a spectral estimator. They focused on spline-based estimators. The work done by Beltrao and Bloomfield (1987) focuses on nonparametric spectral density estimation. They show that minimizing mean integrated square error (MISE) is asymptotically equivalent to minimizing a cross-validatory log-likelihood (CVLL). So an optimal bandwidth can be chosen by minimizing CVLL. All candidate estimators in Beltrao and Bloomfield are nonparametric. Hurvich (1985) proposes a unified FDCV method that can select tuning constants for any types of estimators as long as such the estimator can be computed using the actual data $\{x_t\}$. For example, the method of Hurvich allows us to select between a parametric estimate and a nonparametric estimate, which we will show later has a considerable advantage in the HAC problem. We will introduce the philosophy and an overview of Beltrao and Bloomfield (1987) and focus on Hurvich (1985) as his unified FDCV method is the foundation of our proposed HAC method .

First, let us introduce several measures of the discrepancy between the actual spectral density and its estimate.

The **mean squared logarithmic error (MSLE)** of a spectral density estimate $\hat{f}$ at a fixed frequency ω is defined as

$$MSLE(\hat{f}(\omega)) = E\Big[\Big(\log \hat{f}(\omega) - \log f(\omega)\Big)^2\Big]$$

The **mean integrated squared percentage error (MISPE)** of a spectral density estimate $\hat{f}$ over a frequency band $[0, \pi]$ is defined as

$$MISPE(\hat{f}) = E\Big[\int_0^\pi \Big(\frac{\hat{f}(\omega) - f(\omega)}{f(\omega)}\Big)^2 d\omega\Big]$$

The **mean integrated squared logarithmic error (MISLE)** of a spectral density estimate $\hat{f}$ over a frequency band $[0, \pi]$ is defined as

$$MISLE(\hat{f}) = E\Big[\int_0^\pi \Big(\log \hat{f}(\omega) - \log f(\omega)\Big)^2 d\omega\Big]$$

Beltrao and Bloomfield (1987) and Hurvich (1985) use global measures, on the entire frequency band $[0, \pi]$. Note, however, that the HAC problem is inherently

local to zero frequency. This fact will motivate our modification in Chapter 3 of the FDCV method of Hurvich.

Beltrao and Bloomfield (1987) focus on mean integrated squared percentage error on $[0, \pi]$. However, instead of integrating over Fourier frequencies, Beltrao and Bloomfield obtain a discrete version of the discrepancy by summing over the Fourier frequencies between 0 and π but ignoring the 0-th and $\frac{n}{2}$-th Fourier frequency. It is of the form

$$E\Big[\frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}}\Big(\frac{\hat{f}(\omega_j) - f(\omega_j)}{f(\omega_j)}\Big)^2\Big]$$

where $\tilde{n}$ is the largest integer less than or equal to $\frac{n-1}{2}$. Beltrao and Bloomfield consider the bandwidth selection problem for discrete periodogram average estimates. If we denote discrete periodogram average estimate with bandwidth m at frequency ω_j as $\hat{f}(\omega_j; m)$, and define the specific criterion Bloomfield and Beltrao are using as $MISPE_{BB}$, we have that

$$MISPE_{BB}(m) = E\Big[\Big(\frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}}\frac{\hat{f}(\omega_j; m) - f(\omega_j)}{f(\omega_j)}\Big)^2\Big]$$

$MISPE_{BB}(m)$ is not a feasaible criterion as it depends on the unknown spectral density f. Whittle's approximation for minus twice the Gaussian log-likelihood for spectral density f is

$$A = \sum_{j=1}^{\tilde{n}}\Big[\log(f(\omega_j)) + \frac{I(\omega_j)}{f(\omega_j)}\Big]^2$$

Taking the first derivative of log-likelihood, it is easy to see that A is minimized at $f(\omega_j) = I(\omega_j)$. For discrete periodogram average estimate, this can be achieved by taking sufficiently small m. Hence, such direct minimization of A will not help. The approach of leave-one-out cross-validation can be a solution to this issue. Bloomfield and Beltrao construct a leave-one-out cross-validatory version of A, given by

$$CVLL(m) = \sum_{j=1}^{\tilde{n}}\Big[\log(\hat{f}^{-j}(\omega_j; m)) + \frac{I(\omega_j)}{\hat{f}^{-j}(\omega_j; m)}\Big]^2$$

where $\hat{f}^{-j}(\omega_j; m)$ omits $I(j)$ from $\hat{f}(\omega_j; m)$ when $\hat{f}$ is a periodogram average estimate.

They show that as n goes to ∞,

$$\frac{1}{\tilde{n}} CVLL(m) = \frac{1}{\tilde{n}} \sum_{j=1}^{\tilde{n}} \left[\log f(\omega_j) + \frac{I(\omega_j)}{f(\omega_j)} \right] + \frac{1}{2} MISPE_{BB}(m) + o_p(MISPE_{BB}(m))$$

where the term $o_p(MISPE_{BB}(m))$ is uniform in m.

Therefore, minimizing the $CVLL$ on the left hand side with respect to m is approximately equivalent to minimizing $MISPE_{BB}$ on the right hand side for n large as $\frac{1}{\tilde{n}} \sum_{j=1}^{\tilde{n}} \left[\log f(\omega_j) + \frac{I(\omega_j)}{f(\omega_j)} \right]$ does not depend on m.

Beltrao and Bloomfield restrict the candidates to nonparametric spectral estimates. Hurvich (1985) extends the FDCV approach's applicability by introducing a generalized leave-one-out version of the spectral density estimate. Hurvich's purpose is to develop a method that allows researchers to do tuning parameter/model selection across parametric and nonparametric estimates simultaneously. There are two methods introduced in his paper: an autocovariance-based approach and a DFT-based approach. With the autocovariance-based approach, the researcher will be able to do model selection from any spectral density estimate that be expressed as a function of the sample autocovariances. Those estimates include the lag-weights estimate, discrete periodogram average estimate, and autoregressive estimate using the Yule-Walker method. For example, the autocovariance-based approach allows researchers to choose between a Yule-Walker autoregressive estimate and a periodogram average estimate based on some objective criterion. This approach extended the applicability of the FDCV method developed by Bloomfield and Beltrao, whose approach only allows for model selection within periodogram average estimates. The DFT-based approach of Hurvich is more generally applicable than his autocovariance-based approach. This approach allows the candidates to include any spectral density estimates based on the actual data. Note that not all the spectral density estimates can be expressed using sample autocovariance, such as autoregressive estimate with least-squares or maximum likelihood estimate or restricted maximum likelihood estimate. The DFT-based approach is the one we will use in our FDCV method for HAC standard error estimation due to its

generality. Let us review this approach step by step.

Recall that $MISPE_{BB}$ is the one used by Beltrao and Bloomfield as their discrepancy, Hurvich also considers the discrete version of $MISPE$ as a discrepancy, but he also considers the discrete version of $MISLE$ as his discrepancy:

$$MISPE_H = E\Big[\frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}}\Big(\frac{\hat{f}(\omega_j) - f(\omega_j)}{f(\omega_j)}\Big)^2\Big]$$

$$MISLE_H = E\Big[\frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}}\Big(\log\hat{f}(\omega_j) - \log f(\omega_j)\Big)^2\Big]$$

Note that the definitions of $MISPE_H$ and $MISPE_{MM}$ are different as we no longer require f to be a nonparametric estimate and bandwidth input is dropped in the formula. For $MISPE_H$ and $MISLE_H$, the cross-validatory estimates are

$$CVLL_H = \frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}}\Big[\log\hat{f}^{-j}(\omega_j) + \frac{I(\omega_j)}{\hat{f}^{-j}(\omega_j)}\Big]^2$$

$$CVMSE_H = \frac{1}{\tilde{n}}\sum_{j=1}^{\tilde{n}}\Big\{\Big[\log\hat{f}^{-j}(\omega_j) - \Big(\log I(\omega_j) + C\Big)\Big]^2 - \frac{\pi^2}{6}\Big\}$$

where $C = 0.577216...$ is the Euler constant and $\hat{f}^{-j}$ is the generalized leave-one-out version of the spectral density estimate. Hurvich suggests the use of $CVMSE_H$ as the selection criterion, and the selected spectral estimator is the one with the lowest $CVMSE_H$ value. Now we proceed to describe how the generalized leave-one-out version of the spectral density estimate is obtained.

