The Pennsylvania State University
The Graduate School

THREE ESSAYS ON NONPARAMETRIC INFERENCE FOR LONGITUDINAL DATA AND TIME SERIES DATA

A Dissertation in
Statistics
by
Seonjin Kim

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2013

The dissertation of Seonjin Kim was reviewed and approved* by the following:

Zhibiao Zhao
Assistant Professor of Statistics
Dissertation Advisor, Chair of Committee

David Hunter
Professor of Statistics
Department Head

Runze Li
Distinguished Professor of Statistics

Donald St. P. Richards
Professor of Statistics

Fuqing Zhang
Professor of Meteorology and Statistics

Jingzhi Haung
Associate Professor of Finance

*Signatures are on file in the Graduate School.

Abstract

This dissertation consists of three essays on nonparametric inference problems for dependent data, such as longitudinal data and time series data. In the literature on statistical estimation and inference, the frequently imposed independence assumption allows for technical simplicity, but it leads to serious restrictions in many applications. For example, in longitudinal data analysis and time series analysis, dependence is actually one of the main objectives of interest. The techniques developed for independent data may fail or not be efficient where dependence presents. Moreover, the traditional approach to nonparametric inference problems has some disadvantages due to the structure dependency and estimation difficulty of the limiting variance function, and there exist mathematical obstacles caused by dependence inherited in data. In order to alleviate the aforementioned problems, I have developed novel theories and methodologies for dependent data through self-normalization techniques and quantile regression. The developments involve theoretical notability as well as practical applicability for financial, medical, environmental, and social science.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I would like to thank very much my advisor Zhibiao Zhao. I cannot imagine that I have finished my work without his support and instruction. Also I would like to thank all the faculty at the Department of Statistics at Pennsylvania State University for great lectures contributing a lot to my research and statistical insight. I would like to express my thank to all my committee members who contributed with advice and encouragement. Finally, I want to thank my wife, Youngsun Yu, for her endless supply of support, for which my mere expression of thanks does not suffice.

Seonjin Kim

Chapter 1

Introduction

In the last two decades, one of the most active research areas in statistics is nonparametric modeling. Nonparametric models and methods are useful complements to the traditional well developed parametric counterparts. They allow the users to entertain model flexibility while reducing modeling bias, and partly due to this reason, nonparametric inference has been extensively studied under various settings (Fan and Gijbels, 1996; Fan and Yao, 2003; Li and Racine, 2007). Denote by $\mu(x)$ a nonparametric function of interest from an underlying population (X, Y). For example, in nonparametric mean regression problems, X and Y represent covariates and response, respectively, and $\mu(x) = \mathbb{E}(Y|X = x)$ is the unknown mean regression function. In quantile regression, $\mu(x)$ is the conditional quantile of Y given $X = x$. Let $\hat{\mu}_n(x)$ be a nonparametric estimate of $\mu(x)$ based on observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. Under suitable conditions, we have

$$\sqrt{nb_n}\frac{\hat{\mu}_n(x) - \mu(x) - b_n^2\rho(x)}{s(x)} \xrightarrow{\mathcal{D}} N(0, 1), \tag{1.1}$$

where b_n is an appropriate bandwidth, $b_n^2\rho(x)$ is the bias term, and $s^2(x)$ is the limiting variance function. Throughout we use $\xrightarrow{\mathcal{D}}$ to denote convergence in distribution and $N(0, 1)$ to denote the standard normal distribution.

This dissertation has three contributions. First, for statistical inference of $\mu(x)$, the traditional approach involves consistent estimation of $s^2(x)$ using a bandwidth-dependent nonparametric procedure. However, estimating $s(x)$ is often problematic because of its structure dependency and complicated structure. The first

contribution is to develop some unified self-normalization technique for both the dense and sparse longitudinal models that does not need an extra estimation procedure for $s(x)$. Second, because of dependence inherited in longitudinal data, it is not trivial to develop notable methodology for efficient estimation of $\mu(x)$ or to extend approaches developed for independent data. The second contribution is to develop efficient estimation for the nonparametric mean regression function of longitudinal models. The third contribution is to develop specification testing for that the model dynamics of the original but unobservable data follows a parametric model, when the time series observations are contaminated. These three topics are investigated in Chapter 2–4.

