1-2015

Three essays on financial econometrics

Jiang LIANG

THREE ESSAYS ON FINANCIAL ECONOMETRICS

LIANG JIANG

SINGAPORE MANAGEMENT UNIVERSITY
2015

Three Essays on Financial Econometrics

by
Liang Jiang

Submitted to School of Economics in partial fulfillment of the
requirements for the Degree of Doctor of Philosophy in Economics

Dissertation Committee:

Peter C.B. Phillips (Supervisor/Co-Chair)
Sterling Professor of Economics and Statistics
Yale University
Distinguished Term Professor of Economics
Singapore Management University

Jun Yu (Supervisor/Co-Chair)
Professor of Economics and Professor of Finance
Singapore Management University

Sock Yong Phang
Celia Moh Professor and Professor of Economics
Singapore Management University

Kian Guan Lim
OUB Chair Professor and Professor of Finance
Singapore Management University

Singapore Management University
2015

Abstract

Three Essays on Financial Econometrics

Liang Jiang

This dissertation develops several econometric techniques to address three issues in financial economics, namely, constructing a real estate price index, estimating structural break points, and estimating integrated variance in the presence of market microstructure noise and the corresponding microstructure noise function.

Chapter 2 develops a new methodology for constructing a real estate price index that utilizes all transaction price information, encompassing both single-sales and repeat-sales. The method is less susceptible to specification error than standard hedonic methods and is not subject to the sample selection bias involved in indexes that rely only on repeat sales. The methodology employs a model design that uses a sale pairing process based on the individual building level, rather than the individual house level as is used in the repeat-sales method. The approach extends ideas from repeat-sales methodology in a way that accommodates much wider datasets. In an empirical analysis of the methodology, we fit the model to the private residential property market in Singapore between Q1 1995 and Q2 2014, covering several periods of major price fluctuation and changes in government macroprudential policy. The index is found to perform much better in out-of-sample prediction exercises than either the S&P/Case-Shiller index or the index based on standard hedonic methods. In a further empirical application, the recursive dating method of Phillips, Shi and Yu (2015a, 2015b) is used to detect explosive behavior in the Singapore real estate market. Explosive behavior in the new index is found to arise two quarters earlier than in the other indices.

Chapter 3, based on the Girsanov theorem, obtains the exact finite sample distribution of the maximum likelihood estimator of structural break points in a continuous time model. The exact finite sample theory suggests that, in empirically realistic situations, there is a strong finite sample bias in the estimator of structural break points. This property is shared by least squares estimator of both the absolute structural break point and the fractional structural break point in discrete time models. A simulation-based method based on the indirect estimation approach is proposed to reduce the bias both in continuous time and discrete time models. Monte Carlo studies show that the indirect estimation method achieves substantial bias reductions. However, since the binding function has a slope less than one, the variance of the indirect estimator is larger than that of the original estimator.

Chapter 4 develops a novel panel data approach to estimating integrated variance and testing microstructure noise using high frequency data. Under weak conditions on the underlying efficient price process and the nature of high frequency noise contamination, we employ nonparametric kernel methods to estimate a model that accommodates a very general formulation of the effects of microstructure noise. The methodology pools information in the data across different days, leading to a panel model form that enhances efficiency in estimation and produces a convenient approach to testing the linear noise effect that is conventional in existing procedures. Asymptotic theory is developed for the nonparametric estimates and test statistics.

Table of Contents

Acknowledgements

Pursuing the Ph.D. in Economics at SMU has transformed my life. At this moment, I would like to thank all those who have helped me in this Ph.D. journey.

First and foremost, I wish to express my deepest gratitude to my supervisors, Professor Peter C.B. Phillips and Professor Jun Yu. They are the greatest teachers, inspiring me to explore the uncharted territory of knowledge, and extraordinary mentors, guiding me, with enormous patience and encouragement, through the tough times in the Ph.D. pursuit. Without their enlightening teaching, ardent advising, constant encouragement and invaluable help, this dissertation would not have been completed. They are also my role models of first-class researchers, outstanding supervisors and marvelous human beings with shining qualities. Their everlasting enthusiasm and passionate dedication to reaserch are not only motivational and inspirational for me during this Ph.D. study period, but will have a lasting influence on my future research and life. I am so fortunate to have them as my supervisors and my gratitude to them cannot be expressed in words.

My sincere appreciation goes to two committee members, Professor Kian Guan Lim and Professor Sock Yong Phang, for their insightful suggestions and helpful guidance during the course of this research. They are the distinguished professors and renowned experts, from whose enlightening comments I benefit a greal deal. I feel grateful for the opportunity to learn from them.

My gratitude extends to Professor Sungbae An, Professor Thomas Sargent and Professor Liangjun Su, who made valuable comments on some chapters in this dissertation, and to our Ph.D. program directors Professor Anthony Tay and Professor Yiu Kuen Tse for all the help and support.

I am also very thankful for all my friends and fellow classmates at SMU, in particular, Professor Ye Chen, Professor Yingjie Dong, Jianhui Huang, Professor Yong Li, Professor Cheng Liu, Peng Liu, Xiaobin Liu, Vishrut Rena, Professor Xiaohu Wang, Sijia Xu, Huaxia Zeng, Ming Zeng, Professor Tao Zeng, Professor Yonghui Zhang.

Last but not least, I would like to specially thank my parents and my wife for their endless love, and for always being there for me.

Chapter 1 Introduction

It has been widely accepted that financial markets are the indispensable pillars in the modern economy. In this dissertation, several econometric techniques have been developed to address three statistical issues in the analysis of time series data on financial markets.

The first issue addressed in this dissertation concerns the construction of a suitable real estate price index. Real estate prices are one of the key indicators of economic activity. Indices measuring changes in real estate prices help to inform households about their asset wealth and to make a wide variety of economic decisions that depend on wealth resources. Policy makers rely on the information provided by these indices when designing and formulating monetary and fiscal policies at the aggregate level as well as macro-prudential policies directed at the financial and banking sectors.

In Chapter 2, we propose a new methodology for constructing a real estate price index from transaction data. In our method, all transaction price information, including both single-sales and repeat-sales, is exploited. Meanwhile, our approach is more robust to specification error than standard hedonic methods and is not subject to the sample selection bias involved in indices that rely only on repeat sales. The methodology creates sale pairs on the individual building level, rather than the individual house level as is used in the repeat-sales method, therefore it accommodates much wider datasets than the repeat-sales method. In this regard it is an extension of the ideas of repeat-sales methods. In an empirical analysis of the methodology, we apply the method to the private residential property market in Singapore between Q1 1995 and Q2 2014, covering several periods of major price fluctuation and changes in government macroprudential policy. The new index is found to perform much better in out-of-sample prediction exercises than either the S&P/Case-Shiller index or the index based on standard hedonic methods. In a further empirical application, the recursive dating method of Phillips, Shi and Yu (2015a, 2015b) is used to detect explosive behavior in the Singapore real estate market. Explosive behavior in the new index is found to arise two quarters earlier than in the other indices.

The second issue addressed in this disseration deals with estimating structural break points. Statistical inference of structural breaks has received a great deal of attention in both econometrics and statistics literature over the last several decades. In terms of estimating structure break points, the literature has developed asymptotic theory for estimating the (fractional) structural break point (i.e. the (absolute) structural break point divided by the total sample size), including consistency, rates of convergence, and limit distributions; see, for example, Yao (1987) and Bai (1994). Interestingly and rather surprisingly, the finite sample theory for estimating structure break points seems to have received little attention in the literature. However, there are two pieces of conflicting evidence in simulations. Simulations in Yao (1987) seem to suggest that the asymptotic distribution is not necessarily close to the finite sample distribution, whereas simulations in Bai (1994) seem to suggest there is little bias in the traditional estimator when the true break point is in the middle of the sample.

In Chapter 3, we systematically investigate the finite sample properties and the bias problem in the estimation of structural break points. we use the Girsanov theorem to obtain the exact finite sample distribution of the maximum likelihood estimator of structural break points in a continuous time model. The exact finite sample theory suggests that, that when the true break point is at the middle of the sample, the finite sample distribution is symmetric but can have tri-modality. However, when the true break point occurs earlier (later) than the middle of the sample, the finite sample distribution is skewed to the right (left) and there is a positive (negative) bias. We also establish its connection to the discrete time models considered in the literature. A simulation-based method based on the indirect estimation approach is proposed to reduce the bias both in continuous time and discrete time models. Monte Carlo studies show that the indirect estimation method achieves substantial bias reductions. However, since the binding function has a slope less than one, the variance of the indirect estimator is larger than that of the original estimator.

The third issue addressed in this disseration is on estimating integrated variance in the presence of market microstructure noise and the corresponding microstructure noise function. The last two decades have witnessed substantial progress in financial assets using ultra high frequency data. Much of the research has concentrated on measuring volatility of financial assets from high frequency data. The resulting quantity is known as realized variance (RV) that estimates integrated variance (IV). The method is nonparametric in nature because one does not need to impose any parametric assumption to describe the dynamics of the true efficient price of the underlying asset.

However, estimating IV at very high frequencies makes it necessary to take into account the presence of market microstructure noise. New methods have been proposed to estimate IV that deal with the microstructure noise; prominent examples are Zhang et al (2005), Bandi and Russell (2008) and Bardnorff-Nielsen et al (2008). In the literature, specific parametric assumptions have been made about the microstructure noise. For example, a commonly made assumption is the pure noise which corresponds to the identical and independently distributed (IID) noise that is independent of the true efficient price. As argued in many papers (such as Hansen and Lunde, 2006), this assumption is too strong. For example, Phillips and Yu (2006) showed that the pure noise assumption fails to produce the so-called flat pricing phenomenon, an empirical regularity that is widely observed in data at medium and ultra high frequencies.

In Chapter 4, we provide a new method to model, estimate and test the effect of microstructure noise on estimating the integrated variance using high frequency data. Different from the time series-based approaches adopted in the literature, we develop a novel panel data approach to model the microstructure noise function nonparametrically. Our method is therefore able to allow for a much wider class of assumptions for microstructure noise and provides a convenient way to test the noise specification and to estimate the integrated variance.

Chapter 2 New methodology for constructing real estate price indices applied to the Singapore residential market

2.1 Introduction

Real estate prices are one of the key indicators of economic activity. Indices measuring changes in real estate prices help to inform households about their asset wealth and to make a wide variety of economic decisions that depend on wealth resources. Policy makers rely on the information provided by these indices in their design and formulation of monetary and fiscal policies at the aggregate level as well as macro-prudential policies directed at the financial and banking sectors. Though real estate prices are widely accepted as highly important economic statistics,[1] the construction of a suitable index that will reflect movements in the price of a typical house in the economy presents many conceptual, practical, and theoretical challenges.

First, houses are distinctive, making it particularly difficult to characterize a "typical" house for the development of an index. Different houses have varying characteristics such as location, size, ownership, utilities and indoor/outdoor facil-

[1] The recent literature has witnessed an upsurge of interest in studying real estate markets from perspectives of banking, financial and macroprudential policy. See, for example, the study of the relationship between real estate prices and banking instability (Koetter and Poghosyan, 2010; Reinhart and Rogoff, 2013), the market linkage among different assets (Chan et al., 2011), the impact of macro-prudential policy on housing prices (Shi et al., 2013; Mendicino and Punzi, 2014), the role of housing markets for macroeconomy (Iacoviello, 2005; Musso et al., 2011).

ities. These differences imply that averaging all market transaction prices without controlling for house heterogeneity inevitably produces bias. Second, house transactions are infrequent and sales data are unbalanced for several reasons. Most houses on the market are single-sale houses. Houses that have been sold more than once account for a small portion of the whole market in a typical dataset. Also, houses sold in one period can be quite different from those sold in other periods. These factors unbalance the pricing data and complicate econometric construction of a price index due to problems of heterogeneous, missing, and unequally spaced observations. Third, a typical presumption underlying construction of real estate price indices is that the average quality of properties in the market remains constant over time, whereas quality improvements in housing occurs continuously from advances in materials, design, utilities, and construction technologies. Meanwhile and in spite of ongoing maintenance, older dwellings age with the holding period, leading to some depreciation in house value. These countervailing effects can produce ambiguities regarding what movements in a real estate price index reflect: the underlying market situation or quality changes in the properties that happen to be sold. This problem is exacerbated in a fast growing real estate market where a substantial proportion of sales are new sales released directly from developers.

Two main approaches dominate the literature on real estate price indices: the hedonic regression method and repeat-sales method. The hedonic method assumes that house values can be decomposed into bundles of utility-bearing attributes that contribute to the observed heterogeneity in prices. Observed house prices may then be regarded as the composite sum of elements that reflect implicit structural and locational prices (Rosen, 1974). Hedonic methods for estimating a real estate price index employ regression techniques to control for various sources of heterogeneity in prices using observations on covariates and dummy variables that capture relevant characteristics. However, the choice of the covariates in such hedonic regressions is limited by data availability and involves subjective judgements by the researcher, which may lead to model *specification bias*. Moreover, Shiller (2008) argued that the hedonic approach can lead to spurious regression effects in which the irrelevant hedonic variables are significant. A further complication is that the precise relationship between hedonic information and sales prices is unknown, likely to be complex, and may well be house dependent.

Unlike the hedonic approach, which uses all transaction prices to create an index, the repeat-sales method uses only properties that are sold multiple times in the sample to track market trends. The technique was first introduced for building the real estate price index by Bailey, Muth, and Nourse (1963) and then extended to include time-dependent error variances in seminal and highly influential work by Case and Shiller (1987, 1989). The repeat-sales method seeks to avoid the problem of heterogeneity by looking at the difference in sale prices of the same house. No hedonic variables are needed, so the approach avoids the difficulties of choosing hedonic information and specifying functional forms. However, since the repeat-sales method confines the analysis only to houses that have been sold multiple times, it is natural to question whether repeat-sales are representative of the entire market and whether there exists significant *sample selection bias*. Clapp et al. (1991) and Gatzlaff and Haurin (1997) argued that the properties that are sold more than once could not represent the whole real estate market and the index estimated by the repeat-sales method is most likely subject to some sample selection bias. Moreover, large numbers of observations must be discarded because repeat-sales typically comprise only a small subset of all sales. Not surprisingly, the repeat-sales method has been criticized by researchers (e.g., Case et al., 1991; Nagaraja et al., 2010) for discarding too much data. On the other hand, while repeat-sales themselves may not be representative of the entire market, price changes in repeat-sales may still be representative of the market. Moreover, as argued in Shiller (2008), "there are too many possible hedonic variables that might be included, and if there are n possible hedonic variables, then there are $n!$ possible lists of independent variables in a hedonic regression, often a very large number. One could strategically vary the list of included variables until one found the results one wanted." As a result, Shiller (2008) made the strong claim that “the repeat-sales method is the only way to go” and this assertion has been influential. In the U.S., for instance, indices produced by the repeat-sales method, such as the FHFA and S&P/Case-Shiller home price indices, are now routinely reported in official government and industry statistics and they regularly attract media attention.

A combined approach, called the hybrid model, has been introduced as an alternative method of constructing house price indices. In particular, Case and Quigley (1991) proposed a hybrid model and applied generalized least squares (GLS) to jointly estimate the hedonic and repeat-sales equations. In subsequent work, Quigley (1995) and Englund et al. (1998) proposed to model explicitly the structure of the error terms in their hybrid model to improve the estimated price index. Hill et al. (1997) instead employed an AR(1) process to model the error dynamics of the hybrid model. Nagaraja, Brown and Zhao (2011) also relied on an underlying AR(1) model to build the hybrid model. To answer the question why hybrid models are better, Ghysels et al. (2012) explained that improved estimation in the hybrid model is analogous to the better forecasts gained by forecast combinations. The hedonic model has less sample selection bias but potentially greater specification bias, whereas the repeat-sales model has less specification bias but more sample selection bias. Ideally, some combination of the two might lead to an improved procedure of delivering an index that reduces both sample selection and specification bias.

With this goal in mind, the present paper proposes a new methodology to construct real estate price indices that addresses some of the criticisms of the hedonic and repeat-sales methods. In our approach, the model is designed to control for hedonic information in a general way and pair sale prices at the individual building level, instead of the individual house level as is done in the repeat-sales method. This novel design offers four main advantages. First, the method makes use of all the real estate information in the sample, including both single-sale and repeat-sale homes. This approach contrasts with the use of just a small fraction of the sample that occurs in repeat-sales methods, thereby reducing both sample selection bias and information loss. With this design, the new real estate price index offers robustness against sample selection bias and gains in efficiency. Second, unlike standard hedonic models, a number of fixed effects are included in the framework to control for unobserved hedonic information and the functional form linking price and hedonic information is left unspecified. Both these features make the new index less susceptible to specification error than standard hedonic models. Third, the new model puts greater weight on pairs whose time gaps between sales are smaller, similar to repeat-sales methods; but since our pairs are constructed at the building level, the time gaps in our pairs are much smaller than those in pairs for repeat-sales methods. Consequently, pairs in our approach are typically more informative about price changes than those in repeat-sales methods. Finally, our model involves a simple and convenient GLS estimation procedure that is easy to implement and computationally efficient.

In triadic comparisons of out-of-sample predictions, the new index is found to give superior performance in predicting both repeat-sale home prices and single-sale home prices relative to the S&P/Case-Shiller index and the index constructed from a standard hedonic model. In dyadic comparisons, we find that the S&P/Case-Shiller index performs much better than the index from the hedonic model. These findings indicate that the specification bias in the standard hedonic method has more serious implications than the sample selection bias inherent in the S&P/Case-Shiller index, at least as far as the Singapore residential property market is concerned. When we test for explosive behavior in the three indices, we find evidence of earlier explosive behavior in our index than in the other indices. This finding has some important implications for macroprudential policy that are discussed in the paper.

The remainder of the paper is organized as follows. Section 2 develops the model and the estimation method. In Section 3, the method is applied to build a real estate price index for Singapore and out-of-sample performance of the alternative indices is compared. In Section 4 we test for explosive behavior in the index and the alternative indices using the recursive method of bubble detection developed recently in Phillips, Shi and Yu (2015a, 2015b). The results are discussed in the context of policy measures conducted by the Singapore government to cool the local real estate market. The Appendix 1 provides details of these policy cooling measures. Section 5 concludes. Throughout the paper we use the terminology 'house' to refer to an independent dwelling (apartment, flat, condominium, terraced, duplex, or free-standing) located within a specific building.

2.2 Model and estimation

Let the log price per square foot for the jth sale of the ith house in building p be $y_{i,j,p}$ and $t(i,j,p)$ be the time when the ith house in building p is sold for the jth time. The model design given below in (2.2.1) seeks to explain $y_{i,j,p}$ in terms of constituent components. In particular, we assume that the log price can be modeled as the sum of a log price index component, an unknown function of building level hedonic covariates, a location effect, an individual house effect, other individual house hedonic covariates, plus a partial sum of intervening building specific shocks, and a time-dependent error term. The log price index component is described by the parameter $\beta_{t(i,j,p)}$, which captures the time specific effect of house prices and is the primary parameter of interest. The building level hedonic information (whether observed or not) is denoted as Z_p; and an unknown function $f(Z_p)$ relates this building level information to the individual house price, capturing both observed and unobserved building level effects on price. The location effect is captured by a location variable μ_p, which is assumed to be a fixed effect with respect to the location of the building p, which may well be correlated with covariates. The individual house effect is captured by $h_{i,p}$, which is assumed to be independent over i with mean zero and variance σ_h^2. The building specific shocks at time t are described by the random variables $u_{t,p}$ which have mean zero and variance σ_u^2, and are assumed to be independent of each other across all buildings and for all time periods.

Suppose the total number of time periods (in quarters, say) is T. Then, $t(i,j,p)$

belongs to the set $\{1,\ldots,T\}$. When there is no confusion, we simply write $t(i,j,p)$ as t. Let L be the total number of buildings. Then the model is formulated as

$$y_{i,j,p} = \beta_{t(i,j,p)} + f(Z_p) + \gamma' X_{i,p} + \mu_p + \sum_{k=t(1,1,p)+1}^{t(i,j,p)} u_{k,p} + h_{i,p} + \varepsilon_{i,j,p}, \qquad (2.2.1)$$

where $X_{i,p}$ is the vector of covariates for the ith house in building p, f is a nonparametric function of Z_p, and $\varepsilon_{i,j,p}$ are idiosyncratic shocks that are assumed to be $iid(0,\sigma_\varepsilon^2)$. The covariates $X_{i,p}$ capture the available house level hedonic information (such as the floor number, the number of rooms, and so on) in the data.

The standard hedonic model (Ghysels et al., 2012) can be written as:

$$y_{i,j,z} = \mu_z + \beta_{t(i,j,z)} + \gamma' X_{i,z} + \varepsilon_{i,j,z}, \qquad (2.2.2)$$

where $y_{i,j,z}$ is the log price per square foot for the jth sale of the ith house in area z and $t(i,j,z)$ is the time when the ith house in area z is sold for the jth time. There are a few important differences between our model and the standard hedonic model which we now discuss.

There are still two restrictions implicit in model (2.2.2). First, a parametric form must be imposed to relate the observed building level covariates to the price. In model (2.2.2), a linear specification is adopted. However, any parametric specification is potentially invalid. Second, unobserved building level information cannot be accommodated in model (2.2.2). In the new model (2.2.1), building level hedonic information (Z_p) is included nonparametrically (whether observed or not). Furthermore, individual house fixed effects are not included in the standard hedonic model as they cannot be consistently estimated. In the new model, individual house fixed effects, $h_{i,p}$, are included.

Since (2.2.1) contains more detailed building-level information than (2.2.2) as well as a semiparametric specification, the new model is less susceptible to specification bias. To see this, note that housing heterogeneity arises both at the individual building level and the individual house level. To capture heterogeneity at the building level, it is necessary to include all the relevant hedonic information in (2.2.2). Inevitably some covariates will be omitted in (2.2.2) due to data unavailability and latent variable effects. These covariates are generally correlated with the observed covariates and are absorbed into the error term, $\varepsilon_{i,j,z}$, in (2.2.2). As a result, $\varepsilon_{i,j,z}$ is correlated with $X_{i,z}$ in (2.2.2). Whereas, in the new model, f is left unspecified and Z_p can include all relevant building level information, observed or unobserved, that is related to the house price. Hence, (2.2.2) suffers potential specification bias from missing heterogeneity at the building level and from the use of a particular functional form.

Focusing on houses that have sold more than once, the repeat-sales method of Case and Shiller (1987, 1989) is based on the following model

$$y_{i,j,z} - y_{i,j-1,z} = \beta_{t(i,j,z)} - \beta_{t(i,j-1,z)} + \sum_{k=t(i,j-1,z)+1}^{t(i,j,z)} u_{i,z}(k) + \varepsilon_{i,j,z} - \varepsilon_{i,j-1,z}. \quad (2.2.3)$$

where $u_{i,z}(k) \sim_{iid} N(0,\sigma_u^2)$ is the interval error at time $t(i,j-1,z)+k$ for house i in area z. So the partial sum $\sum_{k=t(i,j-1,z)+1}^{t(i,j,z)} u_{i,z}(k)$ is a Gaussian random walk and is used to model the concatenation of pricing shocks to this house between its $j-1$th and jth sale. Model (2.2.3) may be motivated from the specification

$$y_{i,j,z} = \beta_{t(i,j,z)} + f(X_{i,z}) + \mu_z + \sum_{k=0}^{a_{t(i,j,z)}} u_{i,z}(k) + \varepsilon_{i,j,z}, \quad (2.2.4)$$

where $a_{t(i,j,z)}$ is house age at time $t(i,j,z)$ for the ith house in area z. In this model, the functional form that captures the impact of hedonic information (whether it is observed or not) is f, which is left unspecified. For houses that have been sold multiple times in the sample, taking the difference of model (2.2.4) at two time stamps gives model (2.2.3) as both the hedonic covariates (both observed and unobserved) and the location effect are eliminated by differencing. Only houses that have been sold multiple times in the sample are retained in model (2.2.3). The model was estimated by Case and Shiller (1987, 1989) using a multi-stage method and led to the construction of the S&P/Case-Shiller real estate price index (S&P/Case-Shiller methodology report, 2009).

To facilitate estimation of our model, we take the average of equation (2.2.1) for all sales in the same building at each time period whenever there are sales. This yields

$$\bar{y}_{t,p} = \beta_t + f(Z_p) + \gamma' \bar{X}_{t,p} + \mu_{z(p)} + \sum_{k=t_1(p)+1}^{t} u_{k,p} + \bar{h}_{t,p} + \bar{\varepsilon}_{t,p}, \tag{2.2.5}$$

where $\bar{y}_{t,p}$ is the average price of all transaction prices in building p at time t and $t_1(p)$ is the time when the first sale in building p occurred. Similar to the Case-Shiller method, if there is another time period $t'(> t)$ when the most recent transactions occur in the same building p, we have model (2.2.5) at time t'. Taking the difference of model (2.2.5) at these two time periods, we obtain

$$\bar{y}_{t',p} - \bar{y}_{t,p} = \beta_{t'} - \beta_t + \gamma' \left(\bar{X}_{t',p} - \bar{X}_{t,p}\right) + \sum_{k=t+1}^{t'} u_{k,p} + \bar{h}_{t',p} - \bar{h}_{t,p} + \bar{\varepsilon}_{t',p} - \bar{\varepsilon}_{t,p}. \tag{2.2.6}$$

It is clear from Equation (2.2.6) that we create “pairs” at the building level at periods t and t', and then match the average building price at t' against that at t, after taking account of the hedonic information at the individual house level and a building specific random walk effect.

There are three advantages in our method relative to the repeat-sales method. First, since the repeat-sales method only uses data on repeat-sales, it is assumed that price change in repeat-sales are representative of the whole market. In our model, the full sample is used to construct the index, including both single-sales and repeat-sales. As a result, the approach does not suffer from sample selection bias. Second, given that the full sample has been used, there are consequential efficiency gains compared with the use of a subsample of data, as in the repeat-sales model. Third, the time gap between t and t' in our approach is calculated on a building basis whereas the time gap in the repeat-sales method is based on houses. As a result, the time gaps that appear in our approach are never bigger than and often much smaller than those in the repeat-sales method. Indeed, for a high percentage of cases, $t'-t=1$, as in the empirical application considered later in the paper. Since both methods put more weights on pairs whose time gap is smaller, the pairs in our method turn out to be more informative than those in the repeat-sales methods.

The specification used in our approach based on model (2.2.6) is more detailed and complex than that of the repeat-sales model (2.2.3). But estimation of the new model is accomplished in the same manner as the method of Case and Shiller (1987, 1989) and is therefore a simple procedure to implement. The details of the required calculations are as follows.

1. Run an OLS regression of model (2.2.6) to obtain initial estimates of β_t for all t and γ.

2. Plug these initial estimates into (2.2.6) to calculate the regression residuals, denoted by $\widehat{e}_{t',p}$, which are fitted values of the composite component $\sum_{k=t+1}^{t'} u_{k,p}+\bar{h}_{t',p}-\bar{h}_{t,p}+\bar{\varepsilon}_{t',p}-\bar{\varepsilon}_{t,p}$. Note that

$$E\left(\sum_{k=t+1}^{t'} u_{k,p}+\bar{h}_{t',p}-\bar{h}_{t,p}+\bar{\varepsilon}_{t',p}-\bar{\varepsilon}_{t,p}\right)=0,$$

and

$$\begin{aligned}Var\left(\sum_{k=t+1}^{t'} u_{k,p}+\bar{h}_{t',p}-\bar{h}_{t,p}+\bar{\varepsilon}_{t',p}-\bar{\varepsilon}_{t,p}\right) =&(t'-t)\sigma_u^2\\ &+\left(\frac{1}{n_{t',p}}+\frac{1}{n_{t,p}}\right)(\sigma_h^2+\sigma_\varepsilon^2),\end{aligned} \tag{2.2.7}$$

because the building specific shocks, individual house effects and error terms are all independent of each other. In (2.2.7) $n_{t,p}$ refers to the number of house sales transacted at time t in building p.

3. To calculate the weights to be used in GLS estimation, we run the following regression

$$\widehat{e}^{2}_{t',p} = c + (t'-t)\sigma_u^2 + \left(\frac{1}{n_{t',p}} + \frac{1}{n_{t,p}}\right)(\sigma_h^2 + \sigma_\varepsilon^2) + v_{t',p}, \tag{2.2.8}$$

   where $E(v_{t',p}) = 0$. Then the weights are the reciprocals of the fitted values from model (2.2.8). The diagonal matrix $\widehat{W}$ with weights appearing in the main diagonal is then the estimated weight matrix for GLS estimation.