First, assume that a zero-mean process generates our data and we know that the process has zero mean. Hurvich (1985) defines the leave-one-out version of DFT, J_k^{-j} for $0 \le k \le n-1$ and $1 \le j \le n$ as

$$J_k^{-j} = \begin{cases} J_k & \text{if } k \neq j \text{ and } k \neq n-j \\ \frac{1}{2}(J_{k-1} + J_{k+1}) & \text{if } k = j \text{ and } k = n-j \end{cases}$$

Then we can use the inverse Fourier transform to define the leave-one-out version of data, $\{x_t^{-j}\}$ by

$$x_t^{-j} = \sum_{k=0}^{n-1} J_k^{-j}\exp(i\omega_k t)$$

Based on this leave-one-out version of the data, Hurvich defined the generalized leave-one-out spectral density estimate as

$$\hat{f}^{-j}(\omega_j) = \hat{f}(\omega_j; \{x_t^{-j}\})$$

Unfortunately, the assumption that the mean is known to be zero is a very strong assumption which typically does not hold in practice. Indeed, if this assumption held, then HAC standard errors for $\overline{X_n}$ would not be needed since μ would be known. We will henceforth restrict attention to spectral estimates and cross validation functions that are invariant to the addition of a constant to the data set. If we drop the zero mean assumption, the DFT-based leave-one-out version of the data is not invariant under adding a constant because

$$J_1^{-1} = \frac{1}{2}(J_0 + J_2)$$

and

$$J_0 = \frac{1}{n}\sum_{t=0}^{n-1} x_t = \overline{X_n}$$

Hence, J_1^{-1} is not invariant under adding a constant. To handle this issue, we now redefine the leave-out-one version of the DFT for $j = 0, ..., n-1$ and $k = 0, ..., n-1$ as

$$J_k^{-j} = \begin{cases} J_k & \text{if } k \neq j \text{ and } k \neq n-j \\ \frac{1}{2}(J_{k-1} + J_{k+1}) & \text{if } k = j \text{ or } k = n-j \ (j \neq 1) \\ J_2 & \text{if } k = j = 1 \\ J_{n-2} & \text{if } k = n-1 \text{ and } j = 1 \end{cases}$$

We also need to adjust the computation of the leave out one version of data x_t^{-j} by removing the zero frequency and we therefore redefine

$$x_t^{-j} = \sum_{k=1}^{n-1} J_k^{-j} \exp(i\omega_k t)$$

Under this definition, the DFT-based FDCV is invariant under adding a constant.

Chapter 3

A Unified Cross-Validatory Approach to HAC Standard Error Estimation

3.1 A Unified Cross-Validatory Approach to the Estimation of Spectral Density at Zero Frequency

In Chapter 2, we pointed out that Hurvich (1985) provides the generalized leave-out-one definition of spectrum estimate $\hat{f}^{-j}(\omega_j) = \hat{f}(\omega_j; \{x_t^{-j}\})$, which can be applied for any spectrum estimate $\hat{f}$. This opens up the possibility for model selection to choose models over a larger class of candidates. This section introduces a cross-validatory approach for HAC standard error estimation by providing a unified truncation parameter selection procedure for the spectrum estimate at zero frequency.

Consider the model candidate class **C** of spectrum estimates. **C** is composed of REML-based autoregressive spectrum estimates from order zero to five and lag-weights estimates based on a Parzen kernel with truncation point 1 to $m(n)$, where $m(n)$ increases in n and takes an integer value. All members of this candidate class **C** have a well-defined generalized leave-out-one version of the spectrum estimate.

We select the estimate from $\mathbf{C}$ that minimizes the HAC version of $CVMSE$

$$CVMSE_{HAC}(\hat{f}) = \frac{1}{\lfloor \tilde{n}^c \rfloor} \sum_{j=1}^{\lfloor \tilde{n}^c \rfloor} \left\{ \left[log\hat{f}^{-j}(\omega_j) - \Big(logI(\omega_j) + C \Big) \right]^2 - \frac{\pi^2}{6} \right\}$$

where $c \in (0,1)$ is a constant. In practice, we suggest to take $c = \frac{4}{5}$. The chosen spectrum estimate will be our unified cross-validatory estimate, and we will use it to estimate the spectral density of our time series at zero frequency.

3.2 Discussion

Our purposed FDCV method for the HAC problem and Hurvich (1985) differ in the cross-validation function and the candidate class $\mathbf{C}$. They also differ in the definition of the leave-one-out version of the data set, as shown in Chapter 2.

3.2.1 Cross-Validation Function

The cross-validation function $CVMSE$ of Hurvich (1985) is defined over a frequency band between the 1-st to the $\tilde{n}$-th Fourier frequency and $CVMSE_{HAC}$ is defined over a frequency band between the 1-st to the $(\lfloor \tilde{n}^c \rfloor)$-th Fourier frequency for any $c \in (0,1)$.

Why is such a change essential? Andrews (1991) mentioned the potential application of the FDCV method proposed by Beltrao and Bloomfield (1987) to the HAC problem. However, he considered the method to be not well-suited to HAC standard error estimation as the cross-validation criterion $CVMSE$ is a global measure over the frequency band $[0, \pi]$, while the HAC problem focuses on estimating the density at a single frequency, zero. This modification makes the criterion function asymptotically local to zero frequency. Indeed, the largest Fourier frequency in $CVMSE_{HAC}$ approaches zero for $c \in (0,1)$,

$$\lim_{n\to\infty} \frac{2\pi(\frac{n}{2})^c}{n} = \lim_{n\to\infty} \pi \left(\frac{n}{2}\right)^{c-1} = 0$$

Philosophically, our application of FDCV to the HAC problem is motivated by John Tukey's idea of "borrowing strength." We borrow strength from the neigh-

boring frequencies around zero to obtain a stable estimate of $f(0)$.

From simulations, we suggest taking $c = \frac{4}{5}$ to handle the bias variance trade-off inherent in the frequency range used in the cross-validation function.

3.2.2 Candidate Class

The model candidates class of Hurvich (1985) includes Daniell average periodogram estimates and Yule-Walker autoregressive estimates. Our candidate class **C** consists of lag-weights estimates with the Parzen kernel and REML autoregressive estimates.

For the nonparametric model candidates, our shift from the Daniell periodogram average estimates to lag-weights estimates with the Parzen kernel is motivated by findings of Newey and West (1986, 1994), Andrews (1991) and Andrews and Monahan (1992). All those works are based on lag-weights estimators. Newey and West (1994) conclude, based on a simulation, that the effect of the choice of kernel between Bartlett, Parzen, and QS (quadratic spectral) on the performance of the estimator is negligible. We use the Parzen kernel as it never generates negative estimates as opposed to QS, and it has lower asymptotic variance theoretically compared with Bartlett if the infeasible optimal truncation point were used.

For parametric model candidates, we propose to estimate autoregressive models using REML instead of Yule-Walker, MLE or least-squares (least-squares is used to estimate the AR(1) filter in Andrews and Monahan (1992) and Newey and West (1994)). Our use of REML to estimate the autoregressive model is based on two considerations. First is the reduction of bias of REML compared with other autoregressive estimators. The bias of the Yule-Walker estimator is more prominent than other popular autoregressive estimators in a small sample. Yule-Walker performs poorly for autoregressive models having a root that is close to the unit circle (or more generally when the spectral density function has strong peaks or troughs). For least-squares and MLE, Cheang and Reinsel (2000) showed that given an AR(1) process, the bias of estimation of the autoregressive coefficient will be as much as doubled when the root is close to the unit circle. They also showed that the REML estimates of the autoregressive parameters, which do not

require knowledge of the mean, perform equivalently up to a term of $O(\frac{1}{n})$ compared to MLE or least-squares with the infeasible knowledge of the mean. Notice that one potential problem of the HAC method is that if the data generating mechanism has a sharp peak in the spectral density at zero frequency, using an AR(1) filter based on the least-squares estimator might fail to completely prewhiten the data. It is clear that if the autoregressive estimators are based on REML, it will lead to a better prewhitening. This motivates our use of REML to estimate the autoregressive model.

On the other hand, even though our cross-validation procedure is invariant under adding a constant (whether to use mean-corrected data will not influence the chosen model candidate), we need to apply our chosen spectral estimator to mean-corrected data. This is because the lag-weights estimator is not invariant under adding a constant. For parametric estimators, Yule-Walker, least squares, or MLE estimators are not invariant under adding a constant, so we need to subtract the sample mean when estimating the spectrum using our chosen candidate estimates. However, for Yule-Walker, Least-squares, and MLE, demeaning the data will compromise their performance. One advantage of REML is that it is naturally invariant under adding a constant, which is a desirable property for spectrum estimation of a process with an unknown mean.

Note that REML estimation of the autoregressive model in the time series context has not been heretofore implemented in any software packages as far as we are aware. Many researchers suggest using it in the context of mixed linear effects model, which can be specialized to yield autoregressive model. The restricted likelihood function is not well-defined for non-stationary models. A REML autoregressive estimator constrained for stationarity is not implemented in any widely-available packages, as far as we are aware. We will discuss the philosophy and computation of AR(p) REML estimates under stationarity constraints in Appendix A.