In Chapter 2, unified inference for sparse and dense longitudinal models is investigated. In longitudinal data analysis, statistical inference for sparse data and dense data could be substantially different. We consider the mixed-effects model for longitudinal data:

$$Y_{ij} = \mu(X_{ij}) + v_i(X_{ij}) + \sigma(X_{ij})\epsilon_{ij}, \quad i = 1, \ldots, n;\ j = 1, \ldots, n_i,$$

where $\mu(\cdot)$ is a fixed population mean, $v_i(\cdot)$ is a subject-specific random trajectory with $E\{v_i(x)\} = 0$ and covariance function $\gamma(x, x') = \mathrm{cov}\{v_i(x), v_i(x')\}$, and $\{\epsilon_{ij}\}$ are errors with $E(\epsilon_{ij}) = 0$ and $E(\epsilon_{ij}^2) = 1$. Nonparametric estimate of $\mu(\cdot)$ has different convergence rates and limiting variances under the sparse ($\{n_i\}$ are bounded) and dense scenarios ($n_i \to \infty$). The latter phenomenon poses challenges for statistical inference as a subjective choice between the sparse and dense cases may lead to wrong conclusions.

Existing works treat the two cases separately. Because each subject possesses enough numbers of observations in dense longitudinal data, a conventional estimation approach is to smooth each individual curve first and then construct an estimator based on the smoothed curves (Ramsay and Silverman, 2005; Hall et al., 2006; Zhang and Chen, 2007). By contrast, due to the sparse observations from individual subjects, it is essential for sparse longitudinal data to pool data together (Yao et al., 2005a; Hall et al., 2006; Yao, 2007). Besides the problem caused by the above two scenarios, another challenging issue is that the limiting variance function contains the unknown functions $\gamma(x, x)$ and $\sigma^2(x)$. As shown

by Wu and Zhang (2002), Yao et al. (2005a,b), Müller (2005) and Li and Hsing (2010), covariance estimation requires extra smoothing procedures. Moreover, the procedures are not trivial to implement.

To alleviate aforementioned issues, I develop two self-normalization based methods that can adapt to the sparse and dense cases in a unified framework. The key idea of the first approach is to build a self-normalizer whose limit and convergence rate are automatically adjusted depending on whether the condition follows the sparse or dense setting. In other words, the self-normalizer multiplied by the squared convergence rate of the sparse (dense) case converges the limiting variance of sparse (dense) longitudinal data. The second approach constructs a self-normalizer, which can cancel out the convergence rate and limting variance, based on recursive estimates of the mean function. The related methods have been explored mainly under parametric settings for time series data (Lobato, 2001; Kiefer and Vogelsang, 2005; Shao, 2010). A unified inference can be conducted by both the approaches without deciding whether the data are dense or sparse. Furthermore, an extra smoothing procedures for estimating the unknown functions $\gamma(x, x)$ and $\sigma^2(x)$ is not required in these approaches. Simulations show that the proposed methods outperform some existing methods.

In Chapter 3, I proposed an efficient estimation for time-varying coefficient longitudinal models by optimally combining quantile regressions coupled with a prewhitening transformation. The proposed estimator asymptotically achieves the Cramér-Rao bound, which is the first such result in the literature on nonparametric estimation of longitudinal models. Consider the time-varying coefficient longitudinal model:

$$Y_{ij} = \alpha(t_{ij}) + X_i(t_{ij})^T \beta(t_{ij}) + \varepsilon_{ij}, \quad i = 1, \ldots, n; \ j = 1, \ldots, n_i,$$

where $\alpha(\cdot)$ is a time-varying trend intercept, $\beta(\cdot)$ is a d-dimensional vector of coefficient functions of interest, and ε_{ij} is a noise. The widely used local least-squares (LS) regression (Fan and Gijbels, 1996) and the traditional quantile regression (QR) (Koenker, 2005) method use only partial information: sample average for the LS method and single sample quantile for the QR method. Intuitively, more efficient estimators can be potentially constructed by exploiting more information

from the data. Unlike the LS method, QR provides a method for estimating the whole conditional distribution and thus offers a natural framework for combining information across quantiles. While the literature on estimation via combining quantile information has focused on simple linear models (Koenker, 1984; Portnoy and Koenker, 1989; Zou and Yuan, 2008; Bradic, Fan and Wang, 2011) and nonparametric regression models with i.i.d. symmetric error (Kai, Li and Zou, 2010), our development for the time-varying coefficient longitudinal model involves novel theory and methodology.