4. Using $\widehat{W}$ as the weight matrix, GLS regression of (2.2.6) gives the final estimates of β_t for all t and γ. To be specific, we stack equation (2.2.6) into matrix form as

$$Y = Q\theta + e, \tag{2.2.9}$$

   where $\theta = [\ \beta' \quad \gamma'\ ]'$, β is a T-dimensional coefficient vector with elements β_t, Y is an N-dimensional vector with elements $\bar{y}_{t',p} - \bar{y}_{t,p}$, N is the number of pairs in the building level, and $Q = \begin{bmatrix} D & X \end{bmatrix}$, where D is a selection matrix designed to capture the differential components $\beta_{t'} - \beta_t$ in the model. The matrix D is constructed so that its nth row and tth column element has value -1, corresponding to the house price average in the previous period in the building level (viz., β_t) used at time t, and value 1 for the house price average in the current period in the building level (viz., $\beta_{t\prime}$) used at time t', and value 0 otherwise. In the partition of Q, X is a matrix with each row corresponding to elements of the form $\bar{X}_{t\prime,p} - \bar{X}_{t,p}$. GLS applied to (2.2.9) gives the estimate

$$\hat{\theta} = \left(\hat{\beta}', \hat{\gamma}'\right)' = (Q'\widehat{W}Q)^{-1}(Q'\widehat{W}Y),$$

   whose components are used to extract the house price index.

2.3 Empirical analysis

In this section, we apply the proposed model and the repeat-sales method to real estate price data involving quarterly transactions of private non-landed residential property sales in Singapore from Q1 1995 to Q2 2014. The period is of substantial interest given the fluctuations and growth in property prices in Singapore over this period and because of the extensive policy measures introduced by the government to cool the real estate market whose effectiveness can be gauged by empirical analysis of the real estate price indices.

There are mainly two residential property markets in Singapore: a private residential market and the public residential market that is managed by the Housing and Development Board (HDB). HDB is the statutory board of the Ministry of National Development and HDB flats are heavily subsidized by the Singapore government. Not surprisingly, the HDB market is largely segmented from the private residential market. Given its special nature and strong differentiation from the private market, we have excluded HDB transactions in the construction of the property market price index. The sample used for analysis therefore refers only to the private non-landed property market.[2]

The data source for private house information is the Urban Redevelopment Authority (URA),[3] which is Singapore's urban planning and management authority. The URA property market dataset provides extensive records of information for all transactions in the property market. The sale price (both the total price and the price per square foot) and the transaction period are reported. The district, sector and postal code of every transacted property are also recorded. Other characteristics include floor and unit number, project number, size, sale type, property type, completion year, leasehold tenure length, and location type.

During the sample period our data include some 315,000 transactions and the

[2] Non-landed residential property is the largest and most popular housing form in Singapore, constituting more than 75% of private residential units in the market by Q2 2014.

[3] http://www.ura.gov.sg/

Table 2.1: Summary Statistics of Single-Sale Houses in Singapore

Property Type	No. Houses	Mean	Sd	Min	Max
Apartments	40,097	1177	620	154	5146
Condominiums	106,073	947	459	156	6393
99 years tenure	81,086	939	446	154	5000
999 years tenure	6864	884	375	233	2695
Freehold	58,220	1125	600	202	6393
All	146,170	1010	519	154	6393

Table 2.2: Summary Statistics of Repeat-Sales Houses in Singapore

Property Type	No. Houses	Mean	Sd	Min	Max
Apartments	20,618	901	455	137	4700
Condominiums	49,715	850	404	94	4820
99 years tenure	33,554	864	366	94	4700
999 years tenure	4674	864	317	197	2491
Freehold	32,105	985	454	183	4820
All	70,333	865	420	94	4820

number of the dwellings involved is around 216,000.[4] Among these, about 146,000 houses are single-sales and the remainder, about 70,000 houses, are ones that sold more than once. The number of pairs for repeat-sales is around 97,000. So single-sales dominate repeat-sales in the sample in terms of the number of houses. In addition, the total number of buildings L is 4820[5], which leads to around 81,000 pairs at the building level.

There are two types of private non-landed residential properties in the Singapore real estate market: apartments and condominiums. The main difference between them is that condominiums are equipped with facilities but apartments may not be (Sing, 2001). The total number of condominium houses in our sample is around 155,000 and apartments account for some 60,000. In addition, in terms of ownership type, there are freehold, 999-year leasehold and 99-year leasehold. Most private residential properties transacted in the sample are either freehold or 99-year leasehold. Freehold houses are more expensive than 99-year leasehold houses. We

[4] We delete houses with incomplete information on characteristics. Sales that occur less than a quarter after the previous sale of the same house are also excluded.

[5] We delete buildings in which only one transaction occurs during the whole sample period. The number of buildings deleted is around 300, which implies only 300 single-sale houses are deleted. The loss of information is negligible given that we have around 146,000 single-sale houses in the dataset.

have postal district information in our database which is used to identify house location and zipcode information which is used to identify individual buildings.[6] Table 2.1 and Table 2.2 provide summary statistic information on the sample.

The dataset is well-suited to compare our new method with the standard hedonic method and the S&P/Case-Shiller repeat-sales method for index construction. First, we have the complete record of all transactions and the sample size of total sales is large, enabling us to estimate the proposed model accurately. With estimation error being small, attention can focus on comparing the indices constructed by different methods. Second, the hedonic information in the data is extensive so that many variables and alternative specifications can be included on the right hand side of models (2.2.2) and (2.2.1). Third, there are a very large number of repeat-sales in the data, so that model (2.2.3) can also be estimated accurately. Consequently, we can ignore estimation errors and focus on comparing the out-of-sample performance of different methods. By doing so, we can evaluate the relative magnitude of the price implications of implicit specification bias and sample selection bias in the three methods.

It is worth noting that single-sale properties display different summary statistics from repeat-sales properties. The mean price and the standard deviation for repeat-sale houses is lower than single-sale houses across all categories. This observation seems to support the argument that repeat-sale houses are not a representative random sample of the entire market and may carry a sample selection bias. Furthermore, in spite of the long sample period, about 68% of houses in the sample that have changed hands are single-sale houses. So the repeat-sale method is based on only about 32% of the houses in the sample.

The scatter plot of all house prices per square foot over time is given in Figure 2.1. It is difficult to discern price trends from this scatter plot, especially for houses at the low-end of the market because of the density of the data points. For high-end

[6] There are 27 postal districts and 69 postal sectors in the sample. In Singapore each building is assigned a unique zip code. This location and zipcode information is directly retrievable from the database.

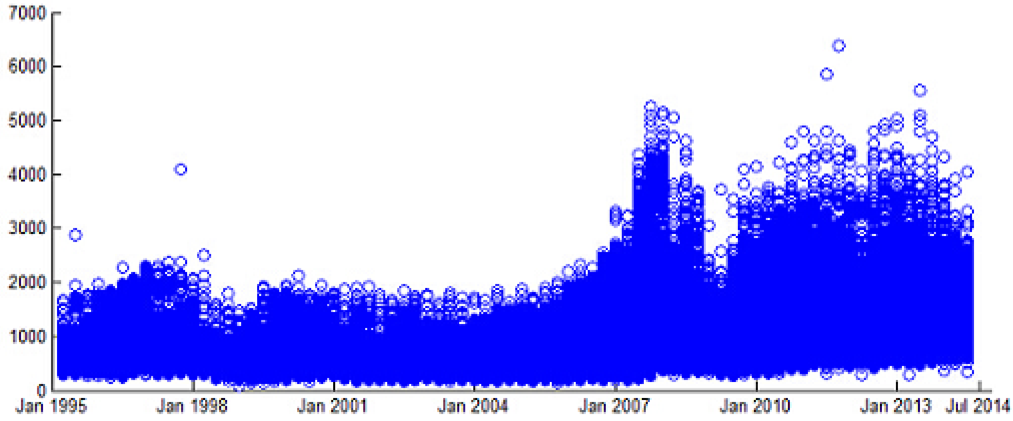

Figure 2.1: Scatter plots of house prices per square foot over January 1995 - June 2014

houses, at least, prices seem to be more stable between 2000 and 2006 than during other periods.

To fit the model in equation (2.2.6), we take account of the following two property characteristics: building zipcode and transaction period. Zipcode information in our database is used to identify buildings. The real estate price index is given by the parametric sequence $\{\beta_t\}$, which delivers the quarterly index from Q1 1995 to Q2 2014 (78 quarters in total). To keep our model as parsimonious as possible in this application, we do not use other hedonic covariates in our empirical analysis and hence the model has the form

$$\bar{y}_{t',p} - \bar{y}_{t,p} = \beta_{t'} - \beta_t + \sum_{k=t+1}^{t'} u_{k,p} + \bar{h}_{t',p} - \bar{h}_{t,p} + \bar{\varepsilon}_{t',p} - \bar{\varepsilon}_{t,p}. \tag{2.3.1}$$

The model can be easily expanded to include additional hedonic information as covariates. We have experimented with other covariates in our dataset and the main empirical findings reported here are qualitatively unchanged. So, for simplicity, we only report results obtained from the above specification.

We follow the estimation procedure described in Section 2 to obtain $\left\{\widehat{\beta}_t\right\}$. Since our purpose is to construct the house price index itself, rather than its logarithm, it is convenient to use the parameterization in Nagaraja, Brown and Zhao (2011) and calculate $\widehat{I}_t = \exp\left(\widehat{\beta}_t\right)$.[7] Finally, we take the first quarter in our sample as the

[7] Although $\widehat{I}_t$ is biased downward for I_t, the biased corrected estimator leads to virtually no change

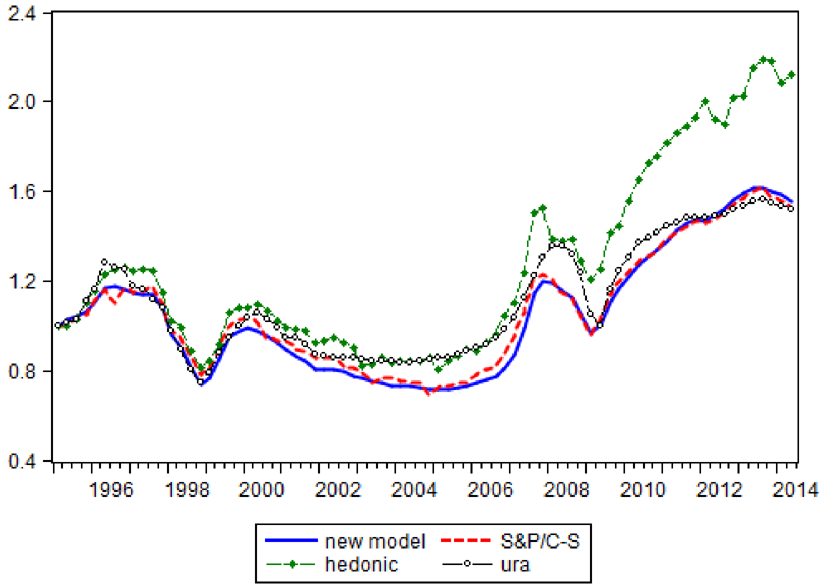

Figure 2.2: Four Real Estate Price Indices for Singapore: Q1 1995 – Q2 2014

reference point for which the price index is set to unity.

For comparison, we apply the hedonic method to all transaction prices and the S&P/Case-Shiller method to repeat-sales prices to build the indices.[8] We plot the proposed index, the S&P/Case-Shiller index, the standard hedonic index and the URA private non-landed residential property price index created by the Urban Redevelopment Authority (URA) in Figure 2.2.[9] As is apparent in the figure, there are some substantial discrepancies among the four indices. In particular, the standard hedonic index is more elevated and appears more volatile than the other indices and seems to diverge from the other indices towards the end of the sample period. This discrepancy may be due to the index's greater susceptibility to specification bias, a possibility that becomes clearer in the out-of-sample analysis below. Also, the URA index has different turning points from the other three indices. For example, over

in our results since the estimation error (and hence the variance estimate that appears in the bias calculation) is small.

[8] We employ the following four property characteristics in the hedonic model: location, transaction periods, property type, and ownership type to construct the hedonic index which is displayed in Figure 2.2. We have experimented with other covariates in our dataset and the main empirical findings reported here are qualitatively unchanged when additional covariates are included.

[9] Since the exact methodology of URA is not sufficiently clear for reproduction, we cannot include the URA index in our out-of-sample exercise.

the period of the global financial crisis, the turning point in the middle 2008 suggested by the URA index is two quarters later than that implied by the other three indices; and the turning point at the beginning of 2009 suggested by the URA index is one quarter later than that implied by the other three indices. Interestingly, our new index and the S&P/Case-Shiller index are very close to each other although our index suggests a longer trough in prices following the outbreak of SARS.

To compare the new index, the standard hedonic index and the S&P/Case-Shiller index and to examine the price implications of the specification bias and sample selection bias, we investigate the out-of-sample predictive power of the three indices.[10] To do so, we divide the observations into training and testing datasets. The testing set contains all the final sales of the houses sold three or more times in our sample period. Among the houses sold twice, their second transactions are randomly placed into the testing set with probability 0.04. We also randomly add the single-sale houses into our testing set with probability 0.24, so that the testing set contains the same number of single-sale houses and repeat-sale houses.[11] All the remaining houses are included in the training set. The resulting testing set contains around 15% of sales in our sample, of which 50% are single-sale houses and the rest are repeat-sales.

We first estimate all models based on the training set and then examine their out-of-sample predictive power on the testing set. Before analyzing the findings, we first explain how price predictions of the repeat-sale homes are obtained using the alternative indices. To calculate the predicted prices of the repeat-sale homes using the new method, we use

$$\hat{Y}_{t',i,p} = \frac{\widehat{I}^{bb}_{t'}}{\widehat{I}^{bb}_{t}} \bar{Y}_{t,p}, \tag{2.3.2}$$

[10] We evaluate the indices by their out-of-sample predictive power rather than their in-sample fitting because out-of-sample perfermance is more important in the context of specification testing. It is also well-known that that good in-sample fits often translate into poor out-of-sample predictions (for a recent discussion, see e.g. Hansen, 2010).

[11] To compare the out-of-sample predictive power of three indices on single sale houses, the test set does not include the single sale houses which are transacted as the first sales in their building. So the single sale houses, which are sold in the same period as the first sale in the building, are automatically included in the training set.

where $\hat{Y}_{t',i,p}$ is the price per square foot for house i in building p at time t', $\widehat{I}_t^{bb}$ is the estimated index from the new model at time t, t is the time period of the previous sale in building p, and $\bar{Y}_{t,p}$ is the average price per square foot for building p at time t in the training set.

For the S&P/Case-Shiller model, given that all single-sales are deleted, we use

$$\hat{Y}_{t',i} = \frac{\widehat{I}_{t'}^{cs}}{\widehat{I}_t^{cs}} Y_{t,i}, \qquad (2.3.3)$$

where $Y_{t,i}$ is the price per square foot for house i at time t, $t' > t$ and $\widehat{I}_t^{cs}$ is the estimated S&P/Case-Shiller index at time t, and t is the time period of the previous sale for house i (which is typically much smaller than t in equation (2.3.2)).

It should be pointed out that the predictive equations (2.3.2) and (2.3.3) are implied by models (2.3.1) and (2.2.3), respectively. From model (2.3.1), the predictive value of the average log price for building p at time t' can be represented as

$$\widehat{\bar{y}}_{t',p} = \bar{y}_{t,p} + \widehat{\beta}_{t'} - \widehat{\beta}_t.$$

When converting the log price to price, the predictive value of the average price for building p at time t' is

$$\widehat{\bar{Y}}_{t',p} = \exp\left\{\widehat{\bar{y}}_{t',p}\right\} = \exp\left\{\bar{y}_{t,p} + \widehat{\beta}_{t'} - \widehat{\beta}_t\right\} = \frac{\exp\left\{\widehat{\beta}_{t'}\right\}}{\exp\left\{\widehat{\beta}_t\right\}} \exp\left\{\bar{y}_{t,p}\right\} = \frac{\widehat{I}_{t'}^{bb}}{\widehat{I}_t^{bb}} \bar{Y}_{t,p}$$

where $\bar{Y}_{t,p}$ is the geometric mean price per square foot for building p at time t in the training set. We take this predictive value $\widehat{\bar{Y}}_{t',p}$ as the predictive value for house i in building p at time t', that is $\hat{Y}_{t',i,p}$. In a similar way, we can derive equation (2.3.3) from (2.2.3).

For the standard hedonic model, we plug the estimated parameters into model (2.2.2) to obtain

$$\widehat{y}_{i,j,z} = \widehat{\mu}_z + \widehat{\beta}_{t(i,j,z)} + \widehat{\gamma}' X_{i,z} \tag{2.3.4}$$

where $\widehat{y}_{i,j,z}$ is the predicted log price for the jth sale of house i in area z and $\widehat{\mu}_z$ is the estimated location dummy variable coefficient for area z. We then follow Nagaraja, Brown and Zhao (2011) to convert the log price into price by means of the transform

$$\widehat{Y}_{i,j,z} = \exp\left\{\widehat{y}_{i,j,z} + \frac{MSR}{2}\right\} \tag{2.3.5}$$

where $MSR = \frac{1}{M}\sum_{i=1}^{M}(y_{i,j,z} - \widehat{y}_{i,j,z})^2$, the mean square residuals and M is the total number of transactions to fit the model.

All three predictive prices are matched against the actual prices observed in the testing set. The root mean squared error (RMSE) and the mean absolute error (MAE) are reported in Table 2.3. Several important findings emerge. First, the S&P/Case-Shiller index performs much better than the standard hedonic index. In particular, compared with the standard hedonic method, the S&P/Case-Shiller method reduces the RMSE by around 40% and reduces the MAE by about 45%. In economic terms, the reduction in the MAE means that the repeat-sales method leads to a reduction of nearly \$100 (per square foot) in pricing error. This number compares to, as reported in Table 2, the mean price of all repeat-sales homes of \$865. Clearly the improvement is economically highly significant. These findings suggest that the sample selection bias present in the repeat-sales method is much less serious than the specification bias in the standard hedonic method, at least as far as house price prediction is concerned. Although we reported evidence that repeat-sales houses are not a representative random sample of the entire market in Singapore, the good out-of-sample performance of the S&P/Case-Shiller index suggests that perhaps the price changes in repeat-sales homes reflects well the changes that occur in single-sale homes.

Second and more importantly, the new model clearly has the best predictive power for repeat-sale homes. In particular, compared with the S&P/Case-Shiller,

Table 2.3: Testing set (with only repeat sales houses included): RMSE & MAE for the Indices (SG dollars)

Loss Function	new model	S&P/C-S	hedonic
RMSE	141	175	291
MAE	92	122	220

our model reduces the RMSE by around 19% and reduces the MAE by about 25%. Compared with the standard hedonic model, our model reduces the RMSE by around 52% and reduces the MAE by about 58%. In economic terms, these reductions in the MAE imply that the the new method leads to a pricing error reduction of $30 (per square foot) relative to the repeat-sale method and $128 (per square foot) relative to the standard hedonic method. All these reductions are substantial. At first glance, it may be surprising that the new model outperforms the repeat-sale method for predicting repeat-sales homes because the two indices are close to each other as shown in Figure 2.2. The superiority of the new method can be explained as follows. When we predict prices of repeat-sale homes, based on the specification of the new model, the average price of the most recent sales of all homes in the same building are used. However, based on the specification of the S&P/Case-Shiller model, one can only use the most recent sale price of the same home, which because of time lags may not reflect the present market as well. Indeed, the time gap in the latter case is usually much larger than the former case, making the most recent sale price of the same home far less relevant for prediction than the average price of the most recent sales of all homes in the same building.

Figures 2.3 and 2.4 plot the histograms for these two sets of time gaps and report the mean, median and standard deviation of the gap time in each case. Apparently, in the new method with probability of around 80% the gap time is 1 or 2 periods with median of 1 period and standard deviation of 2.75. In the repeat-sale method, the distribution of the gap time is much more dispersed with median of 15 periods and standard deviation of 15.48. The average price of all sale prices in the same buildings last quarter can be expected to be far more informative in predicting prices in the current period than the price of the same house 15 periods ago.

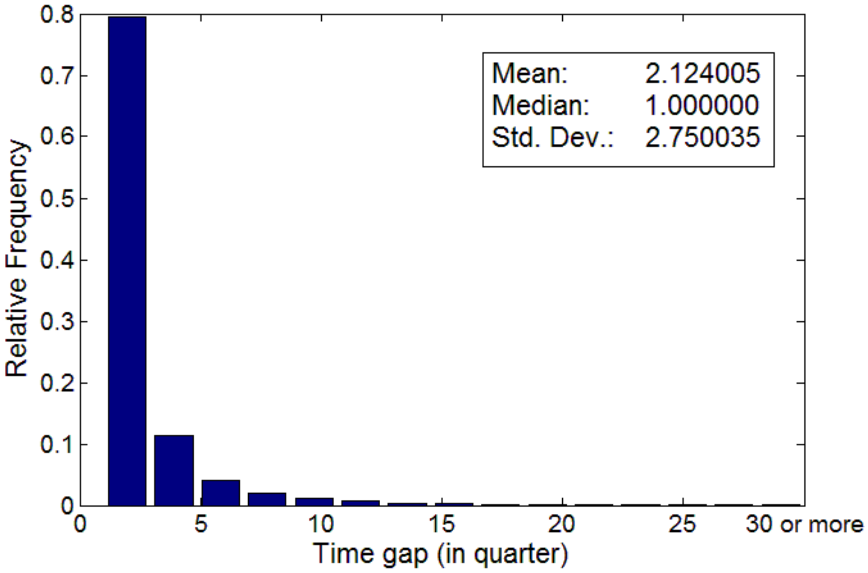

Figure 2.3: Histogram, mean, median and standard deviation of the time gap of sales in the same building.

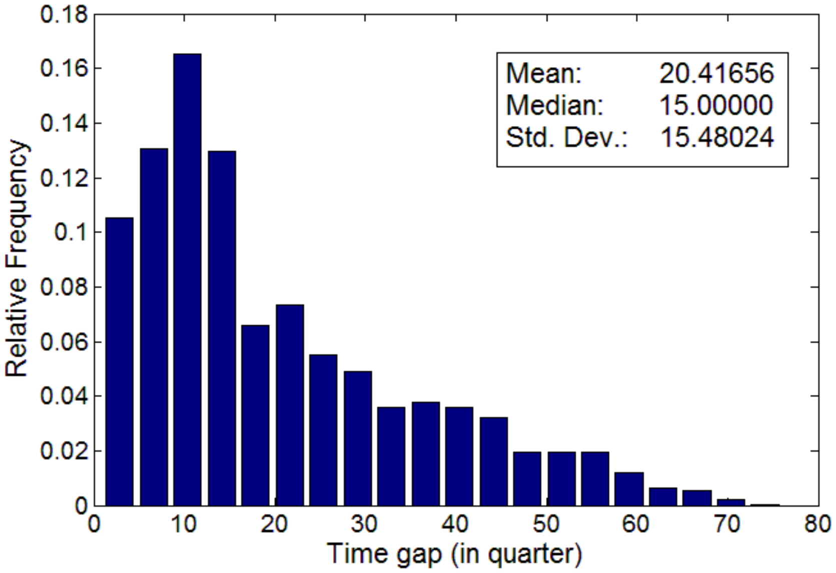

Figure 2.4: Histogram, mean, median and standard deviation of the time gap of sales of the same house.

Next we discuss how to predict prices of single-sale homes using the alternative indices. Since the S&P/Case-Shiller method discards all single-sale information, we cannot use this method to predict the price of single-sale homes. We therefore compare the predictive power of the new model with the standard hedonic model in this case. As before, we use equation (2.3.2) in our model and equation (2.3.4) and (2.3.5) in the hedonic model. The RMSE and MAE are shown in Table 2.4. Again, the new model performs much better in predicting prices of the single-sale homes than the standard hedonic model. Our model reduces the RMSE by around 48% and reduces the MAE by about 54%.

We can also compare the out-of-sample performance of our new model and the standard hedonic model on all houses in the testing set. The RMSE and the MAE are shown in Table 2.5. Our model reduces the RMSE by around 50% and reduces the MAE by about 56%.

Table 2.4: Testing set (only single sale houses included) RMSE & MAE for the Indices (SG dollars)

Loss Function	new model	hedonic
RMSE	156	297
MAE	86	188

Based on this out-of-sample analysis, it is clear that the standard hedonic model suffers from serious specification bias. Two sources of specification bias are expected. First, the attributes of houses or the factors that influence the house price are too many to be recorded in the data, leading to the problem of omission of relevant variables. Second, when covariates are observed, their exact relationship with the house price is almost always unknown and the use of a parametric form is potentially misspecified.

Moreover, the out-of-sample analysis also tells us that discarding single-sale houses from the analysis leads to a significant loss of information for prediction. This is because past prices of single-sale houses in the same building carry useful information. That explains why our new model increases the predictive power considerably relative to the S&P/Case-Shiller even though the two indices appear not to differ so much. To further illustrate this point, we consider a hypothetical (and infeasible) exercise, in which the single-sale houses are not eliminated from the prediction exercise and we predict the price in the testing set with our method and the repeat-sales method. With the repeat-sales method, we use the following fabricated equation (2.3.6) to calculate the predictive price

$$\hat{Y}_{t',i,p} = \frac{\widehat{I_{t'}^{cs}}}{\widehat{I_t^{cs}}} \bar{Y}_{t,p}, \tag{2.3.6}$$

where $\hat{Y}_{t',i,p}$ is the price per square foot for house i in building p at time t', $\widehat{I_t^{cs}}$ is the estimated S&P/Case-Shiller index at time t and t is the time period of the previous sale in building p, and $\bar{Y}_{t,p}$ is the average price per square foot in building p at time t in the training set. There are two main differences between equation (2.3.6) and (2.3.3). The first difference is that $\bar{Y}_{t,p}$ is used to estimate $\hat{Y}_{t',i,p}$ in (2.3.6) instead of $Y_{t,i}$ in (2.3.3). This allows us to predict prices of all houses in the testing set. Whereas (2.3.3) is only applicable to the repeat-sales houses. Secondly, t in (2.3.6) is the time period of the previous sale in building p whereas t in (2.3.3) is the time period of the previous sale of house i. As a result, for the same house i, the time period of the previous sale in building p is potentially much closer to t' than that of the previous sale of house i, even for repeat-sales homes. In our new model and the Case-Shiller model, more recent sales are informative due to the random walk component. Equation (2.3.6) is infeasible for prediction in the Case-Shiller model because the single-sale data have been removed by the S&P/Case-Shiller method. We do this hypothetical comparison only to explain the usefulness of the most recent sales in the same building for prediction.

The RMSE and the MAE from the two models are reported in Table 2.6 when we only predict prices of single-sale houses in the testing set. Tables 2.7, 2.8 give the results when only repeat-sale houses are predicted and all houses are predicted, respectively. By incorporating the information of the most recent sale prices in the same building, both the RMSE and MAE generated by the S&P/Case-Shiller index are substantially reduced. Consequently, although the predictive power of our new

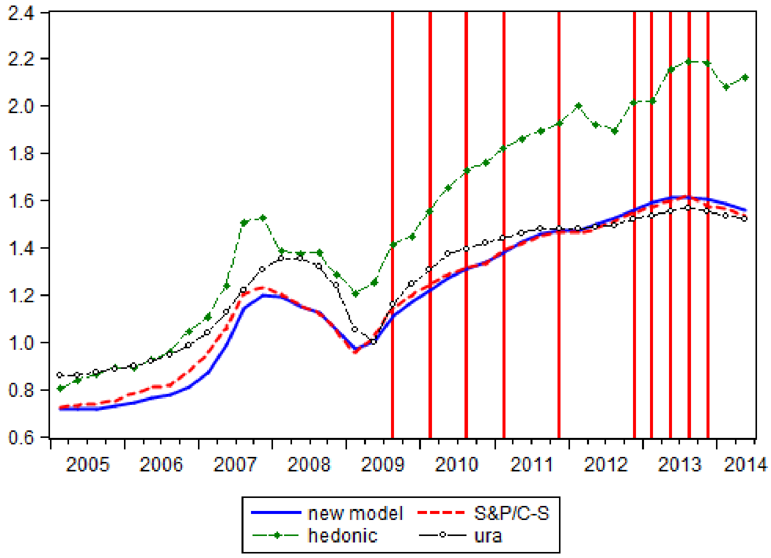

Figure 2.5: Four real estate price indices and the dates of ten rounds of successive macroprudential cooling measures (indicated by vertical lines).

model is still slightly better than the S&P/Case-Shiller model, the outperformance in this case (here evident in MAE) is only marginal because of the use of additional information (infeasibly) in the S&P/Case-Shiller index.

The out-of-sample analysis suggests that our new model captures the overall housing market situation in Singapore better than both the standard hedonic method and the repeat-sales method. As demonstrated before, our new method utilizes all the information, is robust to specification bias, and performs best in out-of-sample analysis. Moreover, the procedure is very convenient to implement in practical work.