Chapter 4

Monte-Carlo Study

This chapter will present the results of a Monte-Carlo study comparing our FDCV and popular kernel-based HAC methods for small sample sizes. We will evaluate the performance of the methods based on the coverage rate of the confidence intervals. The first estimator is the cross-validatory estimator described in Chapter 3. The second estimator is the kernel-based HAC method given by Andrews and Monahan (1992) using the quadratic spectral window, and AR(1) specification in their tuning-parameter selection procedure, and the least-squares based AR(1) prewhitening filter. We call it AM-PW, short for Andrews and Monahan's prewhitening based HAC estimator. The third estimator is the kernel-based HAC estimator given by Newey and West (1994) using Bartlett window as described in Newey and West (1986, 1994), a nonparametric truncation parameter model selection procedure, and the least-squares-based AR(1) prewhitening filter. We call it NW-PW, short for Newey and West prewhitening based estimator. In Section 4.1, we will provide a step-by-step review of AM-PW and NW-PW HAC estimators. Section 4.2 investigates several experiments suggested by Andrews and Monahan (1992) and Den Haan and Levin (1997). We consider two sample sizes, $n = 50, 200$ and report the coverage rates of the HAC-based confidence intervals at nominal rates 90%, 95% and 99% based on 3000 replications. The computations for the AM-PW and NW-PW methods are based on the Sandwich package in R.

4.1 Step-by-step Review of Kernel-based HAC Estimator

The AM-PW and NW-PW methods apply a prewhitening AR filter of order m as an input to the kernel-based automatic truncation parameter selection procedure from Andrews (1991) or Newey and West (1994). In the simulation study of Andrews and Monahan (1992) and Newey and West (1994), they set $m = 1$ and we will use $m = 1$ as well in our simulations. The two methods differ in their truncation parameter selection procedure.

Given data $\{x_t\}$, the two methods select the optimal truncation parameter as follows.

Step 0: Define $\tilde{x}_t = x_t - \sum_{t=0}^{n-1} x_t$.

Step 1: Prewhiten the mean-corrected data

Obtain the prewhitened data $\{\hat{e}_t\}$ using least-squares

$$\tilde{x}_t = \sum_{k=1}^{m} \hat{A}_k \tilde{x}_{t-k} + \hat{e}_t \quad t = m+1, ..., n$$

The case $m = 1$ yields the AR(1) filter used in Andrews and Monahan (1992) and Newey and West (1994) simulation (and the default value for the Sandwich package in R).

$$\tilde{x}_t = \hat{A}\tilde{x}_{t-1} + \hat{e}_t \quad t = 2, ..., n$$

Step 2: Specify a kernel for estimation of the spectral density of $\{\hat{e}_t\}$

Andrews (1991) derives the truncation parameter growth rate that minimizes the asymptotic MSE of the spectral density estimator. The optimal truncation parameter sequence for a given kernel depends on the kernel's smoothness properties, as indicated by the characteristic exponent, q. For the Bartlett kernel, $q = 1$; and for the QS and Parzen kernel, $q = 2$. For a given kernel with characteristic exponent q, the asymptotically optimal bandwidth parameter sequence is given by

$$r^* = c\Big[\Big(\frac{f^{(q)}(0)}{f(0)}\Big)^2 n\Big]^{\frac{1}{2q+1}}$$

where c is a constant that depends on the window function. In particular, we have

$$c = \begin{cases} 1.1447 & \text{Bartlett window} \\ 2.6614 & \text{Parzen window} \\ 1.3221 & \text{QS window} \end{cases}$$

Step 3: Calculate the estimated optimal truncation parameter for the specified kernel in Step 2

To obtain an estimator of the optimal truncation parameter $\hat{r}^*$, we will need to obtain an initial ("pilot") estimator of $f^{(q)}(0)$ and $f(0)$. This is where Andrews (1991) and Newey and West (1994) differ from each other. The initial estimator of $f^{(q)}(0)$ and $f(0)$ is obtained parametrically by Andrews, while it is done non-parametrically by Newey and West.

Andrews suggested to estimate $f(0)$ and $f^{(q)}(0)$ by least-squares fitting of an AR(1) model to $\{\hat{e}_t\}$

Newey and West proposed the initial estimation nonparametrically. The initial estimator of $f(0)$ and $f^{(q)}(0)$ are based on the lag-weights estimator using Bartlett, Parzen or QS. Note that Newey and West claim that different kernels will not significantly influence the construction of the confidence interval based on the simulation result. They recommend to use Bartlett kernel.

Step 4: Calculate the spectral density of the prewhitened data $\{\hat{e}_t\}$ using QS or Bartlett kernel with its associated estimated optimal truncation parameter in step 3. We denote it as $\hat{f}_e(0)$

Step 5: Calculate the HAC estimate of the spectral density of initial observations by

$$\hat{f}(0) = \frac{\hat{f}_e(0)}{|1 - \sum_{k=1}^{m} \hat{A}_k|^2}$$

4.2 Monte-Carlo Results

We conducted five sets of experiments for sample sizes $n = 50, 200$, each with 3000 replications. In the FDCV method for HAC, the class $\mathbf{C}$ of candidate spectrum estimates consisted of the REML-based autoregressive estimates of order 0 to 5 and lag-weights estimates with Parzen kernels from truncation point 1 to

$\lfloor 4(\frac{n}{100})^{\frac{2}{9}} \rfloor$. For each realization, and for each of the candidates $\hat{f}$ in $\mathbf{C}$, I computed its respective $CVMSE_{HAC}(\hat{f})$ with $c = \frac{4}{5}$.

For the REML-based autoregressive estimates, I computed the following model order selection criteria:

$$WHREML(\hat{f}) = \frac{1}{\lfloor \tilde{n}^{\frac{4}{5}} \rfloor} \sum_{j=1}^{\lfloor \tilde{n}^{\frac{4}{5}} \rfloor} \left\{ \left[log\hat{f}^{-j}(\omega_j) - \Big(logI(\omega_j) + C \Big) \right]^2 - \frac{\pi^2}{6} \right\}$$

In this formula, $\hat{f}^{-j}$ is the leave-one-out version of the REML-based autoregressive estimate. The criterion is the same as $CVMSE_{HAC}(\hat{f})$ when $\hat{f}$ is a REML-based autoregressive spectrum estimate. I have renamed this quantity to signal that we are only considering autoregressive spectrum estimates in this case. Here, WH stands for Wahba-Hurvich, $REML$ stands for REML-based autoregressive estimates.

Similarly, for the lag-weights estimate with the Parzen window, I computed the following bandwidth selection criteria:

$$WHPZ(\hat{f}) = \frac{1}{\lfloor \tilde{n}^{\frac{4}{5}} \rfloor} \sum_{j=1}^{\lfloor \tilde{n}^{\frac{4}{5}} \rfloor} \left\{ \left[log\hat{f}^{-j}(\omega_j) - \Big(logI(\omega_j) + C \Big) \right]^2 - \frac{\pi^2}{6} \right\}$$

In this formula, $\hat{f}^{-j}$ is the leave-one-out spectrum estimate for the lag-weights estimates with Parzen kernel. Note that $WHPZ$ is equivalent to $CVMSE_{HAC}$ when $\hat{f}$ is lag-weights estimates with Parzen kernel. Here, PZ stands for Parzen.

I now define a combined criterion function by

$$WHC(\hat{f}) = \begin{cases} WHREML(\hat{f}) & \text{if } \hat{f} \text{ is an autoregressive estimate} \\ WHPZ(\hat{f}) & \text{if } \hat{f} \text{ is a lag-weights estimate} \end{cases}$$

We use the same name for the criterion function as in Hurvich (1985), but the function is now designed for the HAC problem since the criterion is asymptotically local to zero frequency. The estimator selected by WHC is the one that minimizes WHC over the candidate class $\mathbf{C}$.

For $WHREML$, $WHPZ$ and WHC, after obtaining the selected estimator, we will demean data set $\{x_t\}$ and apply the selected estimator on the mean-

corrected time series for estimation at zero frequency. We denote the resulting spectrum estimators as $\hat{f}^{WHREML}(0)$, $\hat{f}^{WHPZ}(0)$ and $\hat{f}^{WHC}(0)$. Their associated estimated standard errors for $\hat{\mu} = \overline{X}_n$ are

$$\hat{\sigma}^{WHREML} = \sqrt{\frac{2\pi \hat{f}^{WHREML}(0)}{n-1}}$$

$$\hat{\sigma}^{WHPZ} = \sqrt{\frac{2\pi \hat{f}^{WHPZ}(0)}{n-1}}$$

$$\hat{\sigma}^{WHC} = \sqrt{\frac{2\pi \hat{f}^{WHC}(0)}{n-1}}$$

where $n-1$ in the denominator is a finite-sample correction (See Andrews (1991)). We will use the standard error to compute the nominal 90%, 95% and 99% confidence intervals for μ and report the observed coverage rates for $WHREML$, $WHPZ$ and WHC. Notice that for AM-PW and NW-PW, we will directly use the results from the Sandwich package, which provides the standard error directly.