The proposed methodology consists of two components. First, we combine information across multiple quantiles under the independent assumption (working independence). Let $\tau_r = r/(k+1)$, $r = 1, \ldots, k$, be k uniformly spaced quantiles. We propose the *weighted local quantile average estimator* (WLQAE) with weights $[\omega_1, \ldots, \omega_k]^T$: $\hat{\beta}_{\mathrm{WLQAE}}(t|\omega) = \sum_{r=1}^{k} \omega_r \hat{\beta}(t|\tau_r)$, where $\hat{\beta}(t|\tau)$ is a τ-th quantile regression estimator. When the optimal weight, which minimizes the asymptotic variance, is used, the limiting variance of the resultant estimator converges to the inverse of the Fisher information of the density of the independent innovations as the number of quantiles $k \to \infty$. Second, in the presence of dependence among within-subject measurements, we adopt a prewhitening technique (Xiao et al., 2003) to transform regression errors into independent innovations. To illustrate the idea, suppose that $\{\varepsilon_{ij}\}_{j=1,\ldots,m_i}$ for subject i follows an AR(p) process: $\varepsilon_{ij} = \sum_{k=1}^{p} \phi_k \epsilon_{ij-k} + e_{ij}$ $j \in \mathbb{N}$, where $\{e_{ij}\}$ are i.i.d. innovations and are also independent of the past $\{\varepsilon_{ij-1}, \varepsilon_{ij-2}, \ldots\}$. We obtain the transformed new response $Y_{ij}^* := Y_{ij} - \sum_{k=1}^{p} \phi_k \epsilon_{ij-k}$ and then apply the optimally WLQAE (OWLQAE) methodology to the prewhitened data $(t_{ij}, X_i(t_{ij}), Y_{ij}^*)$. It is theoretically shown that the prewhitened OWLQAE is more efficient than the working independence OWLQAE. Fully data-driven bandwidth selection and optimal weights estimation are implemented through a two-step procedure. Monte Carlo studies show that the proposed method delivers more robust and superior overall performance than the local least-squares, the weighted local least-squares, and the median quantile regression methods. The Six Cities Study of Air Pollution and Health is used to illustrate the new methodology.

Most existing works on specification testing assume that we have direct observations from the model of interest (Azzalini and Bowman, 1993; Hong and White,

1995; Fan and Li, 1996; Aït-Sahalia, 1996; Fan et al., 2001; Gao and King, 2004; Hong and Li, 2005; Aït-Sahalia et al., 2009). In Chapter 4, we study specification testing for contaminated Markov models. Let $\{X_i\}_{i\in\mathbb{N}}$ be a Markov chain of interest. In some applications, $\{X_i\}$ may not be directly observable and instead we observe a contaminated version $\{Y_i\}$:

$$Y_i = X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where $\{\varepsilon_i\}$ are i.i.d. measurement errors. For example, the above model has been proposed to explain the microstructure noise phenomenon observed in high-frequency financial data. In order to test whether the underlying Markov chain $\{X_i\}$ follows a parametric model, I have developed a conditional expectation generator based specification testing method by combining simultaneous confidence bands (SCB) across different transformations of the data. The evolving model dynamics of the unobservable Markov chain is implicitly coded into the conditional distribution of the observed process. Therefore, we propose measuring the deviation between nonparametric and parametric estimates of conditional regression functions of the observed process, $\mathcal{G}_g(y) = \mathbb{E}[g(Y_i)|Y_{i-1} = y]$, for proper transformations $g(\cdot)$. To determine the critical value, we use the idea of SCB. For a significance level $\alpha \in (0, 1)$, we say that $[l_n(\cdot), u_n(\cdot)]$ is an asymptotic $(1-\alpha)$ nonparametric SCB for $\mathcal{G}_g(y)$ on a given compact set $\mathcal{Y} \subset \mathbb{R}$ if

$$\lim_{n\to\infty} \mathbb{P}\{l_n(y) \le \mathcal{G}_g(y) \le u_n(y), \text{ for all } y \in \mathcal{Y}\} = 1 - \alpha.$$

The nonparametric band $[l_n(\cdot), u_n(\cdot)]$ provides an acceptance region for H_0. If the parametric estimate under H_0 falls outside the band, then reject H_0 at level α. The empirical performance of the proposed method is illustrated through Monte Carlo simulation studies.

In Chapter 5, my future work, quantile regression for locally stationary processes, is briefly introduced. The stationarity plays a pivotal role in the classical time series analysis, but the stationary assumption is often too strong to apply in many applications. In the sea level studies, it is easy to see that the underlying model structure of the see level series has changed over time. Especially, in this case, the non-stationary pattern is the main objective. However, the probabilistic

structure of a non-stationary process at present may not provide any information for future observation of the process so that it is often not clear how to derive meaningful asymptotic properties for non-stationary processes. To alleviate the difficulty, we adopt an idea of locally stationary process. For the neighborhood of a fixed point, a locally stationary process of interest can be approximated by a stationary process under some smoothness conditions so that we can investigate asymptotic properties of the locally stationary process along with the stationary process. Since Dahlhaus (1997), locally stationary models has been actively studied (Davis et al., 2006; Dahlhaus and Subba Rao, 2006; Van Bellegem and Von Sachs, 2008). I also refer to Dahlhaus (2011) for a survey. In addition, estimation of quantiles has recently become an important problem in time series analysis (Draghicescu et.al., 2009; Zhou and Wu, 2009; Zhou, 2010). For example, notable changes in climate variability can make a devastating impact on the environment, and hence the climate extremes are more informative than the mean function. The main goal of this work is to study quantile regression for time series data that contain non-stationary time-varying dependence.