2.4 Cooling measures and explosive behavior

Housing is a highly important sector of the economy and provides the largest form of savings of household wealth in Singapore. Property prices play an important role in consumer price inflation and can therefore have a serious impact on public policy. The private housing sector, property prices and rents also impact measures of Singapore's competitiveness in the world economy. For these and other reasons, the Sin-

Table 2.5: Testing set (all houses included) RMSE & MAE for the Indices (SG dollars)

Loss Function	new model	hedonic
RMSE	149	294
MAE	89	204

Table 2.6: The hypothetical exercise – Testing set (only single sale houses included) RMSE & MAE for the Indices (SG dollars)

Loss Function	new model	S&P/Case-Shiller
RMSE	156	156
MAE	86	87

Table 2.7: The hypothetical exercise – Testing set (only repeat sales houses included) RMSE & MAE for the Indices (SG dollars)

Loss Function	new model	S&P/Case-Shiller
RMSE	141	141
MAE	92	93

Table 2.8: The hypothetical excercise – Testing set (all houses included) RMSE & MAE for the Indices (SG dollars)

Loss Function	new model	S&P/Case-Shiller
RMSE	149	149
MAE	89	90

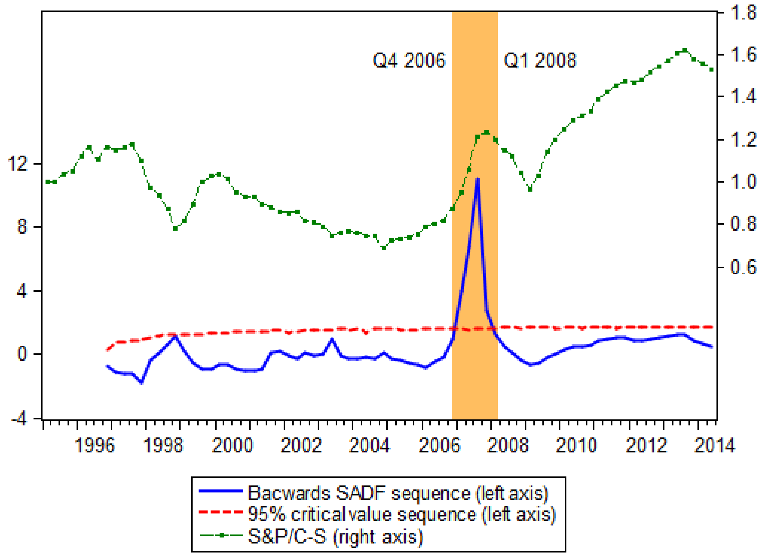

Figure 2.6: Testing for Bubbles in Singapore Real Estate Prices: using the S&P/Case-Shiller index, the BSADF statistic of PSY and the critical values.

gapore government has closely watched movements in housing prices over the last decade and particularly since the house price bubble in the USA. Recently, Singapore implemented ten successive rounds of macro-prudential measures intended to cool down the housing market. These measures were undertaken between September 2009 and December 2013, the first eight of which were targeted directly at the private residential market.

The Appendix 1 summarizes the dates and the nature of these macro-prudential measures. As is evident, a variety of macro-prudential policies have been used by the Singapore government. These include introducing a Seller's Stamp Duty (SSD), lowering the Loan-to-Value (LTV) limit, introducing an Additional Buyer's Stamp Duty (ABSD), and reducing the Total Debt Servicing Ratio (TDSR) and the Mortgage Servicing Ratio (MSR). To visualize the impact of these cooling measures on the dynamics of real estate price movements, Figure 2.5 plots the four price indices for the period between Q1 2008 and Q2 2014, superimposed by vertical lines indicating the introduction of these ten cooling measures.

The primary goal of the macro-prudential policies is to reduce or eliminate emergent price bubbles in the real estate market and bring prices closer in line with fun-

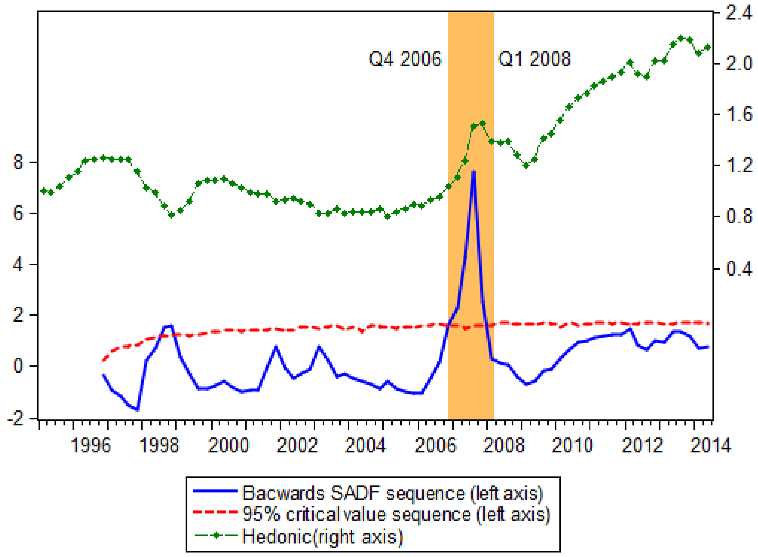

Figure 2.7: Testing for Bubbles in Singapore Real Estate Prices: using the index from the hedonic model, the BSADF statistic of PSY and the critical values.

damental values. Shi et al. (2014) and Mendicino and Punzi (2014) examined the impact of macro-prudential policies on real estate prices. Using the present value model, Diba and Grossman (1988) showed the presence of a rational bubble solution that implies that an explosive behavior in the observed price. If fundamental values are not explosive, then the explosive behavior in prices is a sufficient condition for the presence of bubble. Phillips, Wu and Yu (2011) and Phillips, Shi and Yu (2015a,2015b, PSY hereafter) introduced recursive and rolling window econometric methods to test for the presence of mildly explosive behavior or market exuberance in financial asset prices. These methods also facilitated estimation of the origination and termination dates of explosive bubble behavior. The method of Phillips, Wu and Yu (2011) is particularly effective when there is a single explosive episode in the data while the method of PSY can identify multiple explosive episodes. In the absence of prior knowledge concerning the number of explosive episodes, in what follows we use the PSY method to assess evidence of bubbles in real estate prices.

Bubble behavior and market exuberance and collapse are subsample phenomena. So, PSY proposed the use of rolling window recursive application of right sided unit root tests (against explosive alternatives) using a fitted model for data

$\{X_t\}_{t=1}^n$ of the following form

$$\Delta X_t = \hat{\alpha} + \hat{\beta} X_{t-1} + \sum_{i=1}^{K} \hat{\beta}_i \Delta X_{t-i} + \hat{e}_t. \tag{2.4.1}$$

Details of the procedure and its asymptotic properties are given in PSY. We provide a synopsis here and refer readers to PSY for further information about the specifics of implementation and the procedure properties. Briefly, the unit root test recursion involves a sequence of moving windows of data in the overall sample that expands backward from each observation $t = \lfloor nr \rfloor$ of interest, where n is the sample size and $\lfloor nr \rfloor$ denotes the integer part of nr for $r \in [0,1]$. Let r_1 and r_2 denote the start and end point fractions of the subsample regression. The resulting sequence of calculated unit root test statistics are denoted as $\left\{ADF_{r_1}^{r_2}\right\}_{r_1 \in [0, r_2 - r_0]}$ where r_0 is the minimum window size used in the recursion. and $t = \lfloor Tr \rfloor$ is the point in time for which we intend to test for normal market behavior against exuberance. PSY define the recursive statistic $BSADF_r = \sup_{r_1 \in [0, r_2 - r_0], r_2 = r} \left\{ADF_{r_1}^{r_2}\right\}$. The origination and termination dates of an explosive period are then determined from the crossing times

$$\hat{r}^e = \inf_{r \in [r_0, 1]} \{r : BSADF_r > cv\} \text{ and } \hat{r}^f = \inf_{r \in [\hat{r}^e, 1]} \{r : BSADF_r < cv\}, \tag{2.4.2}$$

where the recursive statistic $BSADF$ crosses its critical value cv. The quantity $\hat{r}^e$ estimates the origination date of an explosive period and $\hat{r}^f$ estimates the termination date of an explosive period. After the first explosive period is identified, the same method may be used to identify origination and termination dates of subsequent explosive episodes in the data.

To assess evidence for potential bubbles in the private real estate market in Singapore, we applied the PSY method first to both the S&P/Case-Shiller index and the index built from the hedonic model with minimum rolling window size $r_0 = 8$, corresponding to two years. Figures 2.6 and 2.7 report the two indices, the corresponding $BSADF$ statistics and the 5% critical values, respectively. The (orange) shaded area corresponds to the explosive period where the $BSADF$ statistic exceeds

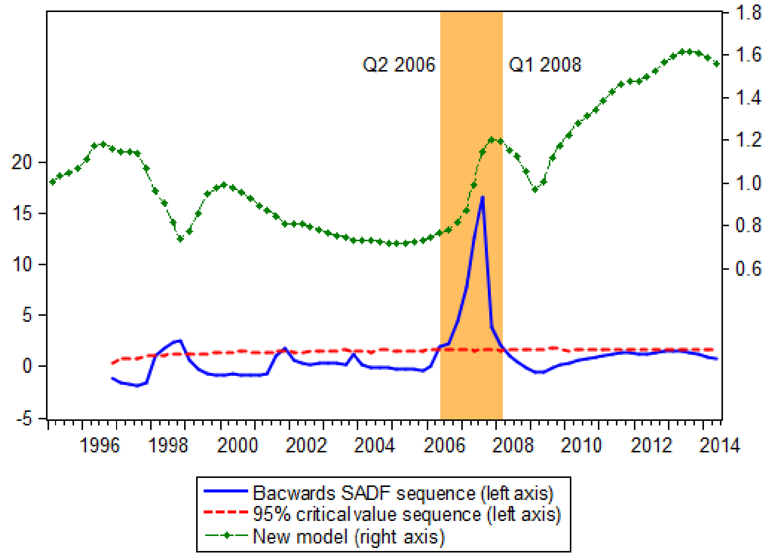

Figure 2.8: Testing for Bubbles in Singapore Real Estate Prices: using the new index, the BSADF statistic of PSY and the critical values.

the critical value. The PSY method identifies an explosive period, namely Q4 2006 to Q1 2008, in both the S&P/Case-Shiller index and the index built from the hedonic model.

We also applied the PSY method to our new index with minimum rolling window size $r_0 = 8$. Figure 2.8 reports the index, the test recursion, and the test 5% critical values. PSY identifies an explosive period in the private real estate market over Q2 2006 to Q1 2008. While the same conclusion date for the explosive period is found for the three indices, our new index suggests that explosive behavior commenced two quarters earlier, a finding that can have important practical implications for policy.

During the period 2006 - 2008, no cooling measures were introduced by the government. If the government had been alerted to the existence of exuberant market conditions in real time during this period, the opportunity would have been available for the implementation of cooling measures to affect the market. If the Case-Shiller index had been used, the government may have been stimulated to introduce cooling measures in Q4 2006, whereas if the new index were available and acted upon, the government may have introduced cooling measures earlier in Q2 2006. Moreover, although all three indices suggest that there were upward movements in price following 2008, between 2009 and 2013, these movements are not determined to be explosive and the PSY detector indicates little or no evidence of explosive behavior after 2009. This tapering in real estate market exuberance coincides with the period September 2009 through December 2013 during which macro-prudential cooling measures were actually implemented by the government and therefore appear to have been effective.

2.5 Conclusion

In order to exploit all available information in real estate markets, this paper provides a new methodology for the estimation of real estate price indices. The proposed new model has some of the advantages of the standard hedonic method as it uses both single-sales and repeat-sales data but it is less prone to specification bias than the standard hedonic model. Moreover, it generalizes the attractive feature of the repeat-sales method by creating sale pairs from within the individual building level, thereby increasing the number of observations used in the index. The model is also easy to estimate. Unlike the maximum likelihood methods of Hill, Knight and Sirmans (1997) and Nagaraja, Brown and Zhao (2011), this approach uses GLS estimation and is computationally efficient with large datasets. Other methods have been suggested to construct sale pairs in the literature – see, for example, McMillen (2012), and Guo et al (2014). Our matching rule is simpler to implement and has the advantage of a semiparametric nature.

We apply our estimation procedure to the real estate market for private residential dwellings in Singapore and examine the model's out-of-sample predictive performance in comparison with indices produced using the repeat-sales methodology of Case and Shiller (1987, 1989) and the standard hedonic method. The findings reveal that, compared with these alternative methodologies, our method has superior performance out-of-sample. We expect our method is well suited to build real estate indices for high density cities where houses are mainly project-based. Each project contains a number of buildings with many units. These units share essentially the same location, facility, design, developer ownership, and utilities, among other common features. In theory, our method can also be applied for single-family homes as long as we can define suitable groups (such as estates) for single-family homes and create sale pairs from the group level. Another useful idea is to use other simple criteria to choose pairs – see Baltagi and Li (2015), for instance, for the use of housing projects. These ideas will be investigated in the future work.

The recursive detection method of Phillips, Shi and Yu (2015a, 2015b) is applied to each of the indices to locate episodes of real estate price exuberance in Singapore. While for all three indices PSY identifies the same bubble, the bubble origination date in the new index comes two quarters earlier than that in the other two indices. Although all three indices grew during 2009 - 2013, the expansion is not explosive, indicating that the ten recent rounds of cooling measure intervention in the real estate market conducted by the Singapore government have been successful in controlling prices.

Chapter 3 On bias in the estimation of structural break points

3.1 Introduction

Statistical inference of structural breaks has received a great deal of attention in both econometrics and statistics literature over the last several decades. Bhattacharya (1994) provides a review of the statistics literature on the problem while Perron (2006) gives a review of the econometrics literature on the same problem. There are also several books devoted to this topic of research, including Csrgő and Horvth (1997), Chen and Gupta (2011). Both strands of the literature have addressed the problem in many aspects, from estimation, testing to computation, from frequentist's methods to Bayesian methods, from one structural break to multiple structural breaks, from univariate settings to multivariate settings. In addition to its statistical implications, the economic and financial implications of structural break problem have also been extensively studied; see, for example Hansen (2001) and Andreou and Ghysels (2009) for excellent reviews.

In terms of estimating structure break points, the literature has developed asymptotic theory for estimating the (fractional) structural break point (i.e. the (absolute) structural break point divided by the total sample size), including consistency, rates of convergence, and limit distributions; see, for example, Yao (1987) and Bai (1994). Interestingly and rather surprisingly, the finite sample theory for estimating structure break points seems to have received little attention in the literature. Is this lack of attention due to the good approximation of the asymptotic distribution to the finite sample distribution in empirically realistic cases and hence there is no need to study the finite sample theory? In particular, is there any bias in the traditional estimator of structural break points? Simulations provided in Yao (1987) seem to suggest that the asymptotic distribution is not necessarily close to the finite sample distribution while simulations provided in Bai (1994) seem to suggest there is little bias in the traditional estimator when the true break point is in the middle of the sample. Or is the lack of attention due to the difficulty in studying the finite sample theory and in approximating the bias, even in the first order?

This paper systematically investigates the finite sample properties and the bias problem in the estimation of structural break points. To the best of our knowledge, our study is the first systematic analysis of the finite sample issues in the literature. We develop the finite sample distribution of the maximum likelihood (ML) estimator of the structural break point in a continuous time model and relate the continuous time model to the discrete time models studied in the literature. We also document the bias both in the continuous time and the discrete time models, and propose an indirect estimation procedure to alleviate the bias via simulations.

Our study makes several contributions to the literature on structural breaks. First, we obtain the finite sample distribution of the ML and least squares (LS) estimators in some simple models and then obtain the bias from the finite sample distribution. It is shown that the bias can be substantial in the ML/LS estimators of the fractional structural break point and the absolute structural break point. The further the fractional structural break point is away from 50%, the more the bias. When the fractional structural break point is smaller (bigger) than 50%, the bias is positive (negative).

Second, we develop a novel approach to obtaining the finite sample distribution. Since the likelihood function and the sum of squared residuals are not differentiable with respect to break point in the discrete time models, the traditional approaches of obtaining the finite sample theory are not feasible. By using the Girsanov theorem, we obtain the likelihood function in a continuous time model with a structural break and then obtain the finite sample distribution of the ML estimator.

Third, we propose to do bias reduction using the indirect estimation procedure. One standard method for bias reduction is to obtain an analytical form to approximate the bias and then bias-correct the original estimator via the analytic approach as in Kendall (1954), Nickell (1981), Yu (2012) for various types of autoregressive models. However, it is difficult to use the analytic approach in this context as the bias formula is difficult to obtain. It is shown that the indirect estimation procedure, without knowing the analytical form to approximate the bias, achieves substantial bias reduction. However, since the binding function has a slope less than one, the variance of the indirect estimator is larger than that of the original estimator. The primary advantage of the indirect estimation procedure lies in its merit in calibrating the binding function via simulations and avoiding the need to obtain an analytic expression for the bias function. Since it is easy to simulate the model and estimate the break point parameter, the indirect estimation is a convenient method for reducing the bias in the estimation of the structural break points.

The rest of the paper is organized as follows. In Section 2, we first briefly review the literature and then develop a continuous time model with a structural break and discuss the finite sample properties of the ML estimator of the structural break point. Section 3 connects the continuous time model to the discrete time models previously considered in the literature. Section 4 introduces the indirect estimation technique and applies it to both the continuous time and the discrete time models with structural break points. In Section 5, Monte Carlo experiments are designed to obtain the bias of traditional estimators in models with structural breaks. We also compare the finite sample performance of the indirect estimation estimate with that of the traditional estimation methods. Sections 6 concludes.

3.2 Bias in a continuous time model

3.2.1 A literature review and motivations

The literature on estimating structural break points is extensive. A partial list of contributions in statistics include Chernoff and Zacks (1964), Hinkley (1969, 1970), Bhattacharya and Brockwell (1976), Ibragimov and Has'minskii (1981), Hawkins et al. (1986), Bhattacharya (1987), and Yao (1987). A key reference is Hinkley (1970) that develops not only the ML method for estimating the absolute break point but also its distributional behavior as the sample sizes before and after the change-point tend to infinity. In econometrics, Jushan Bai and Pierre Perron have made many contributions to the literature through their individual work as well as their collaborative work; see for example, Perron (1989), Bai (1994, 1995, 1997a, 1997b, 2010), Bai and Perron (1998) and Bai et al. (1998). For example, Bai (1994) extends the earlier literature by proposing the least squares (LS) method to estimate the break point in linear processes and develop its large sample theory. Bai and Perron (1998) uses the LS method to estimate linear models with multiple structural breaks.

A simplified model considered in Hinkley (1970) is

$$Y_t = \begin{cases} \mu + \varepsilon_t & \text{if } t \leq k_0 \\ (\mu+\delta)+\varepsilon_t & \text{if } t > k_0 \end{cases}, \tag{3.2.1}$$

where $t = 1, \ldots, T$, ε_t is a sequence of independent and identically distributed (i.i.d.) continuous random variables with zero mean, k_0 is the true value of the absolute structural break point k, constant μ measures the mean of Y_t before break and δ is the size of structural break. Let the probability density function (pdf) of Y_t be $f(Y_t, \mu)$ for $t \leq k_0$ and $f(Y_t, \mu+\delta)$ for $t > k_0$. And denote τ_0 the true value of the fractional structural break point τ, i.e., $\tau_0 = k_0/T$. Under the assumption that the form of function f and parameters μ and δ are all known and at least one observation comes from each distribution, the ML estimator of k_0 is defined as

$$\widehat{k}_{ML} = \arg \max_{k=1,\ldots,T-1} \left\{ \sum_{t=1}^{k} \log f(Y_t, \mu) + \sum_{t=k+1}^{T} \log f(Y_t, \mu + \delta) \right\}. \tag{3.2.2}$$

The corresponding estimator of τ is $\widehat{\tau}_{ML} = \widehat{k}_{ML}/T$. Hinkley (1970) showed that $\widehat{k}_{ML} - k_0$ converges in distribution as the sample sizes before and after the break point tend to infinity. He also pointed out that the distribution of $\widehat{k}_{\infty} - k_0$, where $\widehat{k}_{\infty}$ denotes $\widehat{k}_{ML}$ with infinite sample, has no closed-form expression, and gave a numerical method to compute the distribution. However, this numerical scheme is difficult to handle for small δ since the distribution becomes rather dispersive when δ is small. This difficulty motivates Yao (1987) to develop a limit theory as $\delta \to 0$.

Letting $\delta \to 0$, Yao (1987) derived a sequential limit distribution as

$$\delta^2 I(\mu) \left(\widehat{k}_{\infty} - k_0 \right) \xrightarrow{d} \arg \max_{u \in (-\infty, \infty)} \left\{ W(u) - \frac{1}{2}|u| \right\}, \tag{3.2.3}$$

where $I(\mu)$ is the Fisher information of the density function $f(y, \mu)$, $W(u)$ is a two-sided Brownian motion which will be defined below, and $\xrightarrow{d}$ denotes convergence in distribution. Since $I(\mu)$ depends on the error's distribution, no invariance principle applies to the sequential limit distribution. Yao (1987) also derived the pdf of the sequential limit distribution as

$$g(x) = 1.5e^{|x|}\Phi\left(-1.5|x|^{0.5}\right) - 0.5\Phi\left(-0.5|x|^{0.5}\right),$$

and its cdf as

$$G(x) = 1 + \sqrt{\frac{x}{2\pi}} e^{-x/8} - (x+5)\Phi\left(-0.5\sqrt{x}\right)/2 + 1.5e^{x}\Phi\left(-1.5\sqrt{x}\right), \text{ for } x > 0,$$

$G(x) = 1 - G(-x)$ if $x \leq 0$, where $\Phi(x)$ is the cdf of the standard normal distribution.

For the same model as in Equation (3.2.1), Hawkins et al. (1986) and Bai (1994) studied the LS estimators of k and τ with unknown μ and δ. The LS estimator of k takes the form of

$$\widehat{k}_{LS} = \arg \min_{k=1,\ldots,T-1} S_k^2 = \arg \max_{k=1,\ldots,T-1} V_k^2, \tag{3.2.4}$$

where $S_k^2 = \sum_{t=1}^{k} \left(Y_t - \overline{Y}_k\right)^2 + \sum_{t=k+1}^{T} \left(Y_t - \overline{Y}_k^*\right)^2$ with $\overline{Y}_k$ ($\overline{Y}_k^*$) being the sample mean of the first k (last $T-k$) observations and $V_k^2 = \frac{T(T-k)}{T^2}\left(\overline{Y}_k^* - \overline{Y}_k\right)^2$. The corresponding estimator of τ is $\widehat{\tau}_{LS} = \widehat{k}_{LS}/T$. Hawkins et al. (1986) showed that $T^{\alpha}\left(\widehat{\tau}_{LS} - \tau_0\right) \xrightarrow{p} 0$ for any $\alpha < 1/2$. Bai (1994) improved the rate of convergence by showing that $\widehat{\tau}_{LS} - \tau_0 = O_p\left(\frac{1}{T\delta^2}\right)$. This convergence rate also applies to $\widehat{\tau}_{ML}$ when ε_t is an i.i.d. Gaussian sequence. Because, in the case where $\varepsilon_t \sim$ i.i.d.$N(0,\sigma^2)$, the LS estimator is equivalent to the ML estimator with unknown μ and δ, whose limit theory, as argued in Hinkley (1970), is the same as that of the ML estimator when μ and δ are known as long as μ and δ can be consistently estimated. While $\widehat{\tau}_{LS}$ is consistent, $\widehat{k}_{LS}$ is inconsistent since $\widehat{k}_{LS} - k_0 = O_p\left(\frac{1}{\delta^2}\right)$.

To develop the limit distribution with an invariance principle, δ has to go to zero as $T \to \infty$, as shown in Bai (1994). This kind of limit theory is particularly useful in constructing confidence interval when the size of the break is small. Let δ_T be the size of break that depends on T. Bai showed that if $\varepsilon_t \sim$i.i.d.$(0,\sigma^2)$, $\delta_T \to 0$ and $\frac{\sqrt{T}\delta_T}{\sqrt{\log T}} \to \infty$ as $T \to \infty$,

$$T\left(\delta_T/\sigma\right)^2\left(\hat{\tau}_{LS} - \tau_0\right) \xrightarrow{d} \arg \max_{u\in(-\infty,\infty)} \left\{W(u) - \frac{1}{2}|u|\right\}. \tag{3.2.5}$$

When ε_t is normally distributed, the Fisher information $I(\mu)$ turns out to be σ^{-2}. Therefore, the simultaneous asymptotic distribution in Bai (1994) is the same as the sequential asymptotic distribution in Yao (1987). Bai (1994) also derived the limit distribution when ε_t is a short memory ARMA process, which is the same as shown in Equation (3.2.5) by replacing σ^2 with the long-run variance of ε_t. To obtain the limit distribution, Bai (1994) examined the behavior of normalized objective function in the small neighborhood of the true break point k_0 such that $k = [k_0 + v(\delta_T)^{-2}]$

where v varies in a bounded interval. This is equivalent to the local asymptotic theory of Le Cam (1960).

A study which is closest to ours is Ibragimov and Has'minskii (1981). Ibragimov and Has'minskii analyzed a simple continuous time model

$$dX(t) = \frac{1}{\varepsilon} S(t - \tau_0) dt + dB(t) \tag{3.2.6}$$

where $t \in [0, 1]$, $S(t - \tau_0)$ is a non-stochastic drift term with discontinuity at time τ_0 (i.e. τ_0 is the structural break point), and ε is a small parameter. Let $\lim_{x \to 0+} S(x) - \lim_{x \to 0-} S(x) = \delta$ denote the size of the break. Following the development of the local asymptotic theory of Le Cam, Ibragimov and Has'minskii (1981), under the assumption that a continuous record is available, examined the behavior of the normalized likelihood ratio in the small neighborhood of the true break point τ_0 such that $\tau = \tau_0 + \varepsilon^2 u$ and showed that as $\varepsilon \to 0$,

$$\delta^2 (\hat{\tau}_{ML} - \tau_0) \xrightarrow{d} \arg \max_{u \in (-\infty, \infty)} \left\{ W(u) - \frac{1}{2} |u| \right\}. \tag{3.2.7}$$

Figure 3.1 plots the pdf of the limit distribution given in Yao (1987), Bai (1994), and Ibragimov and Has'minskii (1981). For the purpose of comparison, we also plot the pdf of the standard normal distribution. It can be seen that both distributions are symmetric, suggesting no bias in the limit distribution when estimating the fractional break point using ML/LS. However, relative to the standard normal distribution, the limit distribution has much fatter tails and a much higher peak. The symmetrical property is a result of using the local asymptotic approach to develop the limit distribution in all cases. This property does not help us to understand the finite sample bias in estimating the break points.

The asymptotic arguments above do not take account of asymmetry in the sample before and after the break. To capture the influence of asymmetric information before and after the break, a continuous time model is a natural choice. As long as $\tau_0 \neq 1/2$, the information contained by observations over the time interval $[0, \tau_0]$ and

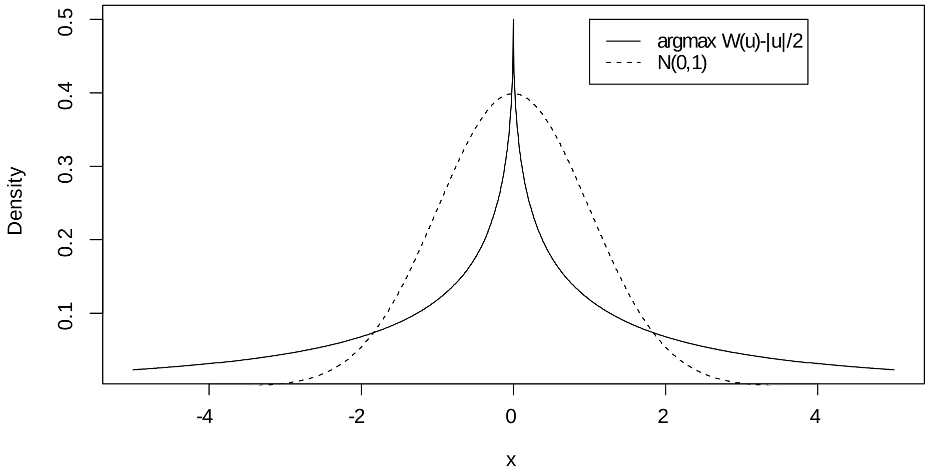

Figure 3.1: The pdfs of $\arg\max_{u\in(-\infty,\infty)} \left\{W(u)-\frac{1}{2}|u|\right\}$ and a standard norm distribution

those over the time interval $[\tau_0, 1]$ are different, even if the continuous records are available and there are infinite number of observations before and after the break. This is because in continuous time models the time span also conveys useful information.