After the observed coverage rates for each method are obtained, we provide the relative efficiency measures for WHC, AM-PW, and NW-PW for each data-generating mechanism under each sample size at the nominal 95% coverage rate. The relative efficiency is a number in $[0, 1]$. In each case, the method with the relative efficiency of 1 is the one with the best performance, and the method with the lowest relative efficiency is the one with the worst performance. To compute the relative efficiency, we will need to construct a measure for badness. We define the badness B of the actual coverage rate p for the nominal coverage rate 95% as

$$B(p) = \begin{cases} 2|logit(p) - logit(0.95)| & \text{if} \quad p \leq 0.95 \\ |logit(p) - logit(0.95)| & \text{if} \quad p > 0.95 \end{cases}$$

where $logit(p) = \log(\frac{p}{1-p})$.

In practice, under-coverage is considered to be worse than over-coverage. To take this asymmetry into account, we use a factor of 2 to penalize under-coverage in $B(p)$, when $p \leq 0.95$.

Let us denote p_1, p_2 and p_3 to be the actual coverage probabilities of WHC,

AM-PW, NW-PW respectively, then we can obtain the relative efficiency of p_i.

$$e(p_i) = \frac{min\{B(p_1), B(p_2), B(p_3)\}}{B(p_i)} \qquad i \in \{1, 2, 3\}$$

Note that $e(p_i) \in [0, 1]$. If $p_i = 1$, then $logit(p_i) = \infty$. In this case, we will assign value 0 to $e(p_i)$

We ran five sets of experiments. The data-generating mechanisms for subsection 4.2.1 are AR(1) processes. For subsection 4.2.2, the data-generating mechanism is a white noise process. For subsection 4.2.3, they are MA(1) processes. The data-generating mechanisms in Subsections 4.2.4 and 4.2.5 were proposed by Den Haan and Levin (1997). Notice that we are primarily comparing results between WHC, AM-PW, and NW-PW. The criteria $WHREML$ and $WHPZ$ are used for supplementary analysis. The **minimum** row in the table of the relative efficiency of each subsection takes the minimum value of relative efficiency statistics of all cases in that experiment.

4.2.1 AR(1) Processes

$$X_t = \phi_1 X_{t-1} + \varepsilon_t \qquad \varepsilon_t \overset{i.i.d}{\sim} N(0,1)$$

Table 4.1: Coverage Probabilities for AR(1) Processes: $n = 50$

ϕ_1	**Method**	**90%**	**95%**	**99%**	ϕ_1	**Method**	**90%**	**95%**	**99%**
	WHC	86.7	91.9	97.4		WHC	81.5	88.7	94.9
	WHREML	86.7	91.9	97.4		WHREML	82.1	88.6	94.6
0.1	WHPZ	87.2	92.5	97.8	0.3	WHPZ	80.6	87.9	95.2
	AM-PW	87.7	93.1	97.7		AM-PW	86.5	92.0	97.2
	NW-PW	85.3	90.8	96.4		NW-PW	85.2	90.6	96.3
	WHC	79.1	85.3	92.8		WHC	75.1	81.8	89.9
	WHREML	80.3	86.2	92.8		WHREML	79.2	84.5	90.9
0.5	WHPZ	76.3	82.9	91.6	0.7	WHPZ	68.7	76.7	87.3
	AM-PW	85.4	90.6	96.3		AM-PW	82.7	88.1	94.3
	NW-PW	84.4	89.9	95.7		NW-PW	81.8	87.5	94.0
	WHC	70.8	77.2	84.7		WHC	68.6	74.4	82.4
	WHREML	75.7	81.5	88.0		WHREML	73.2	78.8	85.7
0.9	WHPZ	46.6	53.6	66.7	0.95	WHPZ	33.4	39.8	51.2
	AM-PW	71.7	77.6	86.9		AM-PW	62.7	70.3	79.5
	NW-PW	71.1	76.9	86.5		NW-PW	62.2	69.9	79.0

Table 4.2: Coverage Probabilities for AR(1) Processes: $n = 200$

ϕ_1	**Method**	**90%**	**95%**	**99%**	ϕ_1	**Method**	**90%**	**95%**	**99%**
	WHC	87.3	93.1	98.3		WHC	85.2	91.3	97.2
	WHREML	87.8	93.3	98.3		WHREML	87.3	92.8	97.8
0.1	WHPZ	86.9	92.7	98.4	0.3	WHPZ	81.2	88.0	95.7
	AM-PW	89.6	94.8	99.0		AM-PW	89.6	94.6	98.9
	NW-PW	88.4	94.2	98.8		NW-PW	88.6	94.1	98.8
	WHC	84.7	91.0	96.7		WHC	85.0	90.5	96.6
	WHREML	87.9	93.1	97.7		WHREML	87.5	92.6	97.6
0.5	WHPZ	77.3	84.4	92.9	0.7	WHPZ	79.4	85.8	93.2
	AM-PW	89.0	94.1	98.8		AM-PW	88.0	93.4	98.3
	NW-PW	88.5	93.8	98.7		NW-PW	87.9	92.9	98.4
	WHC	84.1	89.4	94.8		WHC	82.0	87.3	93.4
	WHREML	85.6	90.8	95.7		WHREML	82.8	88.1	94.1
0.9	WHPZ	67.2	74.8	86.2	0.95	WHPZ	53.1	60.4	73.0
	AM-PW	84.8	89.6	96.0		AM-PW	79.9	86.2	92.9
	NW-PW	84.7	89.4	95.9		NW-PW	79.9	86.0	92.8

Table 4.3: AR(1) Processes Relative Efficiency

ϕ_1	**Method**	$n = 50$	$n = 200$
0.1	WHC	0.66	0.12
	AM-PW	1.00	1.00
	NW-PW	0.52	0.26
0.3	WHC	0.57	0.14
	AM-PW	1.00	1.00
	NW-PW	0.74	0.46
0.5	WHC	0.57	0.29
	AM-PW	1.00	1.00
	NW-PW	0.89	0.80
0.7	WHC	0.65	0.42
	AM-PW	1.00	1.00
	NW-PW	0.94	0.78
0.9	WHC	0.99	0.97
	AM-PW	1.00	1.00
	NW-PW	0.98	0.96
0.95	WHC	1.00	1.00
	AM-PW	0.90	0.91
	NW-PW	0.89	0.90
Minimum	WHC	0.57	0.12
	AM-PW	0.90	0.91
	NW-PW	0.52	0.26

For the AR(1) process, the peak at zero frequency becomes sharper when the AR(1) coefficient ϕ_1 increases. In all the cases, the actual coverage probabilities are smaller than the nominal coverage probabilities. When $\phi = 0.1, 0.3, 0.5, 0.7, 0.9$, the method with relative efficiency of 1 is AM-PW for both $n = 50$ and 200. For $\phi_1 = 0.95$, the method with relative efficiency of 1 is WHC for both $n = 50$ and $n = 200$.

For $\phi = 0.1, 0.3, 0.5, 0.7, 0.9$, the prewhitening-based method is superior because the order of the autoregressive prewhitening filter is the same as the order of the autoregressive data-generating process. In other words, this is precisely the process that AM-PW and NW-PW are built for. We see that even with $WHREML$, where we restrict our choice of candidates to autoregressive models, $WHREML$ is still outperformed by AM-PW and NW-PW. One explanation to that is $WHREML$ does not always choose the true order (In this case, 1) especially in small samples. However, as we will see in subsection 4.2.5, the lack of model selection in the prewhitening filter can lead to undesirable performance in terms of coverage probability of μ.

The bandwidth selection procedure of AM-PW follows the proposal of Andrews (1991) and works well when the prewhitened process has a monotonically decreasing spectral density. The least-squares AR(1) estimator tends to yield an AR(1) coefficient that is biased downward. Therefore, the spectral density of the prewhitened time series from using the least-squares AR(1) filter can still be monotonically decreasing, which favors AM-PW.

Even though AM-PW and NW-PW are built for AR(1) processes due to AR(1) filter application in both methods, they are still outperformed by WHC in the case when $\phi = 0.95$ and by $WHREML$ when $\phi = 0.9$ and 0.95. These simulation results support the motivation behind the application of REML to estimate autoregressive processes in the HAC problem. Hence, we suggest to use REML whenever we are estimating an autoregressive model.

Finally, the gap between the observed coverage rate and the nominal coverage rate of WHC has narrowed when n moves from 50 to 200.

4.2.2 White Noise Process

$$X_t = \varepsilon_t \qquad \varepsilon_t \overset{i.i.d}{\sim} N(0,1)$$

Table 4.4: Coverage Probabilities for White Noise Process

n	**Method**	**90%**	**95%**	**99%**
	WHC	88.8	93.4	98.3
	WHREML	88.9	93.5	98.3
50	WHPZ	89.7	94.5	98.8
	AM-PW	88.1	93.1	97.9
	NW-PW	85.5	90.9	96.4
	WHC	89.5	94.6	99.0
	WHREML	89.7	94.6	99.1
200	WHPZ	89.9	94.9	99.2
	AM-PW	89.7	94.7	99.0
	NW-PW	88.4	94.1	98.8

Table 4.5: White Noise Process Relative Efficiency

Method	$n = 50$	$n = 200$
WHC	1.00	0.76
AM-PW	0.86	1.00
NW-PW	0.46	0.35

For the white noise process, the spectral density function is flat. For both $n = 50$ and 200, the actual coverage probabilities are smaller than the nominal coverage probabilities. The methods with the highest relative efficiency between WHC, AM-PW, and NW-PW are WHC for $n = 50$ and AM-PW with a slight advantage when $n = 200$ (Though based on Table 4.4, all methods provide satisfactory performance when $n = 200$). For WHC, these results show the advantage of the inclusion of nonparametric spectrum estimates as we see from the table that $WHPZ$ has better performance than $WHREML$. The only method that has

issue in this experiment is NW-PW. Its under-coverage of the confidence interval is substantial when $n = 50$. Finally, the gap between the observed coverage rate and the nominal coverage rate of WHC has narrowed when n moves from 50 to 200.

4.2.3 MA(1) Processes

Table 4.6: Coverage Probabilities for MA(1) Processes

		$n = 50$			$n = 200$		
ψ_1	**Method**	**90%**	**95%**	**99%**	**90%**	**95%**	**99%**
	WHC	92.0	95.3	98.4	90.2	95.2	99.1
	WHREML	92.4	95.5	98.3	90.2	95.2	99.0
-0.3	WHPZ	95.2	97.7	99.5	96.0	98.3	99.8
	AM-PW	92.2	95.9	99.1	93.1	97.0	99.7
	NW-PW	86.0	90.9	96.4	89.4	94.5	98.9
	WHC	92.0	95.6	98.7	91.7	95.8	99.1
	WHREML	91.9	95.4	98.6	91.7	95.8	99.1
-0.5	WHPZ	97.2	98.7	99.7	97.0	98.5	99.8
	AM-PW	96.6	98.5	99.7	97.7	99.4	100.0
	NW-PW	84.7	89.7	95.3	89.4	94.5	98.6
	WHC	95.8	97.7	99.3	94.8	97.9	99.7
	WHREML	95.7	97.7	99.3	95.0	98.0	99.7
-0.7	WHPZ	99.2	99.8	100.0	98.4	99.4	100.0
	AM-PW	99.5	99.9	100.0	100.0	100.0	100.0
	NW-PW	85.5	91.0	95.4	87.5	92.4	96.6

Table 4.7: MA(1) Processes Relative Efficiency

ψ_1	**Method**	$n = 50$	$n = 200$
	WHC	1.00	1.00
-0.3	AM-PW	0.31	0.08
	NW-PW	0.05	0.23
	WHC	1.00	1.00
-0.5	AM-PW	0.11	0.09
	NW-PW	0.09	1.00
	WHC	1.00	1.00
-0.7	AM-PW	0.20	0.00
	NW-PW	0.64	0.99
	WHC	1.00	1.00
Minimum	AM-PW	0.11	0.00
	NW-PW	0.05	0.23

For all values of ψ_1, the intervals based on WHC and AM-PW overcover μ and those based on NW-PW undercover μ. When $\psi_1 = -0.3, -0.5, -0.7$, WHC has more reliable performance than AM-PW and NW-PW.

We take different values of the MA(1) coefficient. When the MA(1) coefficient ψ_1 approaches to -1, $f(0)$ approaches 0. This provides an alternative way of doing stress-testing for our method and traditional HAC methods, as we will have a trough of spectral density at zero frequency when ψ_1 is close to -1. Note that when $f(0) = 0$, our assumption of short memory will no longer be true, and the HAC standard error will not be consistent. When $\psi_1 = -0.7$, AM-PW have a coverage probability of 100% even for a 90% confidence interval when $n = 200$.

Finally, the gap between the observed and nominal coverage rates for WHC is narrowed when n moves from 50 to 200 in all cases expect for ψ_1 being close to -1.

4.2.4 MA(2) and MA(3) Processes

$$X_t = \varepsilon_t + \alpha\varepsilon_{t-1} + \beta\varepsilon_{t-q} \qquad q \in \{2,3\} \quad \text{and} \quad \varepsilon \overset{i.i.d}{\sim} N(0,1)$$

Table 4.8: Coverage Probabilities for MA(2) and MA(3) Processes: $n = 50$

α	β	q	**Method**	**90%**	**95%**	**99%**	q	**Method**	**90%**	**95%**	**99%**
			WHC	93.2	96.6	99.0		WHC	95.7	97.7	99.4
			WHREML	93.2	96.5	99.0		WHREML	95.8	97.7	99.4
0.0	-0.3	2	WHPZ	96.7	98.7	99.8	3	WHPZ	97.6	99.2	99.9
			AM-PW	97.2	99.0	99.8		AM-PW	96.9	98.8	99.8
			NW-PW	86.1	90.3	95.5		NW-PW	95.3	97.8	99.6
			WHC	94.0	96.9	99.2		WHC	96.8	98.2	99.4
			WHREML	94.0	96.7	99.2		WHREML	96.5	98.1	99.4
-0.1	-0.3	2	WHPZ	97.8	99.2	99.9	3	WHPZ	99.0	99.7	99.9
			AM-PW	98.2	99.4	99.9		AM-PW	98.7	99.6	99.9
			NW-PW	86.0	90.5	95.9		NW-PW	96.4	98.4	99.7
			WHC	83.3	89.4	95.6		WHC	82.3	89.1	95.3
			WHREML	83.7	89.4	95.9		WHREML	82.8	89.6	95.6
0.0	0.3	2	WHPZ	83.4	89.7	96.1	3	WHPZ	81.5	88.6	95.6
			AM-PW	78.5	85.6	93.7		AM-PW	78.8	86.0	93.8
			NW-PW	81.7	87.6	95.0		NW-PW	77.4	84.4	92.3
			WHC	82.2	88.5	95.3		WHC	81.2	87.6	94.2
			WHREML	82.0	88.5	95.0		WHREML	81.5	87.7	94.4
0.1	0.3	2	WHPZ	82.0	88.7	95.4	3	WHPZ	80.2	87.2	94.3
			AM-PW	80.0	86.7	94.3		AM-PW	76.8	84.2	92.4
			NW-PW	82.8	88.3	95.2		NW-PW	76.2	83.4	92.0

Table 4.9: Coverage Probabilities for MA(2) and MA(3) Processes: $n = 200$

α	β	q	**Method**	**90%**	**95%**	**99%**	q	**Method**	**90%**	**95%**	**99%**
			WHC	91.7	95.9	99.0		WHC	93.4	97.0	99.4
			WHREML	91.7	95.8	99.0		WHREML	93.3	97.0	99.4
0.0	-0.3	2	WHPZ	96.9	98.8	99.8	3	WHPZ	97.5	99.0	99.9
			AM-PW	98.6	99.7	100.0		AM-PW	98.6	99.8	100.0
			NW-PW	89.7	94.3	98.6		NW-PW	89.8	94.1	98.5
			WHC	92.4	96.0	99.1		WHC	94.3	97.5	99.5
			WHREML	92.2	96.1	99.1		WHREML	94.3	97.5	99.5
-0.1	-0.3	2	WHPZ	97.7	98.9	99.9	3	WHPZ	98.3	99.3	100.0
			AM-PW	99.4	100.0	100.0		AM-PW	99.8	100.0	100.0
			NW-PW	89.7	94.0	98.4		NW-PW	88.1	92.9	97.4
			WHC	86.4	91.9	97.6		WHC	87.8	93.1	98.0
			WHREML	86.7	92.0	97.5		WHREML	88.6	93.4	98.1
0.0	0.3	2	WHPZ	84.2	90.1	97.1	3	WHPZ	85.1	91.2	97.6
			AM-PW	80.7	87.8	95.7		AM-PW	81.0	87.9	95.9
			NW-PW	86.1	91.8	97.9		NW-PW	85.2	91.1	97.6
			WHC	85.7	91.4	97.3		WHC	87.9	93.4	97.9
			WHREML	86.4	91.6	97.4		WHREML	88.8	93.7	98.1
0.1	0.3	2	WHPZ	82.0	88.4	96.1	3	WHPZ	84.4	90.7	97.0
			AM-PW	81.9	88.8	96.3		AM-PW	79.3	86.0	94.7
			NW-PW	86.3	92.1	98.1		NW-PW	85.0	91.1	97.3