There is another motivation to consider a continuous time model. To study the finite sample bias for traditional estimators, a typical approach is to consider the first order condition to an extremum problem that defines the associated estimator; see for example, Rilstone et al. (1996) and Bao and Ullah (2007). While this approach covers many popular models, it is not applicable to the problem of estimating the structural break point in discrete time models, regardless if ML or LS is used. This is because the objective functions in (3.2.2) and (3.2.4) are not differentiable and hence no first order condition is available for developing high order expansions. Using continuous time models we can avoid this difficulty.

3.2.2 A continuous time model

The continuous time model considered in the paper is

$$dX(t) = S(t-\tau_0)dt + \sigma dB(t), \tag{3.2.8}$$

with $t \in [0,1]$, where $S(t-\tau_0)$ is a non-stochastic drift term with discontinuity at time τ_0. Let $\lim_{x\to 0+} S(x) - \lim_{x\to 0-} S(x) = \delta$ denote the size of the break. Different from Model (3.2.6) studied in Ibragimov and Has'minskii (1981), we let $\varepsilon = 1$, not $\varepsilon \to 0$. In addition, we have σ in the diffusion function, capturing the noise level. Hence the signal-to-noise ratio in our model is δ/σ, which is a constant, unlike what was assumed in Ibragimov and Has'minskii (1981).

Furthermore, in order to establish a link to the discrete time model in Yao (1987) and Bai (1994), we consider the case in which $S(t-\tau_0)$ only takes two values such that

$$S(t-\tau_0) = \begin{cases} \mu & if\ t \le \tau_0 \\ \mu+\delta & if\ t > \tau_0 \end{cases}, \tag{3.2.9}$$

with $t \in [0,1]$, where τ_0 is the unknown true break point, and $\tau_0 \in [\alpha,\beta]$ with $0 < \alpha < \beta < 1$. Consequently Model (4.2.12) can be rewritten as

$$dX(t) = (\mu + \delta 1_{[t>\tau_0]})dt + \sigma dB(t). \tag{3.2.10}$$

Following Le Cam (1960) and Ibragimov and Has'minskii (1981), we obtain the exact log-likelihood ratio of Model (3.2.10) via the Girsanov Theorem[1]

$$\log\left(\frac{dP_\tau}{dP_{\tau_0}}\right) = \int_0^1 \frac{\delta}{\sigma}\left(1_{[t>\tau]} - 1_{[t>\tau_0]}\right) dB(t) - \frac{1}{2}\int_0^1 \frac{\delta^2}{\sigma^2}\left(1_{[t>\tau]} - 1_{[t>\tau_0]}\right)^2 dt.$$

The ML estimator of τ_0 is

$$\widehat{\tau}_{ML} = \arg\max_{\tau\in(0,1)} \log\left(\frac{dP_\tau}{dP_{\tau_0}}\right).$$

[1] See also Phillips and Yu (2009) for a recent usage of the Girsanov Theorem in estimating continuous time models.

When $\tau \leq \tau_0$, we have

$$\begin{aligned}\log\left(\frac{dP_\tau}{dP_{\tau_0}}\right) &= \frac{\delta}{\sigma}\int_0^1 1_{[\tau<t\leq\tau_0]}dB(t) - \frac{\delta^2}{2\sigma^2}\int_0^1 1_{[\tau<t\leq\tau_0]}dt \\ &= \frac{\delta}{\sigma}\int_\tau^{\tau_0} dB(t) - \frac{\delta^2}{2\sigma^2}\int_\tau^{\tau_0} dt \\ &= \frac{\delta}{\sigma}(B(\tau_0)-B(\tau)) - \frac{\delta^2}{2\sigma^2}(\tau_0-\tau).\end{aligned}$$

When $\tau > \tau_0$, we have

$$\begin{aligned}\log\left(\frac{dP_\tau}{dP_{\tau_0}}\right) &= -\frac{\delta}{\sigma}\int_0^1 1_{[\tau_0<t\leq\tau]}dB(t) - \frac{\delta^2}{2\sigma^2}\int_0^1 1_{[\tau_0<t\leq\tau]}dt \\ &= -\frac{\delta}{\sigma}\int_{\tau_0}^{\tau} dB(t) - \frac{\delta^2}{2\sigma^2}\int_{\tau_0}^{\tau} dt \\ &= \frac{\delta}{\sigma}(B(\tau_0)-B(\tau)) - \frac{\delta^2}{2\sigma^2}(\tau-\tau_0).\end{aligned}$$

Thus, we can write the exact log-likelihood ratio as

$$\log\left(\frac{dP_\tau}{dP_{\tau_0}}\right) = \frac{\delta}{\sigma}(B(\tau_0)-B(\tau)) - \frac{\delta^2}{2\sigma^2}|\tau-\tau_0|. \tag{3.2.11}$$

This implies that the ML estimator of τ_0 is

$$\widehat{\tau}_{ML} = \arg\max_{\tau\in(0,1)}\left\{\frac{\delta}{\sigma}(B(\tau_0)-B(\tau)) - \frac{\delta^2}{2\sigma^2}|\tau-\tau_0|\right\}, \tag{3.2.12}$$

which leads to

$$\widehat{\tau}_{ML} - \tau_0 = \arg\max_{u\in(-\tau_0,1-\tau_0)}\left\{\frac{\delta}{\sigma}(B(\tau_0)-B(\tau_0+u)) - \frac{\delta^2}{2\sigma^2}|u|\right\}.$$

We now define a two-sided Brownian motion as

$$W(u) = \begin{cases} W_1(-u) = B(\tau_0) - B(\tau_0-(-u)) & \text{if } u \leq 0 \\ W_2(u) = B(\tau_0) - B(\tau_0+u) & \text{if } u > 0 \end{cases}, \tag{3.2.13}$$

where $W_1(s) = B(\tau_0) - B(\tau_0 - s)$ and $W_2(s) = B(\tau_0) - B(\tau_0 + s)$ are two independent Brownian motions as they are composed by increments of the Brownian motion $B(\cdot)$ before and after the time τ_0 respectively with $W_1(0) = W_2(0) = 0$.

We then have

$$\begin{aligned}
\widehat{\tau}_{ML} - \tau_0 &= \arg \max_{u \in (-\tau_0, 1-\tau_0)} \left\{ \frac{\delta}{\sigma} W(u) - \frac{\delta^2}{2\sigma^2} |u| \right\} \\
&\overset{d}{=} \arg \max_{u \in (-\tau_0, 1-\tau_0)} \left\{ W\left(u \left(\frac{\delta}{\sigma} \right)^2 \right) - \frac{1}{2} \left| u \left(\frac{\delta}{\sigma} \right)^2 \right| \right\} \\
&\overset{d}{=} \left(\frac{\delta}{\sigma} \right)^{-2} \arg \max_{u \in \left(-\tau_0 \left(\frac{\delta}{\sigma} \right)^2, (1-\tau_0) \left(\frac{\delta}{\sigma} \right)^2 \right)} \left\{ W(u) - \frac{|u|}{2} \right\},
\end{aligned}$$

where $\overset{d}{=}$ denotes equivalence in distribution. Consequently, we obtain

$$\left(\frac{\delta}{\sigma} \right)^2 (\widehat{\tau}_{ML} - \tau_0) \overset{d}{=} \arg \max_{u \in \left(-\tau_0 \left(\frac{\delta}{\sigma} \right)^2, (1-\tau_0) \left(\frac{\delta}{\sigma} \right)^2 \right)} \left\{ W(u) - \frac{1}{2} |u| \right\}, \tag{3.2.14}$$

the exact distribution of the ML estimator $\widehat{\tau}_{ML}$ with a continuous record being available, which is also called in this paper the exact finite sample distribution of $\widehat{\tau}_{ML}$ in the sense that it is obtained with a finite time span before and after the break, which is $(0, \tau_0]$ and $[\tau_0, 1)$ respectively.

It seems that the finite sample distribution given in Equation (3.2.14) is similar to the limit distributions given in Yao (1987), Bai (1994) and Ibragimov and Has'minskii (1981) listed in (3.2.3), (3.2.5) and (3.2.7), respectively. However, there is one critical difference between them. The limit distributions in (3.2.3), (3.2.5) and (3.2.7) correspond to the location of the extremum of $W(u) - \frac{1}{2}|u|$ over the interval of $(-\infty, \infty)$. Since the interval is symmetric about zero, the limit distribution is symmetric about zero. However, the finite sample distribution in (3.2.14) corresponds to the location of the extremum of $W(u) - \frac{1}{2}|u|$ over the interval of $\left[-\tau_0 \left(\frac{\delta}{\sigma} \right)^2, (1-\tau_0) \left(\frac{\delta}{\sigma} \right)^2 \right]$, therefore depends on the true value of break point τ_0. Only when τ_0 is 50%, that is the true break point is exactly at the middle, $\left[-\tau_0 \left(\frac{\delta}{\sigma} \right)^2, (1-\tau_0) \left(\frac{\delta}{\sigma} \right)^2 \right]$ becomes $\left(\frac{\delta}{\sigma} \right)^2 [-50\%, 50\%]$ and symmetric about zero.

In this case the finite sample distribution will be symmetric about zero. If τ_0 is not 50% (either smaller or bigger than 50%), the interval and hence the finite sample distribution will be asymmetric. It is easy to see that the finite sample distribution in (3.2.14) suggests upward bias when $\tau_0 < 1/2$ and downward bias when $\tau_0 > 1/2$, and the further τ_0 away from $1/2$, the larger the bias.

Because of this difference in the interval to locate the extremum, we cannot obtain the pdf or cdf of the finite sample distribution in closed-form. As a result, we obtain the pdf by simulations as for the case of the Dickey-Fuller distributions.

Figure 3.2 plots the densities of $\widehat{\tau}_{ML}$ given in Equation (3.2.14) when $\tau_0 = 0.4, 0.5, 0.6$ (the left, middle and right panel respectively) and the signal-to-noise ratio (δ/σ) is 1. Figure 3.3 - Figure 3.6 plots the densities of $\widehat{\tau}_{ML} - \tau_0$ when the signal-to-noise ratio is 2, 4, 6, 8. There are several interesting observations from these plots. First and most importantly, when $\tau_0 = 50\%$, the densities of $\widehat{\tau}_{ML} - \tau_0$ is always symmetric about zero, no matter what value the signal-to-noise ratio takes. As a result, there is no finite sample bias in this case. However, when τ_0 is not 50%, the density is not symmetric any more. In particular, if τ_0 is less (larger) than 50%, the density is positively (negatively) skewed and there is a upward (downward) bias in $\widehat{\tau}_{ML}$. The smaller the signal-to-noise ratio, the larger the bias. The further τ_0 away from 50%, the larger the bias, although this feature does not show up in the graphs.

Second, there are tri-modality in the finite sample distribution when the signal-to-noise ratio is low (for example when $\delta/\sigma = 1, 2, 4$). The true value is one of the three modes while the two boundary points (0 and 1) are the other two modes. For very small signal-to-noise ratio, for example $\delta/\sigma = 1$, the highest mode is not the true value, but the two boundary points when $\tau_0 = 50\%$; it becomes the left (right) boundary point if τ_0 is smaller (larger) than 50%. However, the highest mode moves to the true value when the signal-to-noise ratio increases in all cases with $\delta/\sigma \geq 2$. It is also found that, the mode on the left boundary point is always larger (smaller) than that on the right boundary point when τ_0 is smaller (larger) than 50%. When

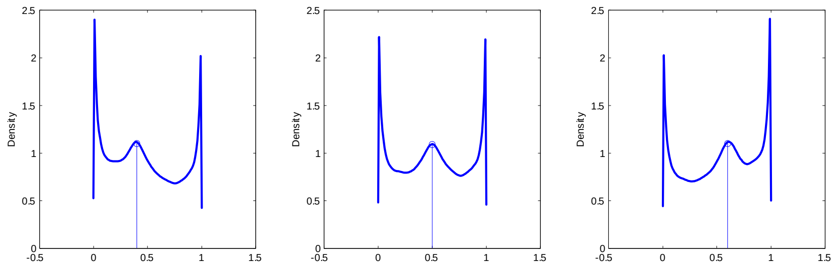

Figure 3.2: The density of $\widehat{\tau}_{ML}$ given in Equation (3.2.14) when $\tau_0 = 0.4, 0.5, 0.6$ (the left, middle and right panel respectively) and the signal-to-noise ratio (δ/σ) is 1. In each panel, the verticle line represents the true value of τ_0.

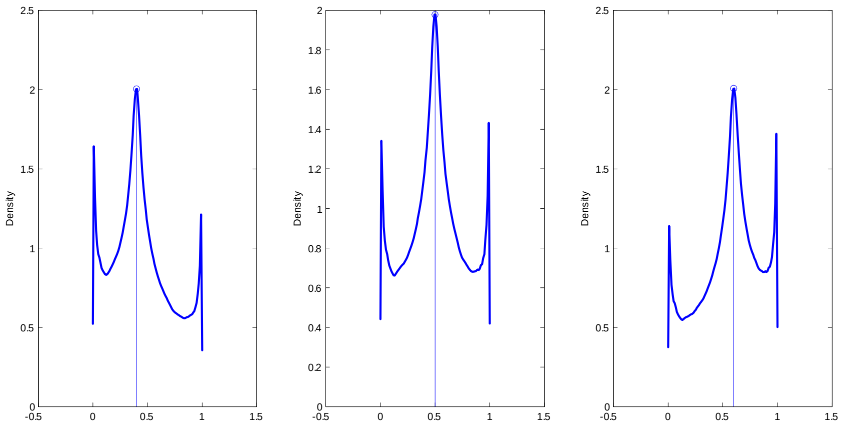

Figure 3.3: The density of $\widehat{\tau}_{ML}$ given in Equation (3.2.14) when $\tau_0 = 0.4, 0.5, 0.6$ (the left, middle and right panel respectively) and the signal-to-noise ratio (δ/σ) is 2. In each panel, the verticle line represents the true value of τ_0.

the signal-to-noise ratio is large enough, tri-modaility becomes unique modality. In this case, the shape of the distribution is similar to that in Figure 1 but is more peaked at the mode.

3.3 Bias in a discrete time model

As reviewed in Section 2, Hinkley (1970), Yao (1987) and Bai (1994) examined the change-in-mean model in the discrete time context.[2] Since the objective functions are not differentiable with respect to k, it is very difficult to obtain the finite sample distribution in the discrete time model. Yao (1987) and Bai (1994) developed

[2] In Bai (1994), ε_t can be a linear process satisfying the summability condition. So Bai's model is more general than Hinkley (1970) and Yao (1987)

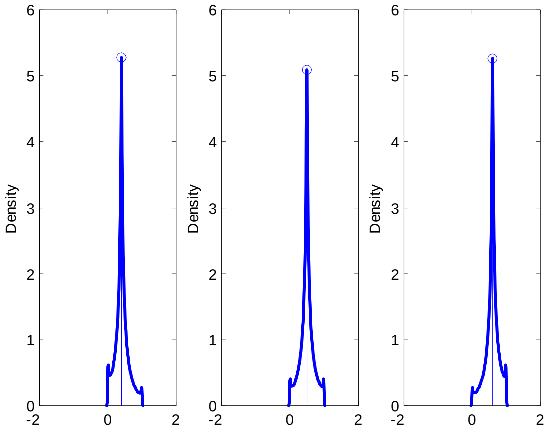

Figure 3.4: The density of $\widehat{\tau}_{ML}$ given in Equation (3.2.14) when $\tau_0 = 0.4, 0.5, 0.6$ (the left, middle and right panel respectively) and the signal-to-noise ratio (δ/σ) is 4. In each panel, the verticle line represents the true value of τ_0.

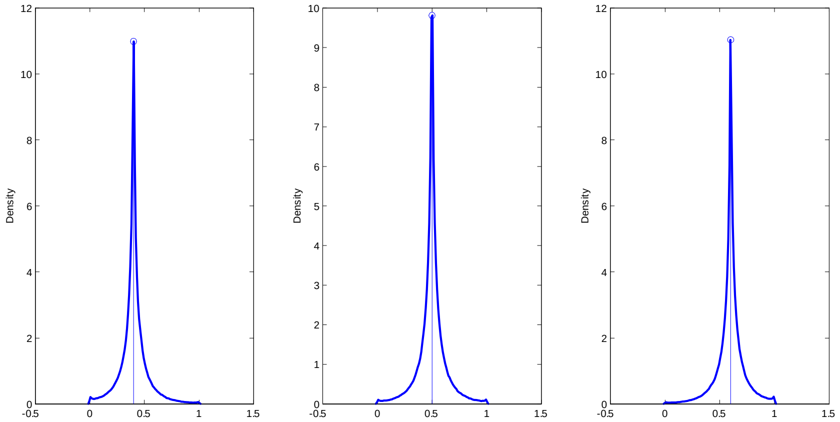

Figure 3.5: The density of $\widehat{\tau}_{ML}$ given in Equation (3.2.14) when $\tau_0 = 0.4, 0.5, 0.6$ (the left, middle and right panel respectively) and the signal-to-noise ratio (δ/σ) is 6. In each panel, the verticle line represents the true value of τ_0.

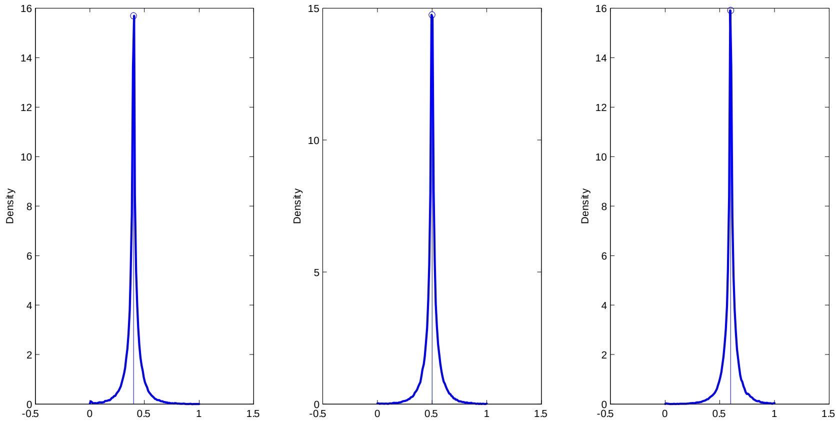

Figure 3.6: The density of $\widehat{\tau}_{ML}$ given in Equation (3.2.14) when $\tau_0 = 0.4, 0.5, 0.6$ (the left, middle and right panel respectively) and the signal-to-noise ratio (δ/σ) is 8. In each panel, the verticle line represents the true value of τ_0.

the large sample properties under the additional assumptions about the size of the structural break δ_T. In this Section, we will study the finite sample properties and the bias of $\widehat{\tau}_{ML}$ and $\widehat{k}_{ML}$ in a discrete time model.

Let us start with the continuous time model specified in Equation (3.2.10). Splitting the interval $[0,1]$ into $1/h$ subintervals so that each interval has a size of h, we then get $T = 1/h$ observations of the stochastic process $X(\cdot)$ at T equally spaced points $\{th\}_{t=1}^{T}$, and have the following exact discrete time representation:

$$X_{th} - X_{(t-1)h} = \begin{cases} \mu h + \sqrt{h}\varepsilon_{th} & \text{if } t = 1, \cdots, \lfloor \tau_0/h \rfloor, \\ (\mu+\delta)h + \sqrt{h}\varepsilon_{th} & \text{if } t = \lfloor \tau_0/h \rfloor + 1, \cdots, T, \end{cases} \tag{3.3.1}$$

where $\varepsilon_{th} \sim$i.i.d.$N(0,\sigma^2)$, $\lfloor \cdot \rfloor$ is the integer-valued function. Considering that ε_{th} is independent of h, we now write it as ε_t.

Letting $Z_t = \left(X_{th} - X_{(t-1)h}\right)/\sqrt{h}$, we obtain

$$Z_t = \begin{cases} \mu\sqrt{h} + \varepsilon_t & \text{if } t \leq \lfloor \tau_0/h \rfloor, \\ (\mu+\delta)\sqrt{h} + \varepsilon_t & \text{if } t > \lfloor \tau_0/h \rfloor. \end{cases} \tag{3.3.2}$$

Whenever h is fixed, the model in equation (3.3.2) is the same as the one in equation

(3.2.1) with ε_t being assumed to follow $N(0,\sigma^2)$, $k_0 = \lfloor \tau_0/h \rfloor$ being the absolute break point.

For the sequential limit distribution of Yao (1987) to be able to provide a good approximation to the finite sample distribution, it is required that the sample size T goes to infinity at a faster rate than that at which the squared structural break size goes to zero. However, in Model (3.3.2), when $h \to 0$, the structural break size $\delta\sqrt{h}$ goes to zero at the rate of $1/\sqrt{T}$. Hence, the sample size T does not go to infinity at a faster rate than that at which the squared structural break size goes to zero. As a result, Yao's limit distribution may not well approximate the finite sample distribution in (3.3.2) when h is small.

The simultaneously double asymptotic distribution given in Bai (1994) is essentially the same as the sequential limit distribution in Yao (1987). To derive the double asymptotic distribution, Bai (1994) assumed that the magnitude of break size goes to zero at a rate larger than $\sqrt{\log T}/\sqrt{T}$. However in Model (3.3.2), when $h \to 0$, $\delta\sqrt{h} = \delta/\sqrt{T} = O\left(1/\sqrt{T}\right)$. This may explain why Bai's limit distribution may not well approximate the finite sample distribution in (3.3.2) when h is small.

On the other hand, our exact finite sample distribution in Equation (3.2.14) can be regarded as a good approximation to the finite sample distribution of the ML estimator of the break point in model (3.3.1) when h is small. It is easy to find that the ML estimator of the break point in model (3.3.1) is the same as the one in Model (3.3.2). Therefore, the finite sample distribution in Equation (3.2.14) could well approximate the finite sample distribution of $\widehat{\tau}_{ML}$ in the discrete time model (3.3.2) when h is small. In particular, we expect there is no bias in $\widehat{\tau}_{ML}$ in the discrete time model (3.3.2) when $\tau_0 = 50\%$. However, we expect a upward (downward) bias in $\widehat{\tau}_{ML}$ in the discrete time model (3.3.2) when τ_0 is smaller (larger) than 50%. Since $\hat{k}_{ML} = \lfloor \widehat{\tau}_{ML} T \rfloor$, we expect the traditional estimator of the absolute break point k in the discrete time model (3.3.2) is also asymmetric and has the bias in finite sample. The bias in $\widehat{\tau}_{ML}$ and $\hat{k}_{ML}$ in the discrete time model will be discussed in detail in the next section.

Consider the special case when $\tau_0 = 50\%$. Notice that the fraction of the sample before and after the break is the same in this case. Also note that Equation (3.2.14) can be written as

$$\left(\frac{\delta}{\sigma}\right)^2 (\hat{\tau} - \tau_0) \overset{d}{=} \arg \max_{u \in \left[-\frac{1}{2}\left(\frac{\delta}{\sigma}\right)^2, \frac{1}{2}\left(\frac{\delta}{\sigma}\right)^2\right]} W(u) - \frac{1}{2}|u|. \tag{3.3.3}$$

This result is similar to the limit theory given by Equation (3.2.5). Given that δ should be replaced by $\delta\sqrt{h}$ and T should be replaced by $1/h$ in (3.2.5), the left hand side in the two equations are identical. The only difference is on the right hand side. The finite sample theory in the continuous time model is the location of the extremum over a finite interval which depends on the signal-to-noise ratio. The limit distribution in the discrete time model is the location of the extremum over an infinite interval. As a result, we expect the finite sample distribution be closer to the limit distribution when the signal-to-noise ratio is large. This expectation can be confirmed by the middle panels in Figure 3.2 - Figure 3.6.

3.4 Bias correction via indirect estimation

The indirect estimation is a simulation-based method, first introduced by Smith (1993), Gourieroux et al. (1993), and Gallant and Tauchen (1996). This method is particularly useful for estimating parameters of a model where the moments and likelihood function of the model are difficult to calculate but the model is easy to simulate. It uses an auxiliary model to capture aspects of the data upon which to base the estimation. The parameters of the auxiliary model can be estimated using either the observed data or data simulated from the true model. Indirect inference chooses the parameters of the true model so that these two sets of parameter estimates of the auxiliary model are as close as possible. Typically, one chooses the auxiliary model that is amenable to estimate and approximate the true model well at the same time.

Gourieroux et al. (1993) and Gallant and Tauchen (1996) established the asymptotic properties of the indirect estimator, including consistency, asymptotic normality, and asymptotic efficiency. McKinnon and Smith (1997) and Gourieroux et al. (2000) developed a particular indirect estimation procedure, where the auxiliary model is chosen to be the true model in order to improve finite sample properties of the original estimator. Arvanitis and Demos (2014) established primitive conditions for finite sample properties of the indirect estimator and also introduced an iterative procedure to further improve the performance of the indirect estimator. The indirect estimation obtains the bias function by simulating from the true model and hence the auxiliary model. In this section, we apply the indirect estimation procedure to do bias correction in estimating τ and k, the fractional and the absolute structural break point. It is important to obtain the bias function via simulations because, from Equations (3.2.14) and (3.3.3), we know that the bias formula and the bias expansion are too difficult to deal with explicitly. The same idea was used to estimate continuous time models in Phillips and Yu (2009) and dynamic panel models in Gourieroux et al. (2010).

The application of the indirect estimation procedure for estimating structural break proceeds as follows. Given a parameter θ (either τ or k), we simulate data $\tilde{y}(\theta) = \{\tilde{y}_0^h, \tilde{y}_1^h, \ldots, \tilde{y}_T^h\}$ from the true model, such as, Equation (3.2.10) or (3.2.1), where $h = 1, ..., H$, with H being the number of simulated paths. Note that T in $\tilde{y}(\theta)$ should be chosen as the same number of the actual data under analysis so that the bias of the original estimator from the actual observations can be calibrated by simulated data.

The indirect estimation method matches the estimator from the actual observations with the one estimated from the simulated data to obtain the indirect estimator. To be specific, let $Q_T(\theta; y)$ be the objective function of the original (biased) estimation method applied to actual data (y) for estimating the parameter θ. The corresponding extremum estimator $\hat{\theta}$ obtained is then denoted as

$$\hat{\theta}_T = \arg\max_{\theta \in \Theta} Q_T(\theta; y),$$

and the corresponding estimator based on the hth simulated path for some fixed θ is

$$\tilde{\theta}_T^h(\theta) = \arg\max_{\theta\in\Theta} Q_T(\theta; y(\theta)),$$

where Θ is a compact parameter space.

The indirect estimator is then defined as

$$\hat{\theta}_{T,H}^{IE} = \arg\max_{\theta\in\Theta} \left\| \hat{\theta}_T - \frac{1}{H}\sum_{h=1}^{H} \tilde{\theta}_T^h(\theta) \right\|,$$

for some distance measure $\|\cdot\|$. When H goes to infinity, it is expected that $\frac{1}{H}\sum_{h=1}^{H} \tilde{\theta}_T^h(\theta)$ $\xrightarrow{p} E(\tilde{\theta}_T^h(\theta))$. Then the indirect estimator becomes

$$\hat{\theta}_T^{IE} = \arg\max_{\theta\in\Theta} \left\| \hat{\theta}_T - b_T(\theta) \right\|$$

where $b_T(\theta) = E(\tilde{\theta}_T^h(\theta)$ is the binding (or bias) function. If $b_T(\theta)$ is invertible, then the indirect estimator can be directly written as

$$\hat{\theta}_T^{IE} = b_T^{-1}(\hat{\theta}_T).$$

To apply the indirect estimation to the observed data, we assume that the true model is given either by the continuous time model given by (3.2.10) or the discrete time model given by (3.2.1). At first, we employ the LS method of Bai (1994) or the ML method to the actual data in order to obtain $\hat{k}_T$. Then the corresponding estimator for the hth simulated path is $\tilde{k}_T^h(k)$ and the indirect estimation estimator is

$$\hat{k}_T^{IE} = \arg\max_{k\in\Theta} \left\| \hat{k}_T - b_T(k) \right\|,$$

where $\hat{k}_T$ is the original estimator of k from the actual data that has T observations, $b_T(k)$ is the binding function with the form

$$b_T(k) = E(\tilde{k}_T^h(k)),$$

which, in practice, can be effectively replaced by $\frac{1}{H}\sum_{h=1}^{H}\tilde{k}_T^h(k)$ since H can be chosen arbitrarily large. If the binding function is invertible, then

$$\hat{k}_T^{IE} = b_T^{-1}\left(\hat{k}_T\right). \tag{3.4.1}$$

Based on $\hat{k}_T^{IE}$, we can define the indirect estimator of the fractional break point as $\widehat{\tau}_T^{IE} = \hat{k}_T^{IE}/T$. Let the corresponding binding function be $b_T(\tau) = b_T(k)/T$. If $b_T(k)$ is invertible, $b_T(\tau)$ is also invertible. Hence,

$$\widehat{\tau}_T^{IE} = b_T^{-1}\left(\widehat{\tau}_T\right), \tag{3.4.2}$$

where $\widehat{\tau}_T$ is the original estimator of τ from the actual data.