Table 4.10: MA(2) and MA(3) Processes Relative Efficiency

α	β	q	**Method**	$n = 50$	$n = 200$	q	$n = 50$	$n = 200$
0.0	-0.3	2	WHC	1.00	1.00	3	1.00	0.64
			AM-PW	0.26	0.07		0.56	0.10
			NW-PW	0.29	0.74		0.95	1.00
-0.1	-0.3	2	WHC	1.00	1.00	3	1.00	1.00
			AM-PW	0.23	0.05		0.42	0.15
			NW-PW	0.37	0.58		0.91	0.98
0.0	0.3	2	WHC	1.00	1.00	3	1.00	1.00
			AM-PW	0.70	0.52		0.75	0.35
			NW-PW	0.82	0.97		0.67	0.56
0.1	0.3	2	WHC	1.00	0.85	3	1.00	1.00
			AM-PW	0.84	0.57		0.77	0.26
			NW-PW	0.97	1.00		0.74	0.47
			WHC	1.00	0.64			
Minimum			AM-PW	0.23	0.05			
			NW-PW	0.29	0.47			

Den Haan and Levin (1997) suggest these generating mechanisms to compare different HAC estimators' robustness against various autocorrelation structures. The parameters are chosen such that the first-order autocorrelation for the prewhitened time series is small, but the higher-order autocorrelations are substantial. Compared with AM-PW and NW-PW, WHC is superior in most situations and yields the best overall performance in this experiment. If the MA coefficient β is negative, then AM-PW tends to lead to substantial over-coverage of μ , but NW-PW tends to under-cover μ. If the MA coefficient β is positive, then both AM-PW and NW-PW tend to substantially undercover μ. WHC generally has better performance when β is negative and slightly better (though still with substantial under-coverage) when β is negative. In particular, we can see from the comparison between $WHREML$ and $WHPZ$ that $WHREML$ has better coverage performance even though our data-generating mechanism is not an autoregressive process.

A drawback of AM-PW here is that the bandwidth it uses is based on an AR(1) model for the prewhitened data $\{\hat{e}_t\}$. As is pointed out by Den Haan and Levin (1996), it is not true in general that the data-dependent bandwidth parameter should solely depend on the first-order autocorrelation of the prewhitened data. The bandwidth selection procedure of AM-PW follows the proposal of Andrews (1991) and works well when the prewhitened process has a monotonically decreasing spectral density (See Den Haan and Levin (1996)). Since this monotonicity may not hold in practice, such a predetermined fitting of an AR(1) model to $\{\hat{e}_t\}$ may have drawbacks, as seen here. Note that WHC avoids the use of pilot estimates as it is based on cross-validation. Overall, for the simulations in this subsection, NW-PW outperforms AM-PW, but WHC is superior.

Finally, the gap between the observed coverage rate and the nominal coverage rate of WHC narrows as n goes from 50 to 200.

4.2.5 AR(2) Processes

$$X_t = \frac{1}{2}\phi X_{t-1} + \frac{1}{2}\phi X_{t-2} + \varepsilon_t \qquad \varepsilon \overset{i.i.d}{\sim} N(0,1)$$

Table 4.11: Coverage Probabilities for AR(2) Processes: $n = 50$

ϕ	**Method**	**90%**	**95%**	**99%**	ϕ	**Method**	**90%**	**95%**	**99%**
	WHC	80.3	87.5	94.2		WHC	75.8	82.5	90.5
	WHREML	80.2	87.1	94.0		WHREML	76.4	82.8	90.6
0.3	WHPZ	79.4	87.1	94.4	0.5	WHPZ	72.5	80.2	89.9
	AM-PW	81.1	87.2	94.7		AM-PW	74.0	81.3	90.0
	NW-PW	81.2	87.4	94.7		NW-PW	75.9	83.1	91.5
	WHC	71.4	77.7	86.8		WHC	65.5	71.4	80.2
	WHREML	74.1	79.6	87.4		WHREML	68.3	74.4	82.5
0.7	WHPZ	61.9	70.3	81.9	0.9	WHPZ	38.8	44.9	57.5
	AM-PW	62.5	70.6	82.3		AM-PW	42.0	48.8	60.5
	NW-PW	66.9	74.0	85.1		NW-PW	45.8	53.0	65.1

Table 4.12: Coverage Probabilities for AR(2) Processes: $n = 200$

ϕ	**Method**	**90%**	**95%**	**99%**	ϕ	**Method**	**90%**	**95%**	**99%**
	WHC	84.6	90.7	96.8		WHC	84.4	90.5	96.2
	WHREML	85.6	91.3	97.2		WHREML	85.7	91.2	96.7
0.3	WHPZ	80.4	87.2	95.3	0.5	WHPZ	77.4	84.0	92.3
	AM-PW	83.1	89.7	96.8		AM-PW	77.5	84.5	93.3
	NW-PW	85.5	91.5	97.7		NW-PW	82.7	89.1	96.3
	WHC	84.5	89.8	95.8		WHC	83.4	88.4	93.4
	WHREML	86.3	91.1	96.4		WHREML	84.3	89.2	94.4
0.7	WHPZ	77.4	83.9	92.8	0.9	WHPZ	59.6	67.6	79.0
	AM-PW	69.2	77.9	87.9		AM-PW	53.5	61.3	73.2
	NW-PW	77.0	83.8	92.8		NW-PW	58.8	66.3	77.9

Table 4.13: AR(2) Processes Relative Efficiency

ϕ	**Method**	$n = 50$	$n = 200$
0.3	WHC	1.00	0.85
	AM-PW	0.97	0.73
	NW-PW	0.99	1.00
0.5	WHC	0.97	1.00
	AM-PW	0.92	0.55
	NW-PW	1.00	0.82
0.7	WHC	1.00	1.00
	AM-PW	0.82	0.46
	NW-PW	0.89	0.59
0.9	WHC	1.00	1.00
	AM-PW	0.68	0.37
	NW-PW	0.72	0.40
Minimum	WHC	0.97	0.85
	AM-PW	0.68	0.37
	NW-PW	0.72	0.40

This set of experiments was proposed by Den Haan and Levin (1997). The overall best method in this experiment is WHC. The inclusion of parametric autoregressive model candidates in FDCV is motivated by AR(1) prewhitening in the HAC literature. The prewhitening filter that AM-PW and NW-PW considered is a fixed first-order filter, where our method allows for model selection in choosing a parametric model. The advantage of such flexibility is not clear in experiment 4.2.1, where the data-generating mechanism is an AR(1) which is exactly the process that AM-PW and NW-PW are designed for. The value of the autoregressive coefficients is the same, taken to be $\frac{1}{2}\phi$. When ϕ increases, the time series will have a root that is close to the unit circle, and the spectrum density will be sharper. Notice that in all cases, the actual coverage probabilities are smaller than the nominal coverage probabilities, which is similar to experiment 4.2.1. The

performance of WHC, AM-PW and NW-PW are similar when $\phi = 0.3, 0.5$ for both $n = 50$ and 200. When $\phi = 0.7$ and 0.9, AM-PW and NW-PW are strongly outperformed by WHC. This shows the advantage of flexibility in selecting the parametric components for the HAC problem. The inclusion of an AR(2) model candidate in class $\mathbf{C}$ is helpful for WHC in the current situation.

In addition, similarly to what we have seen in 4.2.1, when $\phi_1 = 0.95$, the use of REML to estimate the autoregressive model improves the performance when the spectral density has a sharp peak ($\phi = 0.9$) in the simulation.

Finally, the gap between the observed coverage rate and the nominal coverage rate of WHC narrows as n goes from 50 to 200.

4.2.6 Overall Evaluation

We have run simulations with 22 data-generating mechanisms with $n = 50$ and 200. In this subsection, we point out that WHC is the best performing method in most of the 22 cases, and we also consider the worst-case performance for the various methods.

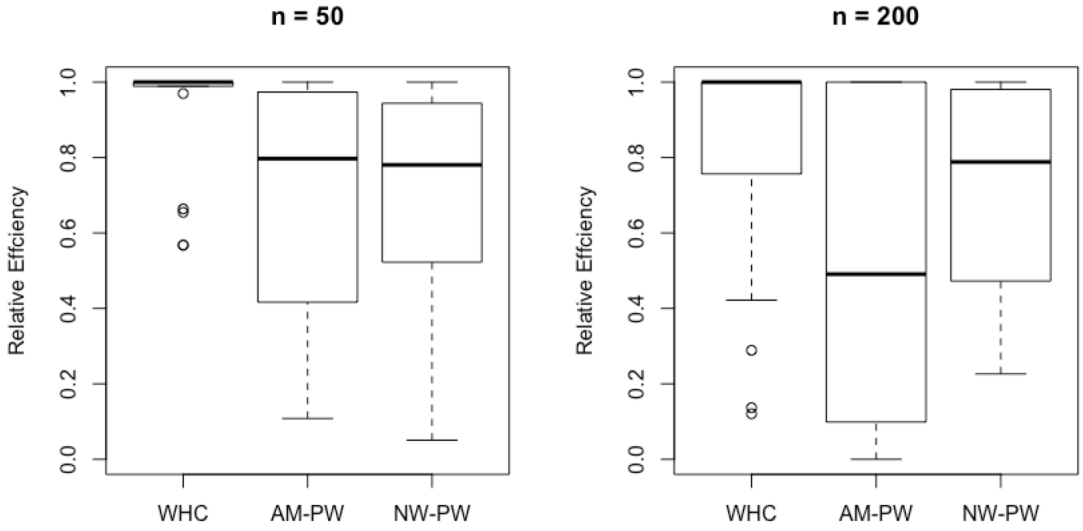
The first table below shows that WHC is the method that achieves a relative efficiency of 1 (the best method) 16 times when $n = 50$ and 12 times when $n = 200$. This is far greater than the number of times that AM-PW and NW-PW achieve a relative efficiency of 1.

Table 4.14: Number of Times a Method
Achieves Relative Efficiency of 1

Method	$n = 50$	$n = 200$
WHC	16	12
AM-PW	5	6
NW-PW	1	4

From the boxplots below, we first notice that the interquartile range is much smaller for WHC than AM-PW and NW-PW. Next, we investigate the worst-case performance of each method. From the boxplots, we see that when $n = 50$, WHC has the best worst-case performance. When $n = 200$, the worst case of

WHC gives a relative efficiency lower than the worst case for NW-PW. However, we should notice that the worst case of WHC occurs in the results of subsection 4.2.1 where the respective data-generating mechanism is AR(1), which is where the AM-PW and NW-PW are designed to do well, due to AR(1) prewhitening.



Another observation from reading the table of coverage probabilities across these 22 cases is that NW-PW has a tendency to undercover compared to WHC and AM-PW.

In view of this discussion, we feel that WHC has the most reliable performance overall.

Chapter 5

Conclusions

We apply a unified frequency domain cross-validation method to select an estimate of the spectral density at zero frequency and study the performance of confidence intervals for the mean based on the resulting HAC standard error. Unlike classical HAC methods, our method does unified model/tuning parameter selection where candidates cut across parametric and nonparametric estimators. In particular, we propose to automatically select the model/tuning parameters from a class of $\mathbf{C}$ consisting of REML-based autoregressive spectrum estimators of order 0 to 5 and lag-weights spectrum estimators with Parzen kernel from truncation point 1 to $m(n)$.

We studied the performance of the confidence interval of our purposed data-driven method compared to other popular plug-in-based approaches like Newey and West (1994) and Andrews and Monahan (1992) in the case of the mean. We found that our method is the best performing and the most reliable method in simulation. More specifically, our method has superior performance when the time series has an autoregressive root that is closed to the unit circle due to our inclusion of the REML-based autoregressive estimators. Moreover, the inclusion of autoregressive spectrum estimates can be advantageous even if our time series is not an autoregressive process, such as a moving average process. Finally, our method is reliable in the case of white noise where the spectral density is constant, due to the inclusion of the nonparametric lag-weights estimators and better choice of the bandwidth parameter.

For future work, we hope to improve the computational efficiency of the REML likelihood evaluation using the preconditioned conjugate gradient algorithm (PCG),

as we will briefly describe in Appendix A. A faster algorithm for computing the REML spectrum estimate will substantially improve the speed of our FDCV algorithm for the HAC problem and make our method more user-friendly.

We might also consider tapering for the nonparametric estimators. This is motivated by the simulation study where we found in most cases except for white noise that even if the spectral density around zero is relatively flat, the FDCV would still more frequently select a parametric estimator. Moreover, we generally found that $WHREML$ has better performance than $WHPZ$ even if our process is not autoregressive. Thus, we would like to improve the performance of the nonparametric spectrum estimators. One related work of this topic on the HAC problem is Smith (2005) who proposes using multitapering for the HAC standard error estimation. However, his ideas have not yet been verified in simulation so far as we are aware. In general, tapering the data will reduce the bias at the cost of inflating the variance. However, if we try to apply the smoothing directly on the non-tapered data, the bias will persist as smoothing only dampens the variance. So, to apply a nonparametric estimator on non-tapered data is to smooth out what has already been biased. We would find the tapers that can noticeably reduce the bias without inflating too much of the variance and then applying the nonparametric estimators on the tapered data.

Finally, HAC focuses on robust standard error estimation in a short memory process. In most of the HAC literature, it is assume that the spectral density at zero frequency is finite and positive. However, if our time series has $f(0)$ being infinite or zero (so that the time series has long memory or is anti-persistent), then the HAC methods based on estimating $f(0)$ are no longer consistent. Details of such problems can be found in Robinson (2005) where an alternative MAC (memory autocorrelation consistent) estimator is considered. We would hope to investigate the possibility of the application of FDCV to the MAC problem.

References

Akaike, Hirotugu (1974). "A new look at the statistical model identification". In: *IEEE transactions on automatic control* 19.6, pp. 716–723.

Andrews, Donald WK (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation". In: *Econometrica: Journal of the Econometric Society*, pp. 817–858.

Andrews, Donald WK and J Christopher Monahan (1992). "An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator". In: *Econometrica: Journal of the Econometric Society*, pp. 953–966.

Barndorff-Nielsen, Ole and Geert Schou (1973). "On the parametrization of autoregressive models by partial autocorrelations". In: *Journal of multivariate Analysis* 3.4, pp. 408–419.

Beltrao, Kaizo and Peter Bloomfield (1987). "Determining the bandwidth of a kernel spectrum estimate". In: *Journal of time series analysis* 8.1, pp. 21–38.

Berk, Kenneth N (1974). "Consistent autoregressive spectral estimates". In: *The Annals of Statistics*, pp. 489–502.

Brockwell, Peter J, Richard A Davis, et al. (1991). *Time series: theory and methods: theory and methods*. Springer Science & Business Media.

Cheang, Wai-Kwong and Gregory C Reinsel (2000). "Bias reduction of autoregressive estimates in time series regression model through restricted maximum likelihood". In: *Journal of the American Statistical Association* 95.452, pp. 1173–1184.

Chen, Willa W and Rohit S Deo (2012). "The restricted likelihood ratio test for autoregressive processes". In: *Journal of Time Series Analysis* 33.2, pp. 325–339.

Chen, Willa W, Clifford M Hurvich, et al. (2006). "On the correlation matrix of the discrete Fourier transform and the fast solution of large Toeplitz systems for long-memory time series". In: *Journal of the American Statistical Association* 101.474, pp. 812–822.

Christiano, Lawrence J and Wouter J Den Haan (1996). "Small-sample properties of GMM for business-cycle analysis". In: *Journal of Business & Economic Statistics* 14.3, pp. 309–327.

Den Haan, Wouter J and Andrew Levin (1996). *Inferences from parametric and non-parametric covariance matrix estimation procedures*. Tech. rep. National Bureau of Economic Research.

— (1997). "A practitioner's guide to robust covariance matrix estimation". In: *Handbook of statistics* 15, pp. 299–342.

Galbraith, RF and JI Galbraith (1974). "On the inverses of some patterned matrices arising in the theory of stationary time series". In: *Journal of applied probability*, pp. 63–71.

Harville, David A (1974). "Bayesian inference for variance components using only error contrasts". In: *Biometrika* 61.2, pp. 383–385.

Hurvich, Clifford M (1985). "Data-driven choice of a spectrum estimate: extending the applicability of cross-validation methods". In: *Journal of the American Statistical Association* 80.392, pp. 933–940.

Lu, Yi and Clifford M Hurvich (2005). "On the Complexity of the Preconditioned Conjugate Gradient Algorithm for Solving Toeplitz Systems with a Fisher–Hartwig Singularity". In: *SIAM journal on matrix analysis and applications* 27.3, pp. 638–653.

Newey, Whitney K and Kenneth D West (1986). *A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix*. Tech. rep. National Bureau of Economic Research.

— (1994). "Automatic lag selection in covariance matrix estimation". In: *The Review of Economic Studies* 61.4, pp. 631–653.

Priestley, Maurice Bertram (1981). *Spectral analysis and time series: probability and mathematical statistics*. 04; QA280, P7.

Robinson, Peter M (2005). "Robust covariance matrix estimation: HAC estimates with long memory/antipersistence correction". In: *Econometric Theory*, pp. 171–180.

Schwarz, Gideon et al. (1978). "Estimating the dimension of a model". In: *Annals of statistics* 6.2, pp. 461–464.

Smith, Richard J (2005). "Automatic positive semidefinite HAC covariance matrix and GMM estimation". In: *Econometric Theory*, pp. 158–170.

Wahba, Grace and Svante Wold (1975). "A completely automatic french curve: fitting spline functions by cross validation". In: *Communications in Statistics-Theory and Methods* 4.1, pp. 1–17.