Following the discussion of the finite sample property in Gourieroux et al. (2000) and Phillips (2012), we impose the following assumption.

Assumption 1. The binding function $b_T(\tau)$, mapping from $(0,1)$ to $b_T(0,1)$, is uniformly continuous and one-to-one.

Under Assumption 1, the binding function $b_T(\cdot)$ is invertible. We have $\widehat{\tau}_T^{IE}$ is "b_T-mean-unbiased", since

$$E\left(b_T\left(\widehat{\tau}_T^{IE}\right)\right) = E\left(\widehat{\tau}_T\right) = E(\widetilde{\tau}_T^h(\tau_0)) = b_T(\tau_0),$$

and

$$b_T^{-1}\left(E\left(b_T\left(\widehat{\tau}_T^{IE}\right)\right)\right) = \tau_0. \tag{3.4.3}$$

By the same reason, $\hat{k}_T^{IE}$ is also "b_T-mean-unbiased", i.e., $b_T^{-1}\left(E\left(b_T\left(\widehat{k}_T^{IE}\right)\right)\right) = k_0$.

Moreover, when $b_T(\cdot)$ is linear, the indirect estimator of τ and k is exactly mean-unbiased since, in (3.4.3), we have

$$b_T^{-1}\left(E\left(b_T(\widehat{\tau}_T^{IE})\right)\right) = E\left(b_T^{-1}\left(b_T(\widehat{\tau}_T^{IE})\right)\right) = E\left(\widehat{\tau}_T^{IE}\right) = \tau_0,$$

which is a especially appealing property in the practice when the binding function is close to linear.

It is important to point out the indirect estimator shares the same consistency property as the original estimator. Since only $\widehat{\tau}_T$ is consistent, hence we can only ensure the consistency of $\widehat{\tau}_T^{IE}$ but not $\widehat{k}_T^{IE}$.

Regarding the efficiency, from Equation (3.4.2) and by the Delta method, we have

$$\mathrm{Var}(\widehat{\tau}_T^{IE}) \approx \left(\frac{\partial b_T(\tau_0)}{\partial \tau}\right)^{-2} \mathrm{Var}(\widehat{\tau}_T). \tag{3.4.4}$$

Hence, the efficiency loss (or gain) is measured by $\frac{\partial b_T(\tau_0)}{\partial \tau}$. If $\left|\frac{\partial b_T(\tau_0)}{\partial \tau}\right| < 1$, $\widehat{\tau}_T^{IE}$ has a bigger variance than $\widehat{\tau}_T$. However, if $\left|\frac{\partial b_T(\phi_0)}{\partial \phi}\right| > 1$, $\widehat{\tau}_T^{IE}$ will have a small variance than $\widehat{\tau}_T$. If the finite sample distribution developed in Section 2 suggests that τ is over estimated when $\tau_0 < 50\%$ and is under estimated when $\tau_0 > 50\%$, the binding function is expected to be flatter than the 45 degrees line. As a result, we expect some efficiency loss from the indirect estimation as the variance of the indirect estimation will be larger than that of the original estimator.

3.5 Monte Carlo results

In this section, we design two Monte Carlo experiments to examine the bias in the LS estimator of k in the discrete time model (3.2.1) and the ML estimator of τ in the continuous time model (3.2.10), and to compare the finite sample performance of the indirect estimator and the original estimators. When inverting the binding function, following Phillips and Yu (2009), we choose a set of grid points for τ, namely, $\tau = [0.1, 0.11, ..., 0.89, 0.9]$ and calculate $b_T(\tau)$ for each τ via simulations. We then use the standard linear interpolation and extrapolation methods to obtain the binding functions in the domain $[0,1]$.[3]

In the first experiment, data are generated from Model (3.2.10), with $\sigma = 1$,

[3] However, if the indirect estimator of τ takes a value outside of the interval $[0,1]$ for one particular replication, such a replication is discarded for both ML and the indirect estimation.

Table 3.1: Monte Carlo comparison of the bias and RMSE of ML and Indirect Estimates. The number of simulated path is set to be 10,000 for indirect estimation. The number of replications is set at 10,000.

Case		Bias		Standard Error		RMSE	
$\frac{\delta}{\sigma}$	τ_0	ML	IE	ML	IE	ML	IE
2	0.3	0.1337	0.0736	0.1408	0.2688	0.1942	0.2787
2	0.5	-0.0016	-0.0025	0.1268	0.2407	0.1268	0.2407
2	0.7	-0.1323	-0.0712	0.1400	0.2669	0.1926	0.2762
4	0.3	0.0518	0.0222	0.1543	0.1870	0.1628	0.1883
4	0.5	0.0021	0.0029	0.1511	0.1820	0.1511	0.1820
4	0.7	-0.0435	-0.0137	0.1479	0.1787	0.1542	0.1792
6	0.3	0.0118	0.0037	0.1100	0.1163	0.1106	0.1164
6	0.5	0.0004	-0.0003	0.1172	0.1228	0.1172	0.1228
6	0.7	-0.0104	-0.0027	0.1092	0.1156	0.1097	0.1156

$\delta = 2,4,6$, $\tau_0 = 30\%, 50\%, 70\%$, $dB(t) \sim iid\ N(0,h)$, where $h = \frac{1}{1000}$. For each combination of δ and τ_0, we obtain the ML estimator of τ from Equation (3.2.12) and the indirect estimator. Our focus is to examine the finite sample properties of $\hat{\tau}$, so it is assumed that the structural break size δ and the standard deviation σ are known during the simulation.

Table 3.1 reports the bias, the standard error, and the root mean squared errors (RMSE) of the ML estimate and the indirect estimate of τ, obtained from 10,000 replications. Some observations can be obtained from the table. Firstly, when $\tau_0 = 50\%$, the ML estimate does not have any noticeable bias in all cases. However, when $\tau_0 \neq 50\%$, ML suffers from a bias problem. For example, when $\tau_0 = 30\%$ and $\delta/\sigma = 2$, the bias is 0.1337 and about 45% of the true value. This is very substantial. In general, the bias becomes larger when τ_0 is further away from 50%, or when the signal-to-noise ratio gets smaller. To the best of our knowledge, such a bias has not been discussed in the literature. Secondly, in all cases when $\tau_0 \neq 50\%$, the indirect estimate substantially reduces the bias. For example, when $\frac{\delta}{\sigma} = 2$ and $\tau_0 = 70\%$, the indirect estimation method removes about half of the bias in ML. Finally, the bias reduction by the indirect estimation method comes with a cost of a higher variance, which causes the RMSE of the indirect estimate slightly higher than its ML counterpart.

In the second experiment, data are generated from Model (3.2.1), with $\sigma = 1$, $\delta = 0.5, 1$, $\tau_0 = 0.3, 0.5, 0.7$, $\varepsilon_t \sim iid\, N(0,1)$, where we choose $T = 50, 80, 100, 120$. For each combination of δ, τ_0 and T, we obtain the LS estimate of k based on Equation (3.2.4) and the indirect estimate for each replication. As in the continuous time model, it is assumed that the structural break size δ and the standard deviation σ are known. The reason we focus on k is because k is a practically important parameter to estimate.

Table 3.2 reports the bias, the standard error, and the root mean squared errors (RMSE) of the ML estimate and the indirect estimator of k, obtained from 10,000 replications. We may draw the following conclusions from Table 3.2. First, when $\tau_0 = 50\%$, the LS estimate does not have any noticeable bias in all cases. However, when $\tau_0 \neq 50\%$, LS suffers from a bias problem. For example, when $T = 50, \tau_0 = 30\%$ and $\delta/\sigma = 0.5$, the bias is nearly 9 while the true value of k is 15. The bias is about 60% of the true value, which is very substantial. In general, the bias becomes larger when τ_0 is further away from 0.5 or when the signal-to-noise ratio gets smaller. To the best of our knowledge, such a bias has not been discussed in the literature. Secondly, in all cases when $\tau_0 \neq 50\%$, the indirect estimate substantially reduces the bias. For example, when $T = 80$, $\frac{\delta}{\sigma} = 0.5$ and $\tau_0 = 30\%$, the indirect estimation method removes more than half of the bias in ML. Finally, the bias reduction by the indirect estimation method comes with a cost of a higher variance, which causes the RMSE of the indirect estimate slightly higher than its ML counterpart.

To understand why the indirect estimation increases the variance, we plot the binding functions in these two models in Figure 3.7 and Figure 3.8, where we also plot the 45 degrees line for the purpose of comparison. Figure 3.7 corresponds to the continuous time model with $\delta = 2, 4, 6$ and Figure 3.8 to the discrete time model with $T = 100$, $\delta = 0.5, 1$. Several conclusions can be made. Firstly, the binding functions always pass through the 45 degrees line at the middle point of τ, suggesting no bias when $\tau = 50\%$ and that the bias becomes smaller when the true

Table 3.2: Monte Carlo comparison of the bias and RMSE of LS and Indirect Estimates. The number of simulated path is set to be 10,000 for indirect estimation. The number of replications is set at 10,000.

Case				Bias		Standard error		RMSE	
T	$\frac{\delta}{\sigma}$	τ_0	k_0	LS	IE	LS	IE	LS	IE
50	0.5	0.3	15	8.9750	6.8050	3.7450	11.6250	9.7250	13.4703
50	0.5	0.5	25	0.0250	-0.0300	3.0950	9.3150	3.0951	9.3150
50	0.5	0.7	35	-8.8650	-6.4750	3.7400	12.0500	9.6216	13.6795
50	1	0.3	15	1.4150	-0.8200	5.0550	6.8500	5.2493	6.8989
50	1	0.5	25	-0.1050	-0.1500	4.5900	5.8700	4.5912	5.8719
50	1	0.7	35	-1.6450	0.4500	5.0950	6.9350	5.3540	6.9496
80	0.5	0.3	24	11.728	5.472	7.544	17.88	13.9448	18.6986
80	0.5	0.5	40	-0.016	-0.632	5.912	12.832	5.9120	12.8476
80	0.5	0.7	56	-12.088	-7.592	5.4432	18.256	13.2570	19.7717
80	1	0.3	24	0.936	-0.352	6.752	7.68	6.8166	7.6881
80	1	0.5	40	-0.008	-0.024	6.2	6.792	6.2000	6.7920
80	1	0.7	56	-0.944	0.208	6.976	7.976	7.0396	7.9787
100	0.5	0.3	30	12.83	4.36	10.66	23.20	16.6807	23.6061
100	0.5	0.5	50	0.35	0.26	8.02	15.13	8.0276	15.1322
100	0.5	0.7	70	-9.22	2.01	10.21	22.02	13.7569	22.1115
100	1	0.3	30	0.72	-0.11	7.28	7.79	7.3155	7.7908
100	1	0.5	50	0.06	0.02	6.49	6.80	6.4903	6.8000
100	1	0.7	70	-0.82	0.09	7.53	8.11	7.5745	8.1105
120	0.5	0.3	36	6.636	-4.724	14.724	24.3	16.1503	24.7549
120	0.5	0.5	60	-0.096	0.252	12.792	20.388	12.7924	20.3896
120	0.5	0.7	84	-6.936	3.816	14.82	24.66	16.3628	24.9535
120	1	0.3	36	0.588	-0.096	7.308	7.656	7.3316	7.6566
120	1	0.5	60	0	-0.024	6.756	6.984	6.7560	6.9840
120	1	0.7	84	-0.504	0.108	7.176	7.524	7.1937	7.5248

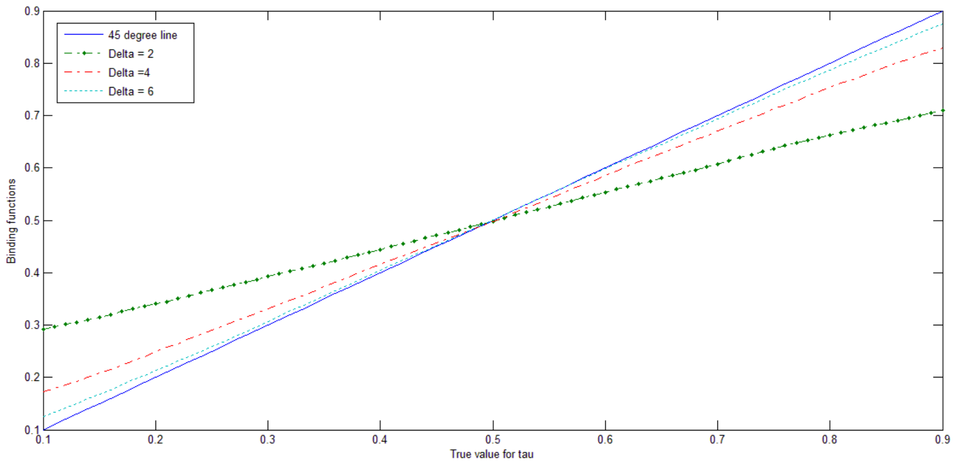

Figure 3.7: Binding function of ML for the continuous time model when $h = 0.001$

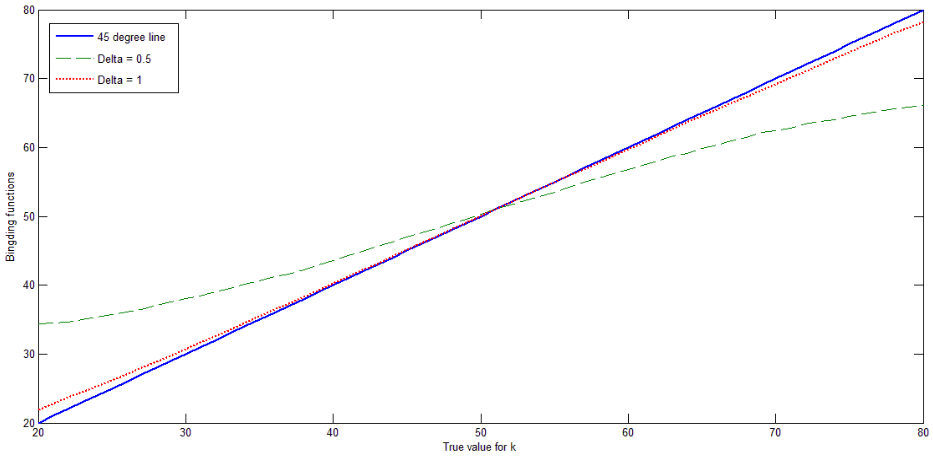

Figure 3.8: Binding function of LS for the discrete time model when T = 100

break point gets close to the middle. Second, the binding functions monotonically increase as τ or k increases, suggesting that the binding functions are invertible. However, in all cases, the binding functions are flatter than the 45 degrees line, explaining why the variance of the indirect estimate is larger than that of the ML estimate. The smaller the signal-to-noise ratio, the flatter the binding function and hence the bigger loss in efficiency. Third, the binding function is not exactly a straight line. It is easy to see the nonlinearity near the two boundaries when $\delta = 0.5$ in the discrete time model. Due to the presence of nonlinearity, the indirect estimation procedure cannot completely remove the bias, although it is "b_T-mean-unbiased".

3.6 Conclusions

This paper is concerned about the finite sample properties in the estimation of structural break points. We find that the finite sample bias is substantial in many practically relevant situations. While the literature on structural break has focused the a great deal of attention to develop asymptotic properties, the finite sample problem has received no attention in this literature, to the best of our knowledge. We hope to fill up this important gap in the literature.

In this paper we address the finite sample problem in several aspects. First we derive the finite sample distribution of the structural break estimator in the continuous time model. We then establish its connection to the discrete time models considered in the literature. It is shown that when the true break point is at the middle of the sample, the finite sample distribution is symmetric but can have trimodality. However, when the true break point occurs earlier than the middle of the sample, the finite sample distribution is skewed to the right and there is a positive bias. When the true break point occurs later than the middle of the sample, the finite sample distribution is skewed to the left and there is a negative bias.

To reduce the bias in finite sample, we obtain the binding functions via simulations and then use the indirect estimation technique to estimate the break parameter. Indirect estimation essentially inverts the binding function at the original estimator obtained from the actual data. It inherits the asymptotic properties of the original estimator but reduces the finite sample bias. Monte Carlo results show that the indirect estimation procedure is effective in reducing the bias of the traditional break point estimators.

The models considered in this paper are very simply in nature. Also, the estimators considered are based on the full sample. Real time (and hence subsample) estimators tend to have more serious finite sample problems. Further studies on developing the finite sample distribution for more realistic models and real time estimators are needed. How to extend the indirect estimation technique in a multiple parameter settings are also useful.

Chapter 4 Panel kernel estimation of integrated variance and the microstructure noise function

4.1 Introduction

The last two decades have witnessed substantial progress in the measurement of price volatility in financial assets using ultra high frequency data. Much of the research has concentrated on the empirical quadratic variation process, a quantity that is now generally known as realized variance (RV) or volatitily. This empirical measure provides a natural estimate of the integrated variance (IV) of the underlying stochastic price process. The method is nonparametric in nature because the quantity is computed directly as the empirical quadratic variation of the data and it is not necessary to impose any parametric assumptions on the dynamics of the true efficient price process for the underlying asset.

A central difficulty in the estimation of IV using very high frequency observations is that such data are typically contaminated by the presence of market microstructure noise. Seeking to address this difficulty, many methods have been proposed to estimate IV, taking account of the microstructure noise contamination. The literature is now large and prominent examples are Zhang et al (2005, ZMA hereafter), Bandi and Russell (2008) and Bardnorff-Nielsen et al (2008, BHLS hereafter). Throughout this literature it has become common to employ specific parametric assumptions about the form of the microstructure noise. In most cases, the noise manifests in the form of a linear term involving the error variance, a form that is justified in the case of pure noise contamination involving indepenendent and identically distributed (IID) additive noise effects that are independent of the true efficient price. It has been argued (e.g., Hansen and Lunde, 2006) that such assumptions are too strong. By way of illustration, Phillips and Yu (2006) showed that the pure noise assumption is incompatible with the so-called flat pricing phenomenon, an empirical regularity that is widely observed in financial asset price data at medium and ultra high frequencies.

The present paper explores a new mechanism for modeling, estimating and testing the impact of microstructure noise contamination. The idea was originally suggested in Phillips and Yu (2006), which provided some empirical evidence supporting the methodology. The approach relies on a functional panel model formulation that represents noise effects through a fixed design nonparametric function, which leads to a convenient and robust IV estimate and a procedure for testing linear and other microstructure noise effects.

The paper contributes by extending the RV literature in several ways. First, the approach leads to a new panel model formulation that pools data over different trading days to enhance estimation of the microstructure process, while also allowing for more general forms of noise contamination. The resulting estimate is more efficient than time series estimates that only use data within a single trading day to adjust for noise. Second, the approach provides new machinery to test the functional form of microstructure noise contamination. Such tests are useful in assessing the nature and potential sources of this contamination because of their implied functional form properties. Third, the paper contributes to the nonparametric panel literature by estimating fixed design (trend) functions in the presence of heteroskedastic and serially correlated errors and by developing the asymptotic theory for these nonparametric procedures.

The paper is organized as follows. Section 2 introduces the model framework and proposes methods of estimation and inference. Section 3 develops the corresponding asymptotic theory for these estimators and tests. Section 4 concludes. Proofs are given in the Appendix 2.

4.2 The framework and our approach

4.2.1 The model

Let $p^*(t)$ be the logarithm of the true efficient price. Without loss of generality, we assume $p^*(t)$ follows a Brownian semimartingale,

$$dp^*(t) = \sigma(t)dB(t), \tag{4.2.1}$$

where $B(t)$ is a standard Brownian motion and $\sigma(t)$ is a càdlàg volatility process whose specification is nonparametric. We wish to estimate the quantity $IV = \int_0^1 \sigma^2(t)dt$, the IV of $p^*(t)$ over a fixed time interval, say a single day. The quantity IV is defined as the limit of the empirical quadratic variation

$$IV = \text{plim}_{\Delta \to 0} \sum_{i=1}^{m} [p^*_{i,m} - p^*_{i-1,m}]^2, \tag{4.2.2}$$

where $p^*_{i,m} = p^*(t_{i,m})$, $0 = t_{0,m} < t_{1,m} < \cdots < t_{m,m} = 1$ is a grid on $[0,1]$, and $\Delta = \sup_i |t_{i,m} - t_{i-1,m}|$ is the maximum grid size.

The observed sample component of (4.2.2) on the given grid is the empirical quadratic variation and is called the realized variance, i.e.,

$$RV^{(m)}(p^*) := \sum_{i=1}^{m} [p^*_{i,m} - p^*_{i-1,m}]^2.$$

It is known that as $\Delta \to 0$ (and hence $m \to \infty$),

$$\sqrt{m}\left[RV^{(m)}(p^*) - IV\right] \xrightarrow{d} MN\left(0, 2\int_0^1 \sigma^4(t)dt\right), \tag{4.2.3}$$

where MN represents mixed normality; see, for example, Barndorff-Nielsen and

Shephard (2002).

Both the consistency and the asymptotic distribution of $RV^{(m)}(p^*)$ require that $p^*_{i,m}$ be observed. At reasonably high frequencies, market microstructure creates challenges for the immediate use of (4.2.2) and (4.2.3) in inference. To capture market microstructure effects, it is frequently assumed that

$$p(t) = p^*(t) + u(t), \tag{4.2.4}$$

where $p(t)$ is the logarithm of the observed price and $u(t)$ represents the influence of microstructure noise.

Many studies assume that the noise process $u(t)$ and the price process $p^*(t)$ are independent; and some research (e.g., Zhou, 1996, Bandi and Russell, 2008, ZMA, 2005) assumes a pure noise structure for $u(t)$, so that $u(t)$ is iid over discrete values of t. Other studies (e.g. Hansen and Lunde, 2006 and Aït-Sahalia, Mykland and Zhang, 2005) assume that $u(t)$ is covariance stationary.

A pure noise assumption substantially simplifies the estimation of IV. To see this, suppose that $u(t) \overset{iid}{\sim} (0, \omega^2)$ over discrete t and $u(t)$ is independent of $p^*(t)$. Then,

$$\mathbb{E}\left(RV^{(m)}(p) | \{p^*(t)\}_0^1\right) = IV + 2\omega^2 m, \tag{4.2.5}$$

which leads to the following representation, with error ε_m,

$$RV^{(m)}(p) = IV + 2\omega^2 m + \varepsilon_m. \tag{4.2.6}$$

The second term in equation (4.2.5) clearly dominates as $m \to \infty$ and $\Delta \to 0$. Thus, in this model, $RV^{(m)}(p)/(2m) \xrightarrow{p} \omega^2$ and a consistent estimate of ω^2 is $RV^{(m)}(p)/(2m)$. Alternative time series techniques, such as the jackknife method of ZMA and the realized kernel method of BHLS, have been proposed to estimate IV consistently. The asymptotic distributions of these alternative estimates have also been obtained under suitable regularity conditions.

When there is no confusion, we simply write $RV_m = RV^{(m)}(p)$. So Equation (4.2.6) may be represented as

$$RV_m = IV + 2\omega^2 m + \varepsilon_m. \tag{4.2.7}$$

When $u(t)$ is not iid or when $u(t)$ and $p^*(t)$ are dependent on each other, $\mathbb{E}\left(RV_m | \{p^*(t)\}_0^1\right)$ may not be a linear function in m. For instance, if $u(t_i)$ has long range dependence with memory parameter $1+d$ such that $e_{it} = (1-L_i)^{1+d} u(t_i)$ is stationary, where $L_i u(t_i) = u(t_{i-1})$, then $\sum_{i=1}^m [u(t_i) - u(t_{i-1})]^2 = \sum_{i=1}^m \left\{(1-L_i)^{-d} e_{it_i}\right\}^2 = O\left(m^{1+d}\right)$. In such cases, it is useful to employ a more general functional relationship between RV_m and m of the form

$$RV_m = IV + f(m) + \varepsilon_m, \tag{4.2.8}$$

where the microstructure noise function $f(m)$ may be linear or nonlinear. It is convenient in practical work to adopt an agnostic position about the source and nature of the time series microstructure noise, leading the noise function $f(m)$ nonparametrically specified.

Following the suggestion in Phillips and Yu (2006), we pursue the idea of estimating IV and testing the microstructure noise function using equation (4.2.8). Suppose RV_m is available at $m = n_0, ..., n_N$. If $f(m)$ is continuous and asymptotically homogeneous of degree γ as $m \to \infty$, then we can represent (4.2.8) in the form

$$RV_m/n_N^\gamma = IV/n_N^\gamma + f(m/n_N) + \varepsilon_m/n_N^\gamma, m = n_0, ..., n_N. \tag{4.2.9}$$

This model is closely related to the time-varying parameter model considered in Robinson (1989),

$$y_t = \beta^\top(t/N)x_t + \sigma(t/N)u_t, u_t \sim iid(0,1), t = 1, ..., N, \tag{4.2.10a}$$

where the time varying design functions $\beta_t \, (:= \beta(t/N))$ and $\sigma_t \, (:= \sigma(t/N))$ have support on $[0,1]$. Robinson proposed nonparametric estimators for β_t and σ_t and established consistency and asymptotic normality for the proposed estimators. He also proposed a consistent estimator of the asymptotic covariance matrix of the β_t estimators. Model (4.2.9) has the additional complication that the scaled error process $\varepsilon_m/n_N^{\gamma}$ will be serially correlated in general.

It may be reasonable to assume the random mechanism that generates the behavior of the microstructure noise has some stability over short time intervals, such as a few days (say D). Accordingly, let $d = 1,...,D$ index these days and denote y_{dn_i} the RV calculated from a grid that contains n_i intra-day returns for day d. Similarly, let $\alpha_d = \int_0^1 \sigma_{dt}^2 dt$ denote the IV for day d where σ_{dt}^2 is the diffusion function of the true efficient price for day d. Since IV is a random variable it varies from day to day.

Using the same setting as Phillips and Yu (2006), we formulate a nonparametric noise function in the framework of a panel data model as

$$y_{dn_i} = \alpha_d + f(n_i) + \varepsilon_{dn_i}, d = 1,...,D, i = 1,...,N, \qquad (4.2.11)$$

where α_d is the fixed effect, and ε_{dn_i} is an induced error process that may be heterogeneous and serially correlated over n_i. It is expected that both $N \to \infty$ and $n_N \to \infty$, so that dual asymptotics operate in the limit. Throughout the paper, we assume that the maximum grid size goes to zero as $n_N \to \infty$.

The nonparametric panel data model (4.2.11) offers new opportunities for estimating IV and testing the microstructure noise function. First, different from standard time series approaches, we pool together the information across different days, leading to a more efficient estimate of IV when the microstructure noise function is stable. Second, the new framework provides a simple but powerful way to test the microstructure noise assumption. For example, a nonlinear microstructure noise function $f(m)$ is evidence against pure noise contamination of the efficient price. The nonparametric panel model (4.2.11) is closely related to the panel data model considered Robinson (2012) and Zhang et al. (2012) where

$$y_{di} = \alpha_d + f\left(i/N\right) + \varepsilon_{di},\ d = 1,...,D, i = 1,...,N,$$

and the ε_{di} are unobservable zero-mean random variables, uncorrelated and homoskedastic across i, but possibly correlated and heteroskedastic over d. Our model has different features stemming from the panel construction, leading to an error process may be heterogeneous and serially correlated over n_i. Addressing these features provides an extension of the nonparametric panel literature, giving a new estimate for the trend function and developing the corresponding asymptotic theory.

It is reasonable to assume ε_{dn_i} has zero mean and its variance is $O\left(n_i\right)$. If $f(n_i)$ is continuous and asymptotically homogeneous of degree γ as $n_i \to \infty$, The noise function is then formulated in standardized form as

$$\frac{y_{dn_i}}{n_N^{\gamma}} = \frac{\alpha_d}{n_N^{\gamma}} + f\left(\frac{n_i}{n_N}\right) + \frac{\varepsilon_{dn_i}}{n_N^{\gamma}}, \tag{4.2.12}$$

where the standardized errors $\varepsilon_{dn_i}/n_N^{\gamma} \sim o_p(1)$ are heterogeneous and serially correlated over n_i.

For the purpose of identification, it is assumed that

$$f\left(\frac{n_0}{n_N}\right) = 0, \tag{4.2.13}$$

so that at some base level n_0 for return calculations there is no microstructure noise effect. Accordingly, Eq. (4.2.13) implies that the RV calculated under n_0 observations yields an unbiased estimator because $\mathbb{E}\left(y_{dn_0}\right) = \alpha_d$. Using Eq. (4.2.13), we have

$$\frac{y_{dn_i} - y_{dn_0}}{n_N^{\gamma}} = f\left(\frac{n_i}{n_N}\right) + \frac{\varepsilon_{dn_i} - \varepsilon_{dn_0}}{n_N^{\gamma}}. \tag{4.2.14}$$

Averaging Eq. (4.2.14) across days leads to

$$\frac{\bar{y}_{An_i} - \bar{y}_{An_0}}{n_N^{\gamma}} = f\left(\frac{n_i}{n_N}\right) + \frac{\bar{\varepsilon}_{An_i} - \bar{\varepsilon}_{An_0}}{n_N^{\gamma}}, \tag{4.2.15}$$

where

$$\bar{y}_{An_i} = \frac{1}{D}\sum_{d=1}^{D} y_{dn_i}, \quad \bar{\varepsilon}_{An_i} = \frac{1}{D}\sum_{d=1}^{D} \varepsilon_{dn_i}.$$

Defining $\varepsilon_{dn_i}/n_N^\gamma =: \sigma_\varepsilon\left(\frac{n_i}{n_N}\right) e_{dn_i}$, Eq. (4.2.15) can be rewritten as

$$\frac{\bar{y}_{An_i} - \bar{y}_{An_0}}{n_N^\gamma} = f\left(\frac{n_i}{n_N}\right) + \sigma_\varepsilon\left(\frac{n_i}{n_N}\right)\bar{e}_{An_i} - \sigma_\varepsilon(\frac{n_0}{n_N})\bar{e}_{An_0}. \tag{4.2.16}$$

Denoting $Y_{n_i} = \frac{\bar{y}_{An_i} - \bar{y}_{An_0}}{n_N^\gamma}$, $\sigma_\xi(\frac{n_i}{n_N}) = \sigma_\varepsilon(\frac{n_i}{n_N})/\sqrt{D}$, and $\xi_{n_i} = \sqrt{D}\bar{e}_{An_i} = \frac{1}{\sqrt{D}}\sum_{d=1}^{D} e_{dn_i}$, we have the following estimable model

$$Y_{n_i} = f\left(\frac{n_i}{n_N}\right) + \sigma_\xi\left(\frac{n_i}{n_N}\right)\xi_{n_i}, \tag{4.2.17}$$

in which the term $\sigma_\varepsilon(\frac{n_0}{n_N})\bar{e}_{An_0}$, which does not depend in n_i, is absorbed into the function $f\left(\frac{n_i}{n_N}\right)$.

4.2.2 A nonparametric estimator of the noise function

Let $F(\tau) = (f(\tau), f'(\tau))'$ where $f'(\tau) = \partial f(\tau)/\partial\tau$. The local linear estimator of F at τ is

$$\hat{F}(\tau) = \left(\hat{f}(\tau), \hat{f}'(\tau)\right)' = \arg\max_{a,b}\sum_{i=1}^{N}\left\{Y_{n_i} - a - b\left(\frac{n_i}{n_N} - \tau\right)\right\}^2 k\left(\frac{\frac{n_i}{n_N} - \tau}{h}\right). \tag{4.2.18}$$

where $k(u)$ is a kernel function and h is a positive scalar bandwidth.

Let $W_\tau =$diag$\left(k\left(\frac{n_1 - \tau n_N}{n_N h}\right), \ldots, k\left(\frac{n_N - \tau n_N}{n_N h}\right)\right)$ and

$$Y = \begin{pmatrix} Y_{n_1} \\ \vdots \\ Y_{n_N} \end{pmatrix}, \quad X_\tau = \begin{pmatrix} 1 & \frac{n_1}{n_N} - \tau \\ \vdots & \vdots \\ 1 & \frac{n_N}{n_N} - \tau \end{pmatrix}.$$

In the matrix form, $\hat{F}(\tau)$ solves

$$\min_F (Y - X_\tau F)' W_\tau (Y - X_\tau F), \tag{4.2.19}$$

and the corresponding analytic form for $\hat{f}(\tau)$ is given by

$$\hat{f}(\tau) = \iota' \left(X_\tau' W_\tau X_\tau\right)^{-1} \left(X_\tau' W_\tau Y\right), \tag{4.2.20}$$

where $\iota = (1,0)'$.

4.2.3 A new estimator of integrated variance

ZMA (2005) designed a jackknife method, based on two different time scales, to estimate IV. Our proposed approach extends the jackknife method of ZMA allowing for a wider class of microstructure noise and a nonparametric formulation of its effects.

From Model (4.2.12), for the fast time scale, we have,

$$\frac{y_{dn_N}}{n_N^\gamma} = \frac{\alpha_d}{n_N^\gamma} + f(1) + \frac{\varepsilon_{dn_N}}{n_N^\gamma}. \tag{4.2.21}$$

And for the slow time scale, we divide the whole sample into the K subsamples and for each subsample, we have J $(:= n_N/K)$ observations. Therefore, for the kth subsample, we have

$$\frac{y_{dJ}^k}{J^\gamma} = \frac{\alpha_d}{J^\gamma} + f(1) + \frac{\varepsilon_{dJ}^k}{J^\gamma}, \tag{4.2.22}$$

so that

$$\frac{1}{K}\sum_{k=1}^K y_{dJ}^k = \alpha_d + J^\gamma f(1) + \frac{1}{K}\sum_{k=1}^K \varepsilon_{dJ}^k.$$

So an estimator for IV is

$$\begin{aligned} \hat{\alpha}_d &= \frac{1}{K}\sum_{k=1}^K y_{dJ}^k - \frac{J^\gamma y_{dn_N}}{n_N^\gamma} \\ &= \left(1 - \frac{1}{K^\gamma}\right)\alpha_d + \frac{1}{K}\sum_{k=1}^K \varepsilon_{dJ}^k - \frac{J^\gamma \varepsilon_{dn_N}}{n_N^\gamma} \\ &= \left(1 - \frac{1}{K^\gamma}\right)\alpha_d + \frac{O(J^{\gamma-\delta})}{\sqrt{K}}\frac{1}{\sqrt{K}}\sum_{k=1}^K e_{dJ}^k - \frac{O(J^{\gamma-\delta})}{K^\delta} e_{dn_N} \xrightarrow{p} \alpha_d \end{aligned} \tag{4.2.23}$$

If $J \to \infty$, $K \to \infty$, $\frac{O(J^{\gamma-\delta})}{\sqrt{K}} = o(1)$, $\frac{O(J^{\gamma-\delta})}{K^\delta} = o(1)$, and $\frac{1}{\sqrt{K}}\sum_{k=1}^K e_{dJ}^k = O_p(1)$. Under pure noise, we have $\gamma = 1$, $\delta = \frac{1}{2}$ so that $\frac{O(J^{\gamma-\delta})}{\sqrt{K}} = o(1)$, $\frac{O(J^{\gamma-\delta})}{K^{\delta}} = o(1)$ and $\frac{1}{\sqrt{K}} \sum_{k=1}^{K} e_{dJ}^{k} = O_p(1)$ when $\frac{J}{K} \to 0$, which corresponds to ZMA (2005).

The above approach requires prior knowledge of γ. In practical applications, we do not know γ unless parametric assumptions are made about the microstructure noise. We propose a feasible version to estimate α_d without making a specific assumption about the effect of microstructure noise. In particular, if f has the parameterized power function function form

$$f(x) = x^{\gamma} f(1),$$

we have

$$\frac{J^{\gamma}}{n_N^{\gamma}} = \frac{f(\frac{J}{n_N})}{f(1)},$$

so that a feasible estimate of IV is

$$\hat{\alpha}_d^f = \frac{1}{K} \sum_{k=1}^{K} y_{dJ}^{k} - \frac{\hat{f}(\frac{J}{n_N}) y_{dn_N}}{\hat{f}(1)}.$$

4.3 Asymptotic theory

Following Robinson (2012), we use the mean squared error (MSE) and mean integrated error (MISE) to capture the goodness of fit of the nonparametric estimates. Minimizing these measures leads to the optimal choice of bandwidth. Before developing the asymptotic theory, we make the following assumptions.

Assumption 1: *The standardized microstructure noise function $f(\tau)$ is asymptotically homogeneous of degree $\gamma > 1/2$, twice continuously differentiable on $[0,1]$ and has bounded 3rd derivative on $[0,1]$.*

Assumption 2: *The kernel functions, $k(u)$ and $\widetilde{k}(u)$ have support on $[-1,1]$, are symmetric around zero, non-negative, continuously differentiable and satisfy*

$$\int_{-1}^{1} k(u)du = 1, \quad \int_{-1}^{1} \widetilde{k}(u)du = 1. \tag{4.3.1}$$

Assumption 3: *As* $N \to \infty$*,* $h \to 0, \widetilde{h} \to 0$ *and* $\liminf_{n\to\infty} n_N h^4 > 0$*,* $\liminf_{n\to\infty} n_N \widetilde{h}^4 > 0$ *where h,* $\widetilde{h}$ *are bandwidths used in the kernel function estimation.*

Assumption 4: *Assume*

$$\frac{\varepsilon_{dn_i}}{n_N^{\gamma}} = \sigma_{\varepsilon}\left(\frac{n_i}{n_N}\right) e_{dn_i} = \sigma\left(\frac{n_i}{n_N}\right) e_{dn_i}/n_N^{\delta}, \tag{4.3.2}$$

where (1) $\delta > 0$ *and* $\sigma(\tau)$ *is nonnegative and twice continuously differentiable on* $[0,1]$*; (2)* $\mathbb{E}\left(e_{dn_i}\right) = 0$*,* $\sup_i \mathbb{E}|e_{dn_i}|^{\beta} < \infty$ *for some* $\beta > 2$*,* $\mathbb{E}\left(e_{dn_i}^2\right) = 1$*,* $\mathbb{E}\left(e_{kn_i} e_{ln_j}\right) = 0$ *for all i and j with* $k \neq l$*, and* $\{e_{dn_i}\}_{i=1}^{\infty}$ *is strictly stationary* α*-mixing with mixing coefficients* α_m *of size -*$\frac{\beta}{\beta-2}$*, so that* $\sum_{m=1}^{\infty} \alpha_m^{1-2/\beta} < \infty$ *for all d. By construction and by Assumptions 3 and 4, we have (1)* $\sigma_{\xi}(\cdot) = \frac{\sigma(\cdot)}{n_N^{\delta}\sqrt{D}}$ *with* $\delta > 0$*; (2)* $\mathbb{E}(\xi_{n_i}) = 0$*,* $\sup_i \mathbb{E}|\xi_{n_i}|^{\beta} < \infty$ *for some* $\beta > 2$*,* $\mathbb{E}(\xi_{n_i})^2 = 1$*, and* $\{\xi_{n_i}\}_{i=1}^{\infty}$ *is strictly stationary* $\alpha-$*mixing with mixing coefficients* α_m *of size -*$\frac{\beta}{\beta-2}$ *and* $\sum_{m=1}^{\infty} \alpha_m^{1-2/\beta} < \infty$*.*

Assumption 5: δ *and h satisfy* $n_N^{\delta-3/2} h^{1/2} D^{1/2} + n_N^{\delta-1/2} h^{5/2} D^{1/2} = o(1)$*.*

Remark 1: Assumption 1 is needed to ensure Model (4.2.11) is compatible upon transformation with Model (4.2.12) so that nonparametric estimation of f over the interval $[0,1]$ is consistent. In effect, the amount of local information about f on $[0,1]$ needs to increase at a suitable rate, so that variance and bias decrease to achieve consistent estimation (c.f., Robinson, 1989).

Remark 2: To simplify the following development, we assume $n_0 < n_1 < \cdots < n_N$ and that the associated grids are all equally spaced with $n_{i+1} - n_i = g < \infty$. The assumption of equal spacing grids is not required for developing the consistency and asymptotic normality of the estimated noise function. But it does facilitate estimation of the covariances.

Theorem 1 *Let* $\hat{f}(\tau)$ *be defined as in Equation (4.2.20). Under Assumptions 1,* *2, 3, 4, as $N \to \infty$, we have*

$$MSE\left\{\hat{f}(\tau)\right\} \sim \frac{h^4 f''(\tau)^2\left\{\int_{-1}^{1} u^2 k(u)du\right\}^2}{4} + \frac{\sigma^2(\tau)\int_{-1}^{1} k^2(u)du}{n_N^{1+2\delta} hD}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right]$$
$$+O\left(\frac{1}{n_N^2}+\frac{h^2}{n_N}\right)^2,$$

where $\rho_{gm} = cov(e_{n_i}, e_{n_i+gm})$.

Remark 3: The term $O\left(\frac{1}{n_N^2}+\frac{h^2}{n_N^2}\right)^2$ arises from the nonparametric approximation error. The term is retained since the variance term has order of magnitude $O\left(\frac{1}{n_N^{1+2\delta}h}\right)$, which may be smaller than the approximation error when δ is large. As a result, to obtain asymptotic normality, we need the additional rate condition on δ and h given in Assumption 5. Applying a central limit theorem for mixing sequences, we obtain the asymptotic distribution for the nonparametric estimator $\hat{f}(\tau)$ given in the following result.

Theorem 2 *Under Assumptions 1, 2, 3, 4, 5 as $N \to \infty$, we have*

$$n_N^{\frac{1}{2}+\delta} h^{\frac{1}{2}}\left[\hat{f}(\tau) - f(\tau) - \frac{h^2 f''(\tau)\int_{-1}^{1} u^2 k(u)du}{2}\right] \xrightarrow{d} N(0, V(\tau)),$$

where

$$V(\tau) = \frac{\sigma^2(\tau)\int_{-1}^{1} k^2(u)du}{D}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right].$$

From Theorems 4.3 and 4.3 optimal bandwidth formulae are obtained as follows.

Theorem 3 *Let Assumptions 1, 2, 3, 4, 5 hold. The optimal h that minimizes the asymptotic MSE (AMSE) and MISE of $\hat{f}$ is, respectively,*

$$h_{AMSE}(\tau) = \left\{\frac{\sigma^2(\tau)\int_{-1}^{1} k^2(u)du\,[1+2\sum_{m=1}^{\infty}\rho_{gm}]}{n_N^{1+2\delta} D f''(\tau)^2\left\{\int_{-1}^{1} u^2 k(u)du\right\}^2}\right\}^{\frac{1}{5}},$$

$$h_{AMISE} = \left\{ \frac{\int_0^1 \sigma^2(\tau)d\tau \int_{-1}^1 k^2(u)du\,[1+2\sum_{m=1}^{\infty}\rho_{gm}]}{n_N^{1+2\delta} D \int_0^1 f''(\tau)^2 d\tau \left\{\int_{-1}^1 u^2 k(u)du\right\}^2} \right\}^{\frac{1}{5}}.$$

Hence, the optimal convergence rate of AMSE and AMISE is $n_N^{-\frac{4}{5}(1+2\delta)}$, which is faster than the usual nonparametric convergence rate $n_N^{-\frac{4}{5}}$ (see for example Cai, 2007).

To calculate confidence intervals, we need to find a consistent estimator for the asymptotic variance of $\hat{f}(\tau)$, i.e., a consistent estimator for

$$\begin{aligned} Avar(\hat{f}(\tau)) &\sim \frac{\sigma^2(\tau)\int_{-1}^1 k^2(u)du}{n_N^{1+2\delta}hD}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right] \\ &\sim \frac{\sigma_\xi^2(\tau)\,[1+2\sum_{m=1}^{\infty}\rho_{gm}]\int_{-1}^1 k^2(u)du}{n_N h}. \end{aligned} \tag{4.3.3}$$

To do so, we first apply the method of Fan and Yao (1998) to estimate the variance function $\sigma_\xi^2(\cdot)$ at the fixed point τ. Let $R_{n_i} = \left(Y_{n_i} - f(\frac{n_i}{n_N})\right)^2$ and $\mathbb{E}(R_{n_i}) = \sigma_\xi^2(\frac{n_i}{n_N})$. Denote the squared residual by $\hat{R}_{n_i} = \left(Y_{n_i} - \hat{f}(\frac{n_i}{n_N})\right)^2$ and let $\hat{\lambda}$, $\hat{\beta}$ solve

$$\arg\max_{\lambda,\beta} \sum_{i=1}^{N} \left\{ \hat{R}_{n_i} - \lambda - \beta\left(\frac{n_i}{n_N} - \tau\right) \right\}^2 \tilde{k}\left(\frac{\frac{n_i}{n_N}-\tau}{\tilde{h}}\right),$$

where $\tilde{k}(\cdot)$ is the kernel and $\tilde{h}$ is the relevant bandwidth. This leads to the local linear estimator $\hat{\sigma}_\xi^2(\tau) = \hat{\lambda}$ for $\sigma_\xi^2(\tau)$.

Theorem 4 *Under assumptions 1, 2, 3, 4, 5 as $N \to \infty$*

$$\hat{\sigma}_\xi^2(\tau) - \sigma_\xi^2(\tau) = O_p\left(\frac{1}{n_N^{1/2+2\delta}\tilde{h}^{1/2}} + \frac{\tilde{h}^2}{n_N^{2\delta}}\right) + o_p\left(\frac{h^2+\tilde{h}^2}{n_N^{\delta}}\right).$$

To estimate ρ_{gm}, we apply the difference-based method proposed in Hart (1989, 1991). Define the centred second differences $\Delta_{n_i} = Y_{n_{i+1}} - 2Y_{n_i} + Y_{n_{i-1}}$, where $i = 1,\ldots,N-1$. By assumption, f is twice continuously differentiable on a compact interval, so that

$$\begin{aligned}\Delta_{n_i} &= \frac{1}{n_N^2} f''\left(\frac{n_i}{n_N}\right)\{1+o(1)\} + \sigma_\xi\left(\frac{n_i}{n_N}\right)(\xi_{n_{i+1}} - 2\xi_{n_i} + \xi_{n_{i-1}}) \\ &+ \left(\sigma_\xi\left(\frac{n_{i+1}}{n_N}\right) - \sigma_\xi\left(\frac{n_i}{n_N}\right)\right)\xi_{n_{i+1}} + \left(\sigma_\xi\left(\frac{n_{i-1}}{n_N}\right) - \sigma_\xi\left(\frac{n_i}{n_N}\right)\right)\xi_{n_{i-11}} \\ &= \sigma_\xi\left(\frac{n_i}{n_N}\right)(\xi_{n_{i+1}} - 2\xi_{n_i} + \xi_{n_{i-1}}) + O_p\left(\frac{1}{n_N^{1+\delta}} + \frac{1}{n_N^2}\right),\end{aligned}$$

which leads to

$$d_{n_i} \equiv \xi_{n_{i+1}} - 2\xi_{n_i} + \xi_{n_{i-1}} = \frac{\Delta_{n_i}}{\sigma_\xi(\frac{n_i}{n_N})} + O_p\left(\frac{1}{n_N} + \frac{1}{n_N^{2-\delta}}\right).$$

Let S_d and S_ξ denote the spectra of the processes $\{d_{n_i}\}$ and $\{\xi_{n_i}\}$, respectively, with $i = 1,\ldots,N$. Then $S_d(\omega) = |1-e^{ig\omega}|^4 S_\xi(\omega)$, where $\omega \in [-\pi,\pi]$. S_ξ can be consistently estimated from d_{n_i}, which, in turn, can be consistently estimated from

$$\hat{d}_{n_i} = \frac{\Delta_{n_i}}{\hat{\sigma}_\xi(\frac{n_i}{n_N})},$$

where $\hat{\sigma}_\xi(\frac{n_i}{n_N})$ is the consistent estimator of $\sigma_\xi(\frac{n_i}{n_N})$.

As in Hart (1989), we define the periodogram of a tapered $\hat{d}_{n_i}$ by

$$\hat{I}_d(\omega) = \frac{1}{T_n}\left|\sum_{j=1}^{N-1} t\left(\frac{n_j}{n_N}\right)\hat{d}_{n_j} e^{-i\omega g j}\right|^2,$$

where $\omega \in [-\pi,\pi]$, $t(\cdot)$ is a tapering function that vanishes at 0 and 1, and $T_n = 2\pi\sum_{j=1}^{N-1} t^2(\frac{n_j}{n_N})$. The estimator of ρ_{gm} is

$$\hat{\rho}_{gm} = \frac{4\pi}{N}\sum_{j=\lfloor\frac{N\kappa}{2\pi}\rfloor}^{\lfloor\frac{N}{2}\rfloor} \cos(\omega_j gm)|1-e^{ig\omega_j}|^{-4}\hat{I}_d(\omega_j),$$

where κ is some small positive number and $\omega_j = \frac{2\pi j}{N}$. When $N \to \infty$, $\kappa \to 0$, and $N\kappa \to \infty$, Hart (1989) shows that $\hat{\rho}_{gm} \xrightarrow{p} \rho_{gm}$. With this consistent estimate of ρ_{gm} we can employ a Newey-West type consistent estimator for $\sigma_\xi^2(\tau)\,[1+2\sum_{m=1}^{\infty}\rho_{gm}]$, viz.,

$$\hat{\sigma}_{\xi}^{2}(\tau)\left[1+2\sum_{m=1}^{M}\left(\frac{M-m}{M}\right)\hat{\rho}_{gm}\right].$$

Remark 4: The reason for using a difference-based rather than residual-based method is that from Theorem 1 $f(\frac{n_i}{n_N})-\hat{f}(\frac{n_i}{n_N})=O_p\left(h^2+\frac{1}{n_N^{\frac{1}{2}+\delta}h^{\frac{1}{2}}}\right)$ whereas the order of σ_ξ is $\frac{1}{n_N^\delta}$, which may lead to inconsistent estimation of ρ_{gm}. In particular, if we use the residual-based method,

$$\begin{aligned}\hat{\xi}_{n_i} &= \frac{Y_{n_i}-\hat{f}(\frac{n_i}{n_N})}{\hat{\sigma}_\xi(\frac{n_i}{n_N})}\\ &= \frac{f(\frac{n_i}{n_N})-\hat{f}(\frac{n_i}{n_N})+\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}}{\hat{\sigma}_\xi(\frac{n_i}{n_N})}\\ &\sim \frac{f(\frac{n_i}{n_N})-\hat{f}(\frac{n_i}{n_N})}{\sigma_\xi(\frac{n_i}{n_N})}+\xi_{n_i}\\ &= O_p\left(n_N^\delta h^2+\frac{1}{n_N^{\frac{1}{2}}h^{\frac{1}{2}}}\right)+\xi_{n_i}.\end{aligned}$$

The difference between $\hat{\xi}_{n_i}$ and ξ_{n_i} has order $n_N^\delta h^2+n_N^{-\frac{1}{2}}h^{-\frac{1}{2}}$, which, depending on δ and h, may not go to zero as $N\to\infty$.

Remark 5: In Hart (1989, 1991) only the stationary case is discussed. Our application allows for a varying conditional variance.

As Altman (1990) and Hart (1991) illustrated, automated bandwidth selection methods, such as cross validation, can perform poorly when dealing with positive correlated data. Following Rice (1984) and Hart (1991), we define the mean average-squared error (MASE) by

$$
\begin{aligned}
M(h) &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\frac{n_i}{n_N})-f(\frac{n_i}{n_N})\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\frac{n_i}{n_N})-Y_{n_i}+\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}(\frac{n_i}{n_N})-Y_{n_i}\right)^2\right]+\frac{1}{N}\sum_{i=1}^{N}\sigma_\xi^2(\frac{n_i}{n_N}) \\
&\quad +\frac{2}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{f}(\frac{n_i}{n_N})-Y_{n_i}\right)\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}\right],
\end{aligned}
$$

where the cross product term

$$
\begin{aligned}
\mathbb{E}\left[\left(\hat{f}(\frac{n_i}{n_N})-Y_{n_i}\right)\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}\right] &= \mathbb{E}\left[\left(\frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\left(Y_{n_j}-Y_{n_i}\right)\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})}\right)\right] \\
&= \mathbb{E}\left[\left(\frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\left(\sigma_\xi(\frac{n_j}{n_N})\xi_{n_j}-\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}\right)\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})}\right)\right] \\
&= \mathbb{E}\left[\left(\frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\sigma_\xi(\frac{n_j}{n_N})\sigma_\xi(\frac{n_i}{n_N})\xi_{n_j}\xi_{n_i}}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})}\right)\right]-\sigma_\xi^2(\frac{n_i}{n_N}) \\
&= \frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\sigma_\xi(\frac{n_j}{n_N})\sigma_\xi(\frac{n_i}{n_N})\mathbb{E}\left[\xi_{n_j}\xi_{n_i}\right]}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})}-\sigma_\xi^2(\frac{n_i}{n_N}) \\
&= \frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\sigma_\xi(\frac{n_j}{n_N})\sigma_\xi(\frac{n_i}{n_N})\rho_{|n_j-n_i|}}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})}-\sigma_\xi^2(\frac{n_i}{n_N}).
\end{aligned}
$$

As in the proof of Theorem 1, denote

$$
\omega_j(\frac{n_i}{n_N})=k(\frac{n_j-n_i}{n_Nh})\left(S_{\frac{n_i}{n_N},2}-S_{\frac{n_i}{n_N},1}(\frac{n_j}{n_N}-\frac{n_i}{n_N})\right),
$$

and

$$
S_{\frac{n_i}{n_N},j}=\sum_{i=1}^{N}k\left(\frac{n_j-n_i}{n_Nh}\right)\left(\frac{n_j}{n_N}-\frac{n_i}{n_N}\right)^j.
$$

Then,

$$\begin{aligned} M(h) &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}\left(\frac{n_i}{n_N}\right)-Y_{n_i}\right)^2\right]-\frac{1}{N}\sum_{i=1}^{N}\sigma_{\xi}^{2}\left(\frac{n_i}{n_N}\right) \\ &+\frac{2}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\sigma_{\xi}(\frac{n_j}{n_N})\sigma_{\xi}(\frac{n_i}{n_N})\rho_{|n_j-n_i|}}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})}. \end{aligned}$$

An unbiased estimator estimate of $M(h)$ is therefore

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}\left(\frac{n_i}{n_N}\right)-Y_{n_i}\right)^2-\frac{1}{N}\sum_{i=1}^{N}\sigma_{\xi}^{2}\left(\frac{n_i}{n_N}\right)+\frac{2}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\sigma_{\xi}(\frac{n_j}{n_N})\sigma_{\xi}(\frac{n_i}{n_N})\rho_{|n_j-n_i|}}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})},$$

which can be obtained by

$$\hat{M}(h)=\frac{1}{N}RSS(h)-\frac{1}{N}\sum_{i=1}^{N}\hat{\sigma}_{\xi}^{2}\left(\frac{n_i}{n_N}\right)+\frac{2}{N}\sum_{i=1}^{N}\frac{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})\hat{\sigma}_{\xi}(\frac{n_j}{n_N})\hat{\sigma}_{\xi}(\frac{n_i}{n_N})\hat{\rho}_{|n_j-n_i|}}{\sum_{j=1}^{N}\omega_j(\frac{n_i}{n_N})},$$

where

$$RSS(h)=\sum_{i=1}^{N}\left(\hat{f}\left(\frac{n_i}{n_N}\right)-Y_{n_i}\right)^2.$$

The optimal h then minimizes the criterion $\hat{M}(h)$.

4.4 Conclusion

The presence of market microstructure noise imposes many challenges for the estimation of integrated variance using high frequency financial data. Under pure noise, the realized variance is a linear function of grid size, giving a linear microstructure noise function that can be eliminated by suitable grid selection, thereby facilitating the estimation of integrated variance. But when the pure noise assumption fails, microstructure noise may produce nonlinear noise function effects on realized variance and different methods are needed to eliminate noise effects. To accommodate such microstructure noise, this paper develops new methods that allow for a nonparametric specification within a panel model context that utilizes daily information over different trading days to assist the estimation of the microstructure noise impact, to test the form of the microstructure noise effects, and to estimate integrated variance in the presence of general microstructure noise. The approach accommodates nonparametric functional forms for the mean and the variance. Consistency and asymptotic normality are established under strong mixing conditions and a data-determined approach to bandwidth selection is suggested.

Chapter 5 Summary of Conclusions

Chapter 2 provides a new methodology for constructing real estate price indices. The proposed new model enjoys some advantages of the standard hedonic method as it uses both single-sales and repeat-sales data but it is less prone to specification bias than the standard hedonic model. Moreover, it generalizes the attractive feature of the repeat-sales method by creating sale pairs within the individual building level, thereby increasing the number of observations used in the index. The model is also easy to estimate, since this approach uses GLS estimation and is computationally efficient with large datasets. Other methods have been suggested to construct sale pairs in the literature – see, for example, McMillen (2012), and Guo et al (2014). Our matching rule is simpler to implement and has the advantage of a semiparametric nature.