Appendix A

Restricted Maximum Likelihood Estimation

There are many popular approaches to estimate an autoregressive model, including Yule-Walker, Burg, least-squares (LS) and maximum likelihood (MLE). Consider an AR(p) process

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + \varepsilon_t \qquad \varepsilon_t \overset{i.i.d}{\sim} N(0, \sigma^2)$$

Let $\phi(z) = 1 - \sum_{j=1}^{p} \phi_j z^j$ be a polynomial of order p and $z \in \mathbb{C}$. A sufficient condition for the existence of a stationary autoregression is that all the roots z_j of $\phi(z)$ lie outside the unit circle (or $|z_j| > 1$ and $1 \leq j \leq p$). Since HAC methods assume that the time series in stationary, we impose this constraint in our estimates of AR(p) models. Yule-Walker and Burg always generate a stationary solution and Burg often has much better estimation performance than Yule-Walker. The restricted likelihood function is not well-defined for non-stationary models, which distinguishes it from the Yule-Walker and Burg methods. All methods described above except for REML have been widely implemented in various statistical software. REML is a method that has been widely applied in the literature of linear mixed effect models, but no package so far as we are aware has a version of REML that constrains for stationarity. We next discuss REML estimation of AR(p) models with a stationarity constraint.

A.1 The Restricted Likelihood for an Linear Regression Model

Consider a general linear regression model

$$y = X\beta + \varepsilon \qquad \varepsilon \sim N(0, H(\phi))$$

where y is an n-dimensional vector, X is a design matrix and $H(\phi)$ is the covariance matrix parametrized by ϕ.

It is well-known that the covariance matrix estimator using MLE is biased. This is related to the degrees of freedom lost in the estimation of the mean components (β). If we estimate the covariance matrix with true β, the estimation would be unbiased. The intuition behind REML is to maximize a modified likelihood that is free of mean components.

Instead of computing likelihood $L(\phi|y)$, we transform y such that the likelihood estimation can bypass estimating β first and can thus produce unbiased estimates for ϕ. If vector a is orthogonal to all columns of design matrix X, i.e.,$a^T X = 0$, then $a^T y$ is known as an error contrast. We can find at most $n-k$ such vectors that are linearly independent. Define $A = (a_1, a_2, ...a_{n-k})$. It follows that $A^T X = 0$ (by orthogonality construction) and $E(A^T y) = 0$. Let $A = I_n X(X^T X)^{-1} X^T$ and we will have $AX = 0$. The error contrast vector

$$w = A^T y = A^T(X\beta + \varepsilon) = A^T \varepsilon \sim N(0, A^T H A)$$

is free of β. In the case of original maximum likelihood, $y \sim N(X\beta, H)$, which is not free of β. Therefore, REML intends to maximize

$$L_w(\phi|A^T y)$$

Harville (1974) derives the formula for $L(\theta|A^T y)$

$$\begin{aligned} L(\theta|A^T y) = &-\frac{1}{2}(n-k)log(2\pi) + \frac{1}{2}log|X^T X| - \frac{1}{2}log|H| \\ &- \frac{1}{2}log|X^T H^{-1} X| - \frac{1}{2}(y - X\hat{\beta})^T H^{-1}(y - X\hat{\beta}) \end{aligned}$$

where $\hat{\beta} = (X^T H^{-1} X)^{-1} X^T H^{-1} y$

Once $H(\theta)$ is obtained from REML, β can be estimated using generalized least

squares and is just

$$\hat{\beta} = (X^T H^{-1} X)^{-1} X^T H^{-1} y$$

A.2 The Restricted Likelihood for an Autoregressive Model

The autoregressive model of order p is a particular case of the model described above. Our design matrix X is simply a vector of 1's, and we specify our error structure to be AR(p).

By Chen and Deo (2012), using the Harville (1974) formula, up to an additive constant, based on $X = (x_1, ..., x_n)^T$, the restricted log-likelihood is given by

$$\begin{aligned} L(X, \phi, \sigma^2) = &-\frac{n-1}{2} log\sigma^2 + \frac{1}{2} log \frac{|\Sigma^{-1}(\phi)|}{|W^T \Sigma^{-1}(\phi) W|} \\ &- \frac{1}{2\sigma^2} \{X^T \Sigma^{-1}(\phi) X - X^T \Sigma^{-1}(\phi) W (W^T \Sigma^{-1}(\phi) W)^{-1} W^T \Sigma^{-1}(\phi) W\} \end{aligned}$$

where $\phi = (\phi_1, ..., \phi_p)^T$, $W = (1, ..., 1)^T$ and $\Sigma(\phi) = \sigma^2 Var(X)$

In particular,

$$\Sigma(\phi) = \sigma^2 \begin{bmatrix} c_0 & c_1 & c_2 & ... & c_{n-2} & c_{n-1} \\ c_1 & c_0 & c_1 & ... & c_{n-3} & c_{n-2} \\ ... & ... & ... & ... & ... & ... \\ c_{n-1} & c_{n-2} & ... & ... & c_1 & c_0 \end{bmatrix}$$

is a Toeplitz matrix.

A.2.1 Computation

The terms in the restricted likelihood that are computationally expensive are $|\Sigma^{-1}(\phi)|$, $W^T \Sigma^{-1}(\phi) W$ and $X^T \Sigma^{-1}(\phi) X$. Notice that inverting a matrix without assuming any structure will cost $O(n^3)$ operations. We investigate ways to decrease the computational cost.

Method 1

This method is based on Chen and Deo (2012) proposal of a stable formula for computing the restricted likelihood. By Barndorff-Nielsen and Schou (1973), $|\sum^{-1}(\phi)| = \Pi_{i=1}^{p} (1 - \phi_{ii}^2)^i$ and the quadratic terms $X^T \Sigma^{-1}(\phi) W$, $X^T \Sigma^{-1}(\phi) X$ and $W^T \Sigma^{-1}(\phi) W$ can be computed using the result of R. Galbraith and J. Galbraith

(1974) that

$$u^T\Sigma^{-1}(\phi)w = \sum_{r=1}^{p}\sum_{s=1}^{p} u_r w_s m_{rs} + \sum_{t=p+1}^{n}\Big(u_t - \sum_{i=1}^{p}\phi_i u_{t-i}\Big)\Big(w_t - \sum_{i=1}^{p}\phi_i w_{t-i}\Big)$$

where u and w are any n-dimensional column vectors,

$m_{rs} = \sum_{j=0}^{r-1}\phi_j\phi_{j+s-r} - \sum_{j=p+1-s}^{p+r-s}\phi_j\phi_{j+s-r}$, $1 \le r \le s \le p$

and $m_{rs} = m_{sr}$

Method 2

An alternative and computationally less costly way is to use the preconditioned conjugate gradient algorithm (PCG) which can speed up the evaluations of the restricted maximum likelihood by solving the Toeplitz system $Ax = b$ (See Lu and Hurvich (2005) and Chen, Hurvich, et al. (2006). We can replace the parts of the Restricted likelihood that contain the product of the inverse of a Toeplitz matrix and a known column by the solution to the system $Ax = b$. The solution is just $x = A^{-1}b$. In our case, the A is Σ and b can be either W or X.

A.2.2 Constraint for Stationarity

We wish to restrict our search region to the parameter space that will produce a stationary solution. One way to do this is to optimize the restricted likelihood function with respect to the partial correlations.

To constrain for stationarity, we can use the Durbin-Levinson recursion to express the autoregressive parameters $\phi = (\phi_1, .., \phi_p)^T$ (stationary) using the partial autocorrelation function (PACF) $\rho = (\phi_{11}, \phi_{22}, ..., \phi_{pp})$ and perform the optimization of the REML likelihood for $\rho \in (-1, 1)^p$. (One can also utilize the Logit transform such that the transformed PACF will be in $(-\infty, \infty)^p$). We take the starting value for REML estimation to be the Burg estimator as it yields a stationary solution. We perform our optimization based on the following steps:

Step 1: Compute the Autoregressive Burg estimate $\hat{\phi}^{Burg} = (\hat{\phi}_1^{Burg}, .., \hat{\phi}_p^{Burg})^T$ and $\hat{\sigma}^2$

Step 2: Transform the Burg estimate $\hat{\phi}^{Burg}$ to PACF $\rho \in (-1, 1)^p$ by first setting $(\phi_{p1}, ..., \phi_{pp}) = \hat{\phi}^{Burg}$

and apply recursion

$$\phi_{h-1,i} = \phi_{h,i} + \phi_{h,h}\phi_{h-1,h-i} \quad i = 1, .., h-1$$

for $h = 2, ..., p$. This recursion follows directly from the transposition of the terms in the Durbin-Levinson recursion. The pacf transformed from the initial Burg autoregressive estimate is

$$\rho^{Burg} = (\phi_{11}, \phi_{22}, ..., \phi_{pp})^T$$

Note that $\hat{\phi}^{Burg}$ and $\hat{\sigma}^2$ is the starting value of our restricted likelihood based on PACF.

Step 3: Do the searching for $\rho \in (-1, 1)^p$ and $\sigma^2 > 0$ and obtain the ρ^* and $(\sigma^2)^*$ that maximizes the restricted log likelihood.

Step 4: Apply Durbin-Levinson Recursion so that we can transform the PACF to ϕ.

$$\phi_{h,i} = \phi_{h-1,i} - \phi_{h,h}\phi_{h-1,h-i} \quad i = 1, 2, ..., h-1$$

This will be our REML estimate that constrained for stationarity.