We apply our estimation procedure to the real estate market for private residential dwellings in Singapore and examine the model's out-of-sample predictive performance in comparison with indices produced using the repeat-sales methodology of Case and Shiller (1987, 1989) and the standard hedonic method. The findings reveal that, compared with these alternative methodologies, our method has superior performance out-of-sample.

The recursive detection method of Phillips, Shi and Yu (2015a, 2015b) is applied to each of the indices to locate episodes of real estate price exuberance in Singapore. While for all three indices PSY identifies the same bubble, the bubble origination date in the new index comes two quarters earlier than that in the other two indices. Although all three indices grew during 2009 - 2013, the expansion is not explosive, indicating that the ten recent rounds of cooling measure intervention in the real estate market conducted by the Singapore government have been successful in controlling prices.

Chapter 3 investigates the finite sample problem in the estimation of structural break points in several aspects. First we derive the finite sample distribution of the structural break estimator in the continuous time model. We then establish its connection to the discrete time models considered in the literature. It is shown that when the true break point is at the middle of the sample, the finite sample distribution is symmetric but can have tri-modality. However, when the true break point occurs earlier than the middle of the sample, the finite sample distribution is skewed to the right and there is a positive bias. When the true break point occurs later than the middle of the sample, the finite sample distribution is skewed to the left and there is a negative bias.

To reduce the bias in finite sample, we obtain the binding functions via simulations and then use the indirect estimation technique to estimate the break parameter. Monte Carlo results show that the indirect estimation procedure is effective in reducing the bias of the traditional break point estimators. But the variance of the indirect estimator is larger than that of the original estimator, since the binding function has a slope less than one.

Chapter 4 examines the estimation of integrated variance and the microstructure noise function using high frequency data. The presence of market microstructure noise imposes many challenges for the estimation of integrated variance using high frequency financial data. Under pure noise, the realized variance is a linear function of grid size, giving a linear microstructure noise function that can be eliminated by suitable grid selection, thereby facilitating the estimation of integrated variance. But when the pure noise assumption fails, microstructure noise may produce nonlinear noise function effects on realized variance and different methods are needed to eliminate noise effects. To accommodate such microstructure noise, this chapter develops new methods that allow for a nonparametric specification within a panel model context that utilizes daily information over different trading days to assist the estimation of the microstructure noise impact, to test the form of the microstructure noise effects, and to estimate integrated variance in the presence of general microstructure noise. The approach accommodates nonparametric functional forms for the mean and the variance. Consistency and asymptotic normality are established under strong mixing conditions and a data-determined approach to bandwidth selection is suggested.

Bibliography

[1] Aït-Sahalia, Y., Mykland, P. A., and Zhang, L., 2005. How often to sample a continuous-time process in the presence of market microstructure noise. Review of Financial studies, 18(2), 351-416.

[2] Altman, N. S., 1990. Kernel smoothing of data with correlated errors. Journal of the American Statistical Association, 85(411), 749-759.

[3] Andreou, E., and Ghysels, E., 2009. Structural breaks in financial time series. In: Handbook of Financial Time Series. Springer Berlin Heidelberg, pp. 839-870

[4] Arvanitis, S. and Demos, A., 2014, On the Validity of Edgeworth Expansions and Moment Approximations for Three Indirect Inference Estimators, Working Paper, Athens University of Economics and Business.

[5] Bai, J., 1994. Least squares estimation of a shift in linear processes. Journal of Time Series Analysis, 15, 453-472.

[6] Bai, J., 1995. Least absolute deviation estimation of a shift. Econometric Theory, 11, 403-436.

[7] Bai, J., 1997a. Estimating multiple breaks one at a time. Econometric Theory, 13, 315-352.

[8] Bai, J., 1997b. Estimation of a change point in multiple regression models. Review of Economics and Statistics, 79, 551-563.

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

[10] Bai, J., and Perron, P., 1998. Estimating and testing linear models with multiple structural breaks. Econometrica, 66, 47-78.

[11] Bai, J., Lumsdaine, R. L., and Stock, J. H., 1998. Testing for and dating common breaks in multivariate time series. The Review of Economic Studies, 65, 395-432.

[12] Bailey, M. J., Muth, R. F., and Nourse, H. O., 1963. A regression method for real estate price index construction. Journal of the American Statistical Association, 58(304), 933-942.

[13] Baltagi, B. and Li, J., 2015. "Cointegration of Matched Home Purchases and Rental Price Indexes – Evidence from Singapore". Working Paper, Singapore Management University.

[14] Bandi, F. M., and Russell, J. R., 2008. Microstructure noise, realized variance, and optimal sampling. The Review of Economic Studies, 75(2), 339-369.

[15] Bao, Y., and A. Ullah, 2007, The second-order bias and mean squared error of estimators in time-series models, Journal of Econometrics, 140(2), 650-669.

[16] Barndorff-Nielsen, O. E., 2002. Econometric analysis of realized volatility and its use in estimating stochastic volatility models. Journal of the Royal Statistical Society: Series B, 64(2), 253-280.

[17] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., and Shephard, N., 2008. Designing realized kernels to measure the ex post variation of equity prices in the presence of noise. Econometrica, 76 (6), 1481-1536.

[18] Bhattacharya, P. K., 1987. Maximum likelihood estimation of a change-point in the distribution of independent random variables: general multiparameter case. Journal of Multivariate Analysis, 23, 183-208.

[19] Bhattacharya, P. K., 1994. Some aspects of change-point analysis. Lecture Notes-Monograph Series, 28-56.

[20] Bhattacharya, P. K, .and Brockwell, P. J., 1976. The minimum of an additive process with applications to signal estimation and storage theory. Z. Wahrsch. verw. Gebiete 37, 51-75.

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

[22] Case, B., H. O. Pollakowski, and S. M. Wachter., 1991. On choosing among house price index methodologies. Real Estate Economics, 19 (3), 286-307.

[23] Case, B., and J. M. Quigley., 1991. The dynamics of real estate prices. The Review of Economics and Statistics, 73 (1), 50-58.

[24] Case, K. E., and R. Shiller., 1987. Prices of single-family homes since 1970: New indexes for four cities. New England Economic Review, 45-56.

[25] Case, K. E., and R. J. Shiller., 1989. The Efficiency of the Market for Single-Family Homes. The American Economic Review, 79 (1), 125-137.

[26] Chan, K. F., S. Treepongkaruna, R. Brooks, and S. Gray., 2011. Asset market linkages: Evidence from financial, commodity and real estate assets. Journal of Banking and Finance, 35 (6), 1415-1426.

[27] Chen, J. and Gupta, A.K., 2011, Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance, Birkhuser.

[28] Chernoff, H., and S. Zacks, 1964, Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time Series, The Annals of Mathematical Statistic, 35(3), 999-1018.

[29] Clapp, J. M., C. Giaccotto, and D. Tirtiroglu., 1991. “Housing price indices based on all transactions compared to repeat subsamples. Real Estate Economics, 19 (3), 270-285.

[30] Csörgö M. and Horvth, L., 1997, Limit theorems in change-point analysis, Wiley.

[31] Diba, Behzad T., and Herschel I. Grossman., 1988. Explosive rational bubbles in stock prices? The American Economic Review, 520-530.

[32] Dong, Y., and Tse, Y., 2015. On Estimating Market Microstructure Noise Variance, Working paper, Singapore Management University.

[33] Englund, P., J. M. Quigley, and C. L. Redfearn., 1998. Improved price indexes for real estate: measuring the course of Swedish housing prices. Journal of Urban Economics, 44 (2), 171-196.

[34] Gallant, A.R., Tauchen, G., 1996. Which moments to match? Econometric Theory, 12, 657–681.

[35] Fan, J., and Yao, Q., 1998. Efficient estimation of conditional variance functions in stochastic regression. Biometrika, 85(3), 645-660.

[36] Gatzlaff, D. H., and D. R. Haurin., 1997. "Sample selection bias and repeat-sales index estimates. The Journal of Real Estate Finance and Economics, 14 (1), 33-50.

[37] Ghysels, E., A. Plazzi, W. N. Torous, and R. Valkanov., 2012. Forecasting real estate prices. Handbook of Economic Forecasting 2.

[38] Gourieroux, C., Monfort, A., and Renault, E., 1993. Indirect estimation. Journal of Applied Econometrics, 8, 85-118.

[39] Gourieroux, C., Phillips, P. C., and Yu, J., 2010. Indirect estimation for dynamic panel models. Journal of Econometrics, 157, 68-77.

[40] Gourieroux, C., Renault, E., Touzi, N., 2000. Calibration by simulation for small sample bias correction. In: Mariano, R.S., Schuermann, T., Weeks, M. (Eds.), Simulation-Based Inference in Econometrics: Methods and Applications. Cambridge University Press, pp. 328–358.

[41] Guo, X., Zheng, S., Geltner, D., and Liu, H., 2014. A new approach for constructing home price indices: The pseudo repeat sales model and its application in China. Journal of Housing Economics, 24, 20-38.

[42] Hansen, B. E., 2001. The new econometrics of structural break: Dating breaks in US labor productivity. Journal of Economic Perspectives, 117-128.

[43] Hansen, P. R., 2010. "A winner's curse for econometric models: on the joint distribution of in-sample fit and out-of-sample fit and its implications for model selection." Manuscript Department of Economics, Stanford University.

[44] Hansen, P. R., and Lunde, A., 2006. Realized variance and market microstructure noise. Journal of Business & Economic Statistics, 24(2), 127-161.

[45] Hardle, W., and Linton, O., 1994. Applied nonparametric methods. Handbook of econometrics, 4, 2295-2339.

[46] Hart, J. D., 1989. Differencing as an approximate de-trending device. Stochastic Processes and their Applications, 31(2), 251-259.

[47] Hart, J. D., 1991. Kernel regression estimation with time series errors. Journal of the Royal Statistical Society. Series B (Methodological), 53 (1), 173-187.

[48] Hill, R. Carter, John R. Knight, and C. F. Sirmans., 1997. Estimating capital asset price indexes. Review of Economics and Statistics, 79 (2), 226-233.

[49] Hawkins, D. L., Gallant, A. R., and Fuller, W., 1986. A simple least squares method for estimating a change in mean. Communications in Statistics-Simulation and Computation, 15, 523-530.

[50] Hinkley, D. V., 1969. Inference about intersection in the two-phase regression problem. Biometrika, 56, 495-504.

[51] Hinkley, D. V., 1970. Inference about the change-point in a sequence of random variables. Biometrika, 57, 1-17.

[52] Iacoviello, Matteo., 2005. House prices, borrowing constraints, and monetary policy in the business cycle. American Economic Review, 739-764.

[53] Ibragimov, I. A., and Has'minskii, R. Z., 1981. Statistical Estimation. Springer.

[54] Kendall, M. G., 1954. Note on bias in the estimation of autocorrelation. Biometrika, 41, 403-404.

[55] Koetter, Michael, and Tigran Poghosyan., 2010. Real estate prices and bank stability. Journal of Banking and Finance, 34 (6), 1129-1138.

[56] Le Cam, L., 1960. Locally asymptotically normal families of distributions. University of California Publications in Statistics, 3, 37–98.

[57] MacKinnon, J. G., and Smith Jr, A. A., 1998. Approximate bias correction in econometrics. Journal of Econometrics, 85, 205-230.

[58] McMillen, D.P., 2012. Repeat sales as a matching estimator. Real Estate Economics, 40 (4) 745–772.

[59] Mendicino, Caterina, and Maria Teresa Punzi., 2014. House prices, capital inflows and macroprudential policy. Journal of Banking and Finance, 49, 337–355.

[60] Musso, Alberto, Stefano Neri, and Livio Stracca., 2011. Housing, consumption and monetary policy: How different are the US and the euro area? Journal of Banking and Finance, 35 (11), 3019-3041.

[61] Nagaraja, C. H., L. D. Brown, and S. M. Wachter., 2010. "House price index methodology." Working Paper, University of Pennsylvania.

[62] Nagaraja, C. H., L. D. Brown, and L. H. Zhao., 2011. An autoregressive approach to house price modeling. The Annals of Applied Statistics, 5 (1), 124-149.

[63] Nickell, S., Biases in dynamic models with fixed effects. Econometrica, 49, 1417–1426.

[64] Perron, P., 1989. The great crash, the oil price shock, and the unit root hypothesis. Econometrica, 57, 1361-1401.

[65] Perron, P., 2006. Dealing with structural breaks. Palgrave Handbook of Econometrics 1, pp. 278-352.

[66] Phillips, P. C., 2012. Folklore theorems, implicit maps, and indirect estimation. Econometrica 80, 425-454.

[67] Phillips, P. C. B., S. Shi, and J. Yu., 2015a. Testing for multiple bubbles: Historical episodes of exuberance and collapse in the S&P 500. International Economic Review, forthcoming.

[68] Phillips, P. C. B., S. Shi, and J. Yu., 2015b. Testing for Multiple Bubbles: Limit Theory of Real Time Detector, International Economic Review, forthcoming.

[69] Phillips, P. C. B., Wu, Y., and J. Yu., 2011. Explosive behavior in the 1990s Nasdaq: When did exuberance escalate asset values? International Economic Review, 52, 201–226.

[70] Phillips, P. C. B., and Yu, J., 2006. Comment, Journal of Business & Economic Statistics, 24 (2), 202-208.

[71] Phillips, P.C.B., and Yu, J., 2009. Maximum likelihood and gaussian estimation of continuous time models in finance. In: Handbook of Financial Time Series, 707–742.

[72] Phillips, P. C., and Yu, J., 2009. Simulation-based estimation of contingent-claims prices. Review of Financial Studies, 22, 3669-3705.

[73] Quigley, J. M., 1995. A simple hybrid model for estimating real estate price indexes. Journal of Housing Economic, 4 (1), 1-12.

[74] Reinhart, Carmen M., and Kenneth S. Rogoff., 2013. Banking crises: An equal opportunity menace. Journal of Banking and Finance, 37 (11), 4557-4573.

[75] Rilstone, P., V.K. Srivastava and A. Ullah, 1996, The second order bias and mean squared error of nonlinear estimators, Journal of Econometrics, 72(2), 369–395.

[76] Robinson, P. M., 1989. Nonparametric estimation of time-varying parameters (pp. 253-264). Springer Berlin Heidelberg.

[77] Robinson, P. M., 2012. Nonparametric trending regression with cross-sectional dependence. Journal of Econometrics, 169(1), 4-14.

[78] Rice, J., 1984. Bandwidth choice for nonparametric regression. The Annals of Statistics, 12 (4), 1215-1230.

[79] Rosen, S., 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. Journal of Political Economy, 82 (1), 34-55.

[80] Shiller, R. J., 2008. Derivatives markets for home prices. No. w13962. National Bureau of Economic Research.

[81] Shi, Song, Jyh-Bang Jou, and David Tripe., 2014. Can interest rates really control house prices? Effectiveness and Implications for Macroprudential Policy. Journal of Banking and Finance, 47, 15–28.

[82] Sing, T., 2001. Dynamics of the condominium market in Singapore. International Real Estate Review, 4 (1), 135-158.

[83] Smith, A. A., 1993. Estimating nonlinear time series models using simulated vector autoregressions. Journal of Applied Econometrics, 8, 63-84.

[84] S&P/Case-Shiller Home Price Indices-Index Methodology. Nov. 2009, http://www.standardandpoors.com/.

[85] White, H., 2001. Asymptotic Theory for Econometricians (revised edition.) Academic Press. San Diego.

[86] Yao, Y. C., 1987. Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. The Annals of Statistics, 15 (3), 1321-1328.

[87] Yu, J., 2012, Bias in the Estimation of the Mean Reversion Parameter in Continuous Time Models, Journal of Econometrics, 169, 114-122

[88] Zhang, L., Mykland, P. A., and Aït-Sahalia, Y., 2005. A tale of two time scales. Journal of the American Statistical Association, 100 (472), 1394-1411.

[89] Zhang, Y., Su, L., and Phillips, P. C., 2012. Testing for common trends in semiparametric panel data models with fixed effects. The Econometrics Journal, 15(1), 56-100.

[90] Zhou, B., 1996. High-frequency data and volatility in foreign-exchange rates. Journal of Business & Economic Statistics, 14(1), 45-52.

Appendix

.1 Date and the content of recent real estate market cooling measures introduced in Singapore

1. 2009/9/14

   - Reinstatement of the confirmed list for the 1st half 2010 government land sales programme.
   - Removal of the interest absorption scheme and interest-only housing loans.
   - Non-extension of the January 2009 budget assistance measures for the property market.

2. 2010/2/20

   - Introduction of a Seller's Stamp Duty (SSD) on all residential properties and lands sold within one year of purchase.
   - Loan-to-Value (LTV) limit lowered from 90% to 80% for all housing loans.

3. 2010/8/30

   - Holding period for imposition of SSD increased from one year to three years.

- Minimum cash payment increased from 5% to 10% and the LTV limit decreased to 70% for buyers with one or more outstanding housing loans.
- The extended SSD does not affect HDB lessees as the required Minimum Occupation Period for HDB flats is at least 3 years.

4. 2011/1/14

   - Increase the holding period for imposition of SSD from three years to four years.
   - Raise SSD rates to 16%, 12%, 8% and 4% for residential properties sold in the first, second, third and fourth year of purchase respectively.
   - Lower the LTV limit to 50% on housing loans for property purchasers who are not individuals.
   - Lower the LTV limit on housing loans from 70% to 60% for second property.

5. 2011/12/8

   - Introduction of an Additional Buyer's Stamp Duty (ABSD).
   - Developers purchasing more than four residential units and following through on intention to develop residential properties for sale would be waived ABSD.

6. 2012/10/6

   - Mortgage tenures capped at a maximum of 35 years.
   - For loans longer than 30 years or for loans that extend beyond retirement age of 65 years: LTV lowered to 60% for first mortgage and to 40% for second and subsequent mortgages.
   - LTV for non-individuals lowered to 40.%

7. 2013/1/12

    - Higher ABSD rates.
    - Decrease the LTV limit for second/third loan to 50/40% from 60%; non-individuals' LTV to 20% from 40.%
    - Mortgage Servicing Ratio (MSR) for HDB loans now capped at 35% of gross monthly income (from 40%); MSR for loans from financial institutions capped at 30.%

8. 2013/6/28: Introduction of Total Debt Servicing Ratio (TDSR). The total monthly repayments of debt obligations should not exceed 60% of gross monthly income.

9. 2013/8/27

    - Singapore Permanent Resident (SPR) Households need to wait three years, before they can buy a resale HDB flat.
    - Maximum tenure for HDB housing loans is reduced to 25 years. The MSR limit is reduced to 30% of the borrower's gross monthly income.
    - Maximum tenure of new housing loans and re-financing facilities for the purchase of HDB flats is reduced to 30 years. New loans with tenure exceeding 25 years and up to 30 years will be subject to tighter LTV limits.

10. 2013/12/9

    - Reduction of cancellation fees From 20% to 5% for executive condominiums.
    - Resale levy for second-timer applicants.
    - Revision of mortgage loan terms. Decrease MSR from 60% to 30% of a borrower's gross monthly income.

.2 Proofs in chapter 4

.2.1 Proof of theorem 1

Letting

$$S_{\tau,j} = \sum_{i=1}^{N} k\left(\frac{n_i - \tau n_N}{n_N h}\right)\left(\frac{n_i}{n_N} - \tau\right)^{j},$$

we have

$$\begin{aligned}
\hat{f}(\tau) &= \iota' \begin{pmatrix} S_{\tau,0} & S_{\tau,1} \\ S_{\tau,1} & S_{\tau,2} \end{pmatrix}^{-1} \left(X_\tau' W_\tau Y\right) \\
&= \iota' \begin{pmatrix} S_{\tau,0} & S_{\tau,1} \\ S_{\tau,1} & S_{\tau,2} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} k(\frac{n_i - n_N\tau}{n_N h}) Y_{n_i} \\ \sum_{i=1}^{N} k(\frac{n_i - n_N\tau}{n_N h}) \left(\frac{n_i}{n_N} - \tau\right) Y_{n_i} \end{pmatrix} \\
&= \left(S_{\tau,0}S_{\tau,2} - S_{\tau,1}^2\right)^{-1} \left[\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(S_{\tau,2} - S_{\tau,1}\left(\frac{n_i}{n_N} - \tau\right)\right) Y_{n_i}\right],
\end{aligned}$$

which can be rewritten as

$$\hat{f}(\tau) = \frac{\sum_{i=1}^{N} \omega_i Y_{n_i}}{\sum_{i=1}^{N} \omega_i}, \tag{.2.1}$$

where

$$\omega_i = k(\frac{n_i - n_N\tau}{n_N h})\left(S_{\tau,2} - S_{\tau,1}(\frac{n_i}{n_N} - \tau)\right).$$

By the definition of ω_i, we have

$$\sum_{i=1}^{N} \omega_i \left(\frac{n_i}{n_N} - \tau\right) = 0. \tag{.2.2}$$

By (.2.1) and (.2.2), the bias of $\hat{f}(\tau)$ is

$$\begin{aligned}
\mathbb{E}\left(\hat{f}(\tau)\right) - f(\tau) &= \frac{\sum_{i=1}^{N} \omega_i \left[f(\frac{n_i}{n_N}) - f(\tau) - f'(\tau)\left(\frac{n_i}{n_N} - \tau\right)\right]}{\sum_{i=1}^{N} \omega_i} \\
&= \frac{\sum_{i=1}^{N} \omega_i \left(\frac{n_i}{n_N} - \tau\right)^2 f''(\tau)}{2\sum_{i=1}^{N} \omega_i} + \frac{\sum_{i=1}^{N} \omega_i \left(\frac{n_i}{n_N} - \tau\right)^3 f'''(\zeta_i)}{6\sum_{i=1}^{N} \omega_i} \\
&= B_1 + B_2,
\end{aligned}$$

where ζ_i is between $\frac{n_i}{n_N}$ and τ, and

$$B_1 = \frac{\sum_{i=1}^{N} \omega_i \left(\frac{n_i}{n_N} - \tau \right)^2 f''(\tau)}{2 \sum_{i=1}^{N} \omega_i},$$

$$B_2 = \frac{\sum_{i=1}^{N} \omega_i \left(\frac{n_i}{n_N} - \tau \right)^3 f'''(\zeta_i)}{6 \sum_{i=1}^{N} \omega_i}.$$

By definition,

$$\begin{aligned} B_1 &= \frac{\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i}{n_N} - \tau\right)^2 f''(\tau) S_{\tau,2} - \sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i}{n_N} - \tau\right)^3 f''(\tau) S_{\tau,1}}{2\left(S_{\tau,0}S_{\tau,2} - S_{\tau,1}^2\right)} \\ &:= B_{11} - B_{12}, \end{aligned}$$

where

$$\begin{aligned} B_{11} &= \frac{\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i}{n_N} - \tau\right)^2 f''(\tau) S_{\tau,2}}{2\left(S_{\tau,0}S_{\tau,2} - S_{\tau,1}^2\right)} \\ &= h^2 \frac{f''(\tau)\frac{1}{n_N h}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i - n_N\tau}{n_N h}\right)^2 \frac{1}{n_N h^3} S_{\tau,2}}{2\left(\frac{S_{\tau,0}}{n_N h}\frac{S_{\tau,2}}{n_N h^3} - \frac{S_{\tau,1}^2}{n_N^2 h^4}\right)} \\ &\sim \frac{h^2 f''(\tau)\int_{-1}^{1} u^2 k(u) du}{2}, \end{aligned}$$

since

$$
\begin{aligned}
\left|\frac{S_{\tau,1}}{n_N h_1^2}\right| &= \left|\frac{S_{\tau,1}}{n_N h_1^2} - \int_{-1}^{1} uk(u)du\right| \\
&= \left|\frac{S_{\tau,1}}{n_N h_1^2} - \frac{1}{h}\int_0^1 k\left(\frac{u-\tau}{h}\right)\left(\frac{u-\tau}{h}\right)du\right| \qquad (.2.3) \\
&= \left|\frac{1}{h}\sum_{i=1}^{N}\int_{\frac{n_{i-1}}{n_N}}^{\frac{n_i}{n_N}} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i - n_N\tau}{n_N h}\right)du - \frac{1}{h}\sum_{i=1}^{N}\int_{\frac{n_{i-1}}{n_N}}^{\frac{n_i}{n_N}} k\left(\frac{u-\tau}{h}\right)\left(\frac{u-\tau}{h}\right)du\right| \\
&\leq \frac{1}{h}\sum_{i=1}^{N}\int_{\frac{n_{i-1}}{n_N}}^{\frac{n_i}{n_N}}\left|k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i - n_N\tau}{n_N h}\right) - k\left(\frac{u-\tau}{h}\right)\left(\frac{u-\tau}{h}\right)\right|du \\
&\leq \frac{1}{h}\sum_{i\in I}\int_{\frac{n_{i-1}}{n_N}}^{\frac{n_i}{n_N}} C\left|\frac{n_i - n_N u}{n_N h}\right|du \qquad (.2.4) \\
&\leq \frac{C}{h}\frac{1}{h}\sum_{i\in I}\int_{\frac{n_{i-1}}{n_N}}^{\frac{n_i}{n_N}} \max\left|\frac{n_i}{n_N} - u\right|du \\
&= \frac{C}{h}\frac{1}{h}\sum_{i\in I}\frac{1}{n_N}\frac{1}{n_N} \\
&= O(\frac{1}{n_N h}),
\end{aligned}
$$

and, similarly,

$$
\frac{1}{n_N h^{j+1}} S_{\tau,j} = \int_{-1}^{1} u^j k(u)du + O\left(\frac{1}{n_N h}\right) \qquad (.2.5)
$$

where (.2.3) is from the change of variable and (.2.4) is from the mean value theorem, and I denotes the set of observations with non-zero weights and $\#I = O(n_N h)$, where $\#I$ represents the cardinality.

Therefore, by (.2.5)

$$
\begin{aligned}
B_{12} &= \frac{\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i - n_N\tau}{n_N h}\right)^3 f''(\tau)S_{\tau,1}}{2\left(S_{\tau,0}S_{\tau,2} - S_{\tau,1}^2\right)} \\
&= h^2 \frac{f''(\tau)\frac{1}{n_N h}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i - n_N\tau}{n_N h}\right)^3 \frac{1}{n_N h^2}S_{\tau,1}}{2\left(\frac{S_{\tau,0}}{n_N h}\frac{S_{\tau,2}}{n_N h^3} - \frac{S_{\tau,1}^2}{n_N^2 h^4}\right)} \\
&= h^2 O\left(\frac{1}{n_N^2 h^2}\right) = O\left(\frac{1}{n_N^2}\right)
\end{aligned}
$$

since $\int_{-1}^{1} u^j k(u) = 0$ when j is odd and

$$B_1 \sim \frac{h^2 f''(\tau) \int_{-1}^{1} u^2 k(u) du}{2} + O\left(\frac{1}{n_N^2}\right). \tag{.2.6}$$

Similarly, it can be shown that

$$B_2 \sim h^3 O\left(\frac{1}{n_N h}\right) \sim O\left(\frac{h^2}{n_N}\right).$$

So

$$\mathbb{E}\left(\hat{f}(\tau)\right) - f(\tau) \sim \frac{h^2 f''(\tau) \int_{-1}^{1} u^2 k(u) du}{2} + O\left(\frac{1}{n_N^2} + \frac{h^2}{n_N}\right). \tag{.2.7}$$

To calculate the variance of $\hat{f}(\tau)$, note that

$$\begin{aligned} \hat{f}(\tau) - f(\tau) &= \frac{\sum_{i=1}^{N} \omega_i \left[f\left(\frac{n_i}{n_N}\right) - f(\tau)\right]}{\sum_{i=1}^{N} \omega_i} + \frac{\sum_{i=1}^{N} \omega_i \sigma_\xi(\frac{n_i}{n_N}) \xi_{n_i}}{\sum_{i=1}^{N} \omega_i} \\ &\sim \frac{h^2 f''(\tau) \int_{-1}^{1} u^2 k(u) du}{2} + \frac{\sum_{i=1}^{N} \omega_i \sigma_\xi(\frac{n_i}{n_N}) \xi_{n_i}}{\sum_{i=1}^{N} \omega_i} + o\left(\frac{1}{n_N}\right), \end{aligned}$$

by Eq. (.2.7).

Hence,

$$\begin{aligned} \hat{f}(\tau) - f(\tau) - \frac{h^2 f''(\tau) \int_{-1}^{1} u^2 k(u) du}{2} &\sim \frac{\sum_{i=1}^{N} \omega_i \sigma_\xi(\frac{n_i}{n_N}) \xi_{n_i}}{\sum_{i=1}^{N} \omega_i} + o(\frac{1}{n_N}) \\ &\sim \frac{\sum_{i=1}^{N} k(\frac{n_i - n_N \tau}{n_N h}) \left(S_{\tau,2} - S_{\tau,1}\left(\frac{n_i}{n_N} - \tau\right)\right) \sigma_\xi(\frac{n_i}{n_N}) \xi_{n_i}}{\left(S_{\tau,0} S_{\tau,2} - S_{\tau,1}^2\right)} \\ &\quad + o\left(\frac{1}{n_N}\right) \\ &\sim V_1 - V_2 + o\left(\frac{1}{n_N}\right), \end{aligned} \tag{.2.8}$$

where

$$V_1 = \frac{S_{\tau,2} \sum_{i=1}^{N} k\left(\frac{n_i - n_N \tau}{n_N h}\right) \sigma_\xi(\frac{n_i}{n_N}) \xi_{n_i}}{\left(S_{\tau,0} S_{\tau,2} - S_{\tau,1}^2\right)},$$

$$V_2 = \frac{S_{\tau,1}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left(\frac{n_i}{n_N} - \tau\right)\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i}}{\left(S_{\tau,0}S_{\tau,2} - S_{\tau,1}^2\right)}.$$

By Assumption 4,

$$\begin{aligned}
|V_2| &\leq \frac{S_{\tau,1}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left|\frac{n_i}{n_N} - \tau\right|\sigma_\xi(\frac{n_i}{n_N})\,\|\xi_{n_i}\|}{\left(S_{\tau,0}S_{\tau,2} - S_{\tau,1}^2\right)} \\
&\leq O\left(\frac{1}{n_N^\delta\sqrt{D}}\right)\frac{\frac{1}{n_N h^2}S_{\tau,1}\frac{1}{n_N h}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\left|\frac{n_i - n_N\tau}{n_N h}\right|}{\left(\frac{S_{\tau,0}}{n_N h}\frac{S_{\tau,2}}{n_N h^3} - \frac{S_{\tau,1}^2}{n_N^2 h^4}\right)} \\
&\sim O_p\left(\frac{1}{n_N^{1+\delta}h\sqrt{D}}\right) \qquad (.2.9)
\end{aligned}$$

since $\xi_{n_i} = O_p(1)$ by Assumption 4, $\frac{1}{n_N h^2}S_{\tau,1} = O(\frac{1}{n_N h})$ and $\int_{-1}^{1}|u|k(u)du$ is bounded.

The dominant term is V_1. To see this, note that $\mathbb{E}(V_1) = 0$ and

$$\begin{aligned}
V_1 &= \frac{S_{\tau,2}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\frac{\sigma(n_i/n_N)}{n_N^\delta\sqrt{D}}\xi_{n_i}}{\left(S_{\tau,0}S_{\tau,2} - S_{\tau,1}^2\right)} \\
&= \frac{\frac{1}{n_N h^3}S_{\tau,2}\frac{1}{n_N h}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}}{\left(\frac{S_{\tau,0}}{n_N h}\frac{S_{\tau,2}}{n_N h^3} - \frac{S_{\tau,1}^2}{n_N^2 h^4}\right)n_N^\delta\sqrt{D}} \\
&\sim \frac{\frac{1}{n_N h}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}}{n_N^\delta\sqrt{D}}. \qquad (.2.10)
\end{aligned}$$

So

$$\begin{aligned}
\mathbb{E}(V_1)^2 &\sim \mathbb{E}\left(\frac{\frac{1}{n_N h}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}}{n_N^\delta\sqrt{D}}\right)^2 \\
&= \frac{\frac{1}{n_N^2 h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)k\left(\frac{n_j - n_N\tau}{n_N h}\right)\sigma\left(\frac{n_i}{n_N}\right)\sigma\left(\frac{n_i}{n_N}\right)\mathbb{E}(\xi_{n_i}\xi_{n_j})}{n_N^{2\delta}D},
\end{aligned}$$

which leads to

$$\begin{aligned}
(n_N^{2\delta} D)\mathbb{E}(V_1)^2 &\sim \frac{1}{n_N^2 h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right) k\left(\frac{n_j - n_N\tau}{n_N h}\right) \sigma\left(\frac{n_i}{n_N}\right) \sigma\left(\frac{n_j}{n_N}\right) \mathbb{E}(\xi_{n_i}\xi_{n_j}) \\
&= \frac{1}{n_N^2 h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right) k\left(\frac{n_j - n_N\tau}{n_N h}\right) \sigma^2\left(\frac{n_i}{n_N}\right) \mathbb{E}(\xi_{n_i}\xi_{n_j}) \\
&\quad + \frac{1}{n_N^2 h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} \left\{ k\left(\frac{n_i - n_N\tau}{n_N h}\right) k\left(\frac{n_j - n_N\tau}{n_N h}\right) \sigma\left(\frac{n_i}{n_N}\right) \right. \\
&\quad \left. \times \left[\sigma\left(\frac{n_i}{n_N}\right) - \sigma\left(\frac{n_i}{n_N}\right)\right] \mathbb{E}(\xi_{n_i}\xi_{n_j}) \right\} \\
&= A + H,
\end{aligned}$$

where

$$\begin{aligned}
A &= \frac{1}{n_N^2 h^2}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)^2 \sigma^2\left(\frac{n_i}{n_N}\right) \mathbb{E}(\xi_{n_i}^2) \\
&\quad + \frac{2}{n_N^2 h^2}\sum_{i=1}^{N}\sum_{j>i}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right) k\left(\frac{n_j - n_N\tau}{n_N h}\right) \sigma^2\left(\frac{n_i}{n_N}\right) \mathbb{E}(\xi_{n_i}\xi_{n_j}) \\
&= A_1 + A_2.
\end{aligned}$$

Since $\mathbb{E}(\xi_{n_i}^2) = 1$ by Assumption 4, we have

$$\begin{aligned}
A_1 &= \frac{1}{n_N^2 h^2}\sum_{i=1}^{N} k(\frac{n_i - n_N\tau}{n_N h})^2 \sigma^2\left(\frac{n_i}{n_N}\right) \\
&\sim \frac{1}{n_N h}\int_{-1}^{1} k^2(u)\sigma^2(\tau + hu)du \sim \frac{\sigma^2(\tau)}{n_N h}\int_{-1}^{1} k^2(u)du. \qquad (.2.11)
\end{aligned}$$

For A_2, as in Robinson (1991) and Cai (2007), we define a sequence $\{b_i\}_{i=1}^{\infty}$ such that $b_N \to \infty$ with $\frac{b_N}{n_N h} \to 0$ as $N \to \infty$.

Then,

$$\begin{aligned}
A_2 &\sim \frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{1\le j-i\le \lfloor b_N/g\rfloor} k\left(\frac{n_i-n_N\tau}{n_N h}\right)k\left(\frac{n_j-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)\mathbb{E}(\xi_{n_i}\xi_{n_j})\\
&\quad+\frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{j-i>\lfloor b_N/g\rfloor} k\left(\frac{n_i-n_N\tau}{n_N h}\right)k\left(\frac{n_j-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)\mathbb{E}(\xi_{n_i}\xi_{n_j})\\
&= A_{21}+A_{22},
\end{aligned}$$

where, by the mixing inequality, $|E(\xi_{n_i}\xi_{n_j})|\le C\alpha_{|n_j-n_i|}^{1-2/\beta}$, we have

$$\begin{aligned}
|A_{22}| &\le \frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{j-i>\lfloor b_N/g\rfloor} k\left(\frac{n_i-n_N\tau}{n_N h}\right)k\left(\frac{n_j-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)C\alpha_{(n_j-n_i)}^{1-2/\beta}\\
&\le \frac{2C}{n_N^2 h^2}\sum_{i=1}^{N-1} k\left(\frac{n_i-n_N\tau}{n_N h}\right)\sum_{m>\lfloor b_N/g\rfloor}\alpha_{|m|}^{1-2/\beta}\sim o\left(\frac{1}{n_N h}\right),
\end{aligned}$$

where the last inequality is from the mixing condition that $\sum_{m=1}^{\infty}\alpha_m^{1-\frac{2}{\beta}}<\infty$ and the boundness of k and σ.

Denote

$$\begin{aligned}
A_{21} &\sim \frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{1\le j-i\le \lfloor b_N/g\rfloor} k^2\left(\frac{n_i-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)\mathbb{E}(\xi_{n_i}\xi_{n_j})\\
&\quad+\frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{1\le j-i\le \lfloor b_N/g\rfloor}\left\{k\left(\frac{n_i-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)\right.\\
&\quad\left.\times\left[k(\frac{n_j-n_N\tau}{n_N h})-k(\frac{n_i-n_N\tau}{n_N h})\right]\mathbb{E}(\xi_{n_i}\xi_{n_j})\right\}\\
&= A_{21}^a+A_{21}^b,
\end{aligned}$$

where

$$\begin{aligned}
|A_{21}^{b}| &\leq \frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{1\leq j-i\leq \lfloor b_N/g\rfloor}\left\{k\left(\frac{n_i-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)\right.\\
&\qquad \left.\times\left|k\left(\frac{n_j-n_N\tau}{n_N h}\right)-k\left(\frac{n_i-n_N\tau}{n_N h}\right)\right||\mathbb{E}(\xi_{n_i}\xi_{n_j})|\right\}\\
&\leq \frac{C}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{m=1}^{\lfloor b_N/g\rfloor}k\left(\frac{n_i-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)\left|\frac{m}{n_N h}\right|\alpha_m^{1-2/\beta}\\
&\leq \frac{C}{n_N^2 h^2}\sum_{i=1}^{N-1}k\left(\frac{n_i-n_N\tau}{n_N h}\right)\sigma^2\left(\frac{n_i}{n_N}\right)\sum_{m=1}^{\lfloor b_N/g\rfloor}\alpha_m^{1-2/\beta}\left|\frac{\lfloor b_N/g\rfloor}{n_N h}\right|\\
&\sim \frac{C}{n_N h}\int_{-1}^{1}k(u)\sigma^2(\tau+uh)du\sum_{m=1}^{\lfloor b_N/g\rfloor}\alpha_m^{1-2/\beta}\left|\frac{\lfloor b_N/g\rfloor}{n_N h}\right|\sim o(\frac{1}{n_N h}),
\end{aligned}$$

since $\sum_{m=1}^{\infty}\alpha_m^{1-2/\beta}<\infty$ and $\lfloor b_N/g\rfloor/(n_N h)\to 0$.

Meanwhile,

$$\begin{aligned}
A_{21}^{a} &= \frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{1\leq j-i\leq\lfloor b_N/g\rfloor}k^2(\frac{n_i-n_N\tau}{n_N h})\sigma^2\left(\frac{n_i}{n_N}\right)\mathbb{E}(\xi_{n_i}\xi_{n_j})\\
&= \frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}\sum_{m=1}^{\lfloor b_N/g\rfloor}k^2(\frac{n_i-n_N\tau}{n_N h})\sigma^2\left(\frac{n_i}{n_N}\right)\rho_{gm}\\
&= \frac{2}{n_N^2 h^2}\sum_{i=1}^{N-1}k^2(\frac{n_i-n_N\tau}{n_N h})\sigma^2\left(\frac{n_i}{n_N}\right)\sum_{m=1}^{\lfloor b_N/g\rfloor}\rho_{gm}\\
&= \frac{2}{n_N h}\int_{-1}^{1}k^2(u)\sigma^2(\tau+uh)du\sum_{m=1}^{\lfloor b_N/g\rfloor}\rho_{gm}\\
&\sim \frac{2\sigma^2(\tau)\sum_{m=1}^{\infty}\rho_{gm}\int_{-1}^{1}k^2(u)du}{n_N h}+o\left(\frac{1}{n_N h}\right). \qquad (.2.12)
\end{aligned}$$

In sum, by (.2.11) and (.2.12),

$$A\sim\frac{\sigma^2(\tau)\int_{-1}^{1}k^2(u)du}{n_N h}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right]+o\left(\frac{1}{n_N h}\right).$$

Furthermore,

$$\begin{aligned}
|H| &\le \frac{1}{n_N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\left\{k\left(\frac{n_j-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\right.\\
&\quad \left.\times\left|\sigma\left(\frac{n_j}{n_N}\right)-\sigma\left(\frac{n_i}{n_N}\right)\right| |\mathbb{E}(\xi_{n_i}\xi_{n_j})|\right\} \qquad (.2.13)\\
&\le \frac{C}{n_N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i-n_N\tau}{n_Nh}\right)k\left(\frac{n_j-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\left|\frac{n_j}{n_N}-\frac{n_i}{n_N}\right|\alpha_{|n_j-n_i|}^{1-2/\beta}\\
&\le \frac{C}{n_N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\left\{k\left(\frac{n_j-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\right.\\
&\quad \left.\times\left[\left|\frac{n_j}{n_N}-\tau\right|+\left|\frac{n_i}{n_N}-\tau\right|\right]\alpha_{|n_j-n_i|}^{1-2/\beta}\right\}\\
&\le \frac{2hC}{n_N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i-n_N\tau}{n_Nh}\right)k\left(\frac{n_j-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\alpha_{|n_j-n_i|}^{1-2/\beta} \qquad (.2.14)\\
&\le \frac{2hC}{n_N^2h^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\alpha_{|n_j-n_i|}^{1-2/\beta}\sim o\left(\frac{1}{n_Nh}\right),
\end{aligned}$$

where the inequality (.2.14) is due to Assumption 2. To have non-zero kernel weights, $-1 \le \frac{n_i-n_N\tau}{n_Nh} \le 1$, which implies that $\left|\frac{n_i}{n_N}-\tau\right| \le h$. The last inequality is from the boundness of k and σ.

Thus,

$$\mathbb{E}(V_1)^2 \sim \frac{\sigma^2(\tau)\int_{-1}^{1}k^2(u)du}{n_N^{1+2\delta}hD}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right]+o\left(\frac{1}{n_N^{1+2\delta}h}\right). \qquad (.2.15)$$

Finally, by (.2.8), (.2.9) and (.2.15),

$$var\left\{\hat{f}(\tau)-f(\tau)-\frac{h^2f''(\tau)\int_{-1}^{1}u^2k(u)du}{2}\right\}\sim\frac{\sigma^2(\tau)\int_{-1}^{1}k^2(u)du}{n_N^{1+2\delta}hD}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right],$$

so that we have

$$\begin{aligned}
MSE\left\{\hat{f}(\tau)\right\} &= \frac{h^4f''(\tau)^2\left\{\int_{-1}^{1}u^2k(u)du\right\}^2}{4}+\frac{\sigma^2(\tau)\int_{-1}^{1}k^2(u)du}{n_N^{1+2\delta}hD}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right]\\
&\quad +O\left(\frac{1}{n_N^2}+\frac{h^2}{n_N}\right)^2.
\end{aligned}$$

Q.E.D.

.2.2 Proof of theorem 2

From (.2.8), (.2.9), (.2.10) and (.2.15), we have

$$n_N^{\frac{1}{2}+\delta}h^{\frac{1}{2}}\left[\hat{f}(\tau)-f(\tau)-\frac{h^2f''(\tau)\int_{-1}^{1}u^2k(u)du}{2}\right]\sim\frac{1}{\sqrt{n_Nh}}\sum_{i=1}^{N}k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}.$$

Note that k has a compact support $[-1,1]$ and I denotes the set of observations with non-zero weights and $\#I=O(n_Nh)$, where $\#I$ represents the cardinality. So,

$$\begin{aligned}\frac{1}{\sqrt{n_Nh}}\sum_{i=1}^{N}k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i} &= \frac{1}{\sqrt{n_Nh}}\sum_{n_i\in I}k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}\\ &= \frac{\sqrt{\#I}}{\sqrt{n_Nh}}\frac{1}{\sqrt{\#I}}\sum_{n_i\in I}k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}.\end{aligned}$$

Denote $Z_{n_i}=k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}$, we have $\mathbb{E}(Z_{n_i})=0$ and

$$\begin{aligned}\bar{\sigma}_N^2 &= var\left[\frac{1}{\sqrt{\#I}}\sum_{n_i\in I}Z_{n_i}\right]\\ &= var\left[\frac{\sqrt{n_Nh}}{\sqrt{\#I}}\frac{1}{\sqrt{n_Nh}}\sum_{i=1}^{N}k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}\right]\\ &= \frac{n_Nh}{\#I}var\left[\frac{1}{\sqrt{n_Nh}}\sum_{i=1}^{N}k\left(\frac{n_i-n_N\tau}{n_Nh}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i}\right]\\ &\sim \frac{\sigma^2(\tau)\int_{-1}^{1}k^2(u)du}{CD}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right], \end{aligned}\tag{.2.16}$$

where C is the limit of $\frac{\#I}{n_Nh}$ and the last step is from the proof of Theorem 1.

Also, for some $\beta>2$,

$$\begin{aligned}\mathbb{E}|Z_{n_i}|^\beta &= \mathbb{E}\left|k(\frac{n_i-n_N\tau}{n_Nh})\sigma(\frac{n_i}{n_N})\xi_{n_i}\right|^\beta\\ &\leq C\mathbb{E}\,|\xi_{n_i}|^\beta\leq C\sup_i\mathbb{E}\,|\xi_{n_i}|^\beta<\infty,\end{aligned}\tag{.2.17}$$

where the first inequality is from the boundness of k and σ and the last inequality is due to Assumption 4.

By Theorem 5.20 of White (2001), (.2.16) and (.2.17) lead to

$$\frac{1}{\sqrt{\#I}}\sum_{n_i\in I} Z_{n_i} \xrightarrow{d} N(0,\bar{\sigma}_N^2),$$

which implies

$$\frac{1}{\sqrt{n_N h}}\sum_{i=1}^{N} k\left(\frac{n_i - n_N\tau}{n_N h}\right)\sigma\left(\frac{n_i}{n_N}\right)\xi_{n_i} \xrightarrow{d} \sqrt{C}N(0,\bar{\sigma}_N^2) = N(0,V(\tau)),$$

where

$$V(\tau) = \frac{\sigma^2(\tau)\int_{-1}^{1}k^2(u)du}{D}\left[1+2\sum_{m=1}^{\infty}\rho_{gm}\right].$$

Q.E.D.

.2.3 Proof of theorem 4

From the proof of Theorem 1, we know that

$$\hat{\sigma}_\xi^2(\tau) = \frac{\sum_{i=1}^{N} w_i \hat{R}_{n_i}}{\sum_{i=1}^{N} w_i}, \tag{.2.18}$$

where

$$w_i = \tilde{k}(\frac{n_i - n_N\tau}{n_N\tilde{h}})\left(\tilde{S}_{\tau,2} - \tilde{S}_{\tau,1}(\frac{n_i}{n_N}-\tau)\right),$$

and

$$\tilde{S}_{\tau,j} = \sum_{i=1}^{N}\tilde{k}\left(\frac{n_i - \tau n_N}{n_N\tilde{h}}\right)(\frac{n_i}{n_N}-\tau)^j.$$

From (.2.18),

$$\begin{aligned}
\hat{\sigma}_{\xi}^{2}(\tau)-\sigma_{\xi}^{2}(\tau) &= \frac{\sum_{i=1}^{N} w_i\left(\hat{R}_{n_i}-\sigma_{\xi}^{2}(\tau)\right)}{\sum_{i=1}^{N} w_i} \\
&= \frac{\sum_{i=1}^{N} w_i\left((R_{n_i}-\sigma_{\xi}^{2}(\tau))+(\hat{R}_{n_i}-R_n)\right)}{\sum_{i=1}^{N} w_i} \\
&= \frac{\sum_{i=1}^{N} w_i\left[\left(\sigma_{\xi}^{2}(\frac{n_i}{n_N})\xi_{n_i}^{2}-\sigma_{\xi}^{2}(\frac{n_i}{n_N})\right)+\left(\sigma_{\xi}^{2}(\frac{n_i}{n_N})-\sigma_{\xi}^{2}(\tau)\right)+(\hat{R}_{n_i}-R_n)\right]}{\sum_{i=1}^{N} w_i} \\
&= I_1+I_2+I_3,
\end{aligned}$$

where

$$I_1=\frac{\sum_{i=1}^{N} w_i\left(\sigma_{\xi}^{2}(\frac{n_i}{n_N})\xi_{n_i}^{2}-\sigma_{\xi}^{2}(\frac{n_i}{n_N})\right)}{\sum_{i=1}^{N} w_i},$$

$$I_2=\frac{\sum_{i=1}^{N} w_i\left(\sigma_{\xi}^{2}(\frac{n_i}{n_N})-\sigma_{\xi}^{2}(\tau)\right)}{\sum_{i=1}^{N} w_i},$$

$$I_3=\frac{\sum_{i=1}^{N} w_i\left(\hat{R}_{n_i}-R_{n_i}\right)}{\sum_{i=1}^{N} w_i}.$$

Note that, from the proof of Theorem 1 and by Assumption 4

$$\begin{aligned}
I_1 &= \frac{\sum_{i=1}^{N} w_i\left(\sigma_{\xi}^{2}(\frac{n_i}{n_N})\left(\xi_{n_i}^{2}-1\right)\right)}{\sum_{i=1}^{N} w_i} \\
&= \frac{\sum_{i=1}^{N} \tilde{k}(\frac{n_i-n_N\tau}{n_N\tilde{h}})\left(\tilde{S}_{\tau,2}-\tilde{S}_{\tau,1}(\frac{n_i}{n_N}-\tau)\right)\left(\sigma_{\xi}^{2}(\frac{n_i}{n_N})\left(\xi_{n_i}^{2}-1\right)\right)}{\left(\tilde{S}_{\tau,0}\tilde{S}_{\tau,2}-\tilde{S}_{\tau,1}^{2}\right)} \\
&\sim \frac{1}{n_N\tilde{h}}\sum_{i=1}^{N}\tilde{k}(\frac{n_i-n_N\tau}{n_N\tilde{h}})\left(\sigma_{\xi}^{2}(\frac{n_i}{n_N})\left(\xi_{n_i}^{2}-1\right)\right) \\
&= \frac{\frac{1}{\sqrt{n_N\tilde{h}}}\frac{1}{\sqrt{n_N\tilde{h}}}\sum_{i=1}^{N}\tilde{k}(\frac{n_i-n_N\tau}{n_N\tilde{h}})\left(\sigma^{2}(\frac{n_i}{n_N})\left(\xi_{n_i}^{2}-1\right)\right)}{n_N^{2\delta}}.
\end{aligned}$$

As in the proof of Theorem 2, denote $Z_{n_i}=\tilde{k}(\frac{n_i-n_N\tau}{n_N\tilde{h}})\left(\sigma^{2}(\frac{n_i}{n_N})\left(\xi_{n_i}^{2}-1\right)\right)$, $E(Z_{n_i})=0$. If $Var\left(\frac{1}{\sqrt{n_N\tilde{h}}}\sum_{i=1}^{N}Z_{n_i}\right)=0$, $I_1=o_p(\frac{1}{n_N^{1/2+2\delta}\tilde{h}^{1/2}})$. If $Var\left(\frac{1}{\sqrt{n_N\tilde{h}}}\sum_{i=1}^{N}Z_{n_i}\right)>0$, for some $\beta>2$,

$$\begin{aligned}
\mathbb{E}|Z_{n_i}|^{\beta} &= \mathbb{E}\left|\tilde{k}(\frac{n_i - n_N\tau}{n_N\tilde{h}})\left(\sigma^2(\frac{n_i}{n_N})\left(\xi_{n_i}^2 - 1\right)\right)\right|^{\beta} \\
&\leq C\mathbb{E}\left|\xi_{n_i}^2 - 1\right|^{\beta} \\
&\leq C2^{\beta-1}(\mathbb{E}\left|\xi_{n_i}^2\right|^{\beta} + 1) \\
&\leq C\sup_i \mathbb{E}\left|\xi_{n_i}^2\right|^{\beta} \\
&< \infty.
\end{aligned}$$

Thus, as in the proof of Theorem 2, by theorem 5.20 in White (2001), we have

$$I_1 = O_p(\frac{1}{n_N^{1/2+2\delta}\tilde{h}^{1/2}}).$$

In sum, we have

$$I_1 = O_p(\frac{1}{n_N^{1/2+2\delta}\tilde{h}^{1/2}}), \tag{.2.19}$$

and

$$
\begin{aligned}
I_2 &= \frac{\sum_{i=1}^{N} w_i \left(\sigma_\xi^2(\frac{n_i}{n_N}) - \sigma_\xi^2(\tau)\right)}{\sum_{i=1}^{N} w_i}, \\
&= \frac{\sum_{i=1}^{N} w_i \left(\sigma^2(\frac{n_i}{n_N}) - \sigma^2(\tau)\right)}{n_N^{2\delta}\sum_{i=1}^{N} w_i} \\
&= \frac{\sum_{i=1}^{N} w_i \left(\sigma^2(\tau) + 2\sigma'(\tau)(\frac{n_i}{n_N} - \tau) + \sigma''(\eta_i)(\frac{n_i}{n_N} - \tau)^2 - \sigma^2(\tau)\right)}{n_N^{2\delta}\sum_{i=1}^{N} w_i} \qquad (.2.20) \\
&= \frac{\sum_{i=1}^{N} w_i \left(\sigma''(\eta_i)(\frac{n_i}{n_N} - \tau)^2\right)}{n_N^{2\delta}\sum_{i=1}^{N} w_i} \qquad (.2.21) \\
&\leq \frac{C\sum_{i=1}^{N} w_i(\frac{n_i}{n_N} - \tau)^2}{n_N^{2\delta}\sum_{i=1}^{N} w_i} \qquad (.2.22) \\
&= \frac{C\sum_{i=1}^{N} \tilde{k}(\frac{n_i - n_N\tau}{n_N\tilde{h}}) \left(\tilde{S}_{\tau,2} - \tilde{S}_{\tau,1}(\frac{n_i}{n_N} - \tau)\right)(\frac{n_i}{n_N} - \tau)^2}{n_N^{2\delta}\left(\tilde{S}_{\tau,0}\tilde{S}_{\tau,2} - \tilde{S}_{\tau,1}^2\right)} \\
&\sim \frac{C\sum_{i=1}^{N} \tilde{k}(\frac{n_i - n_N\tau}{n_N\tilde{h}})(\frac{n_i}{n_N} - \tau)^2\tilde{S}_{\tau,2}}{n_N^{2\delta}\left(\tilde{S}_{\tau,0}\tilde{S}_{\tau,2} - \tilde{S}_{\tau,1}^2\right)} \\
&= O(\frac{\tilde{h}^2}{n_N^{2\delta}}), \qquad (.2.23)
\end{aligned}
$$

where (.2.20) is obtained by the Taylor expansion, (.2.21) is from $\sum_{i=1}^{N} w_i(\frac{n_i}{n_N} - \tau) = 0$ and (.2.22) is from the boundness of $\sigma''(\cdot)$.

Further, by the definition of $\hat{R}_{n_i}$ and R_{n_i},

$$
\begin{aligned}
I_3 &= \frac{\sum_{i=1}^{N} w_i \left[2\left(f(\frac{n_i}{n_N}) - \hat{f}(\frac{n_i}{n_N})\right)\sigma_\xi(\frac{n_i}{n_N})\xi_{n_i} + \left(f(\frac{n_i}{n_N}) - \hat{f}(\frac{n_i}{n_N})\right)^2\right]}{\sum_{i=1}^{N} w_i} \\
&= I_{31} + I_{32},
\end{aligned}
$$

where,

$$\begin{aligned} I_{31} &= \frac{2\sum_{i=1}^{N} w_i \left(f(\frac{n_i}{n_N}) - \hat{f}(\frac{n_i}{n_N})\right) \sigma_{\xi}(\frac{n_i}{n_N})\xi_{n_i}}{\sum_{i=1}^{N} w_i} \\ &= \frac{2\sum_{i=1}^{N} w_i \left(f(\frac{n_i}{n_N}) - \hat{f}(\frac{n_i}{n_N})\right) \sigma(\frac{n_i}{n_N})\xi_{n_i}}{n_N^{\delta}\sum_{i=1}^{N} w_i} \\ &= o_p(\frac{h^2+\tilde{h}^2}{n_N^{\delta}}), \end{aligned} \tag{.2.24}$$

where the last equality is from the proof of Theorem 1 in Fan and Yao (1998).

Similarly, from the proof of Theorem 1 in Fan and Yao (1998), we have

$$I_{32} = o_p\left(\frac{h^2+\tilde{h}^2}{n_N^{\delta}}\right). \tag{.2.25}$$

Hence, from (.2.19), (.2.23), (.2.24) and (.2.25),

$$\hat{\sigma}_{\xi}^2(\tau) - \sigma_{\xi}^2(\tau) = O_p\left(\frac{1}{n_N^{1/2+2\delta}\tilde{h}^{1/2}} + \frac{\tilde{h}^2}{n_N^{2\delta}}\right) + o_p\left(\frac{h^2+\tilde{h}^2}{n_N^{\delta}}\right).$$

Q.E.D