Chapter 2

Unified inference for sparse and dense longitudinal models

2.1 Introduction

Longitudinal models have extensive applications in biomedical, psychometric and environmental sciences (Fitzmaurice et al., 2004; Wu and Zhang, 2006). In longitudinal studies, repeated measurements are recorded over time from subjects, and therefore measurements from the same subject are correlated. One popular framework is to assume that the observations from each subject are noisy discrete realizations of an underlying process $\{\xi(\cdot)\}$:

$$Y_{ij} = \xi_i(X_{ij}) + \sigma(X_{ij})\epsilon_{ij} \quad (i = 1, \ldots, n;\ j = 1, \ldots, n_i). \tag{2.1}$$

Here Y_{ij} is the measurement at time X_{ij} from subject i, $\{\xi_i(\cdot)\}$ are independent realizations of an underlying process $\{\xi(\cdot)\}$, ϵ_{ij} are errors with $E(\epsilon_{ij}) = 0$ and $E(\epsilon_{ij}^2) = 1$, n_i is the number of measurements collected on subject i, and n is the total number of subjects.

There are two typical approaches to taking between-subject variation into account: functional principal component analysis (Yao et al., 2005a,b; Yao, 2007; Ma et al., 2012) and the mixed-effects approach (Wu and Zhang, 2002; Zhang and Chen, 2007). The basic idea of the latter is to decompose $\{\xi_i(\cdot)\}$ into a fixed population mean $\mu(\cdot) = E\{\xi_i(\cdot)\}$ and a subject-specific random trajectory $v_i(\cdot)$ with

$E\{v_i(x)\} = 0$ and covariance function $\gamma(x, x') = \text{cov}\{v_i(x), v_i(x')\}$. Then (2.1) becomes

$$Y_{ij} = \mu(X_{ij}) + v_i(X_{ij}) + \sigma(X_{ij})\epsilon_{ij} \quad (i = 1, \ldots, n;\ j = 1, \ldots, n_i). \tag{2.2}$$

The goal is to estimate the population mean $\mu(\cdot)$ and construct a confidence interval for it.

Depending on the number of measurements within subjects, model (2.2) has two scenarios: dense and sparse longitudinal data. Dense longitudinal data allow $n_i \to \infty$ and a conventional estimation approach is to smooth each individual curve and then construct an estimator based on the smoothed curves (Ramsay and Silverman, 2005; Hall et al., 2006; Zhang and Chen, 2007). In sparse longitudinal data, the n_i are either bounded or independent and identically distributed with $E(n_i) < \infty$, and due to the sparse observations from individual subjects, it is essential to pool data together (Yao et al., 2005a; Hall et al., 2006; Yao, 2007).

In practice, the boundary between dense and sparse cases may not always be clear, and such ambiguity could pose challenges for statistical inference, since different researchers may likely classify the same data set differently. To address this issue, Li and Hsing (2010) proposed a unified weighted local linear estimator of $\mu(x)$. However, as shown in Section 2.2, the latter estimator has different convergence rates and limiting variances under the two scenarios. Therefore, to construct a confidence interval for $\mu(x)$, one should make a subjective decision whether to treat the data as sparse or dense. In Section 2.2, we show that the constructed confidence intervals based on a sparse or dense assumption could differ substantially, depending on many unknown factors. Another challenging issue is that the limiting variance function contains the unknown functions $\gamma(x, x)$ and $\sigma^2(x)$. As shown by Wu and Zhang (2002), Yao et al. (2005a,b), Müller (2005) and Li and Hsing (2010), covariance estimation requires extra smoothing procedures.

We develop two unified nonparametric approaches that can successfully solve the aforementioned issues. First, we establish a unified convergence theory so that inference can be conducted without deciding whether the data are dense or sparse. Second, the unknown limiting variance is canceled out through a self-normalization technique, and thus the proposed methods do not require estimation of the func-

tions $\gamma(x,x)$ and $\sigma^2(x)$. The first approach introduces a unified self-normalized central limit theorem that can adapt to both cases. The second approach constructs a self-normalizer based on recursive estimates of the mean function. The related methods have been explored mainly under parametric settings for time series data (Lobato, 2001; Kiefer and Vogelsang, 2005; Shao, 2010). In the longitudinal setting, our development of the self-normalization method is more attractive due to the sparse and dense scenario and the more complicated structure such as the within-subject covariance and overall noise variance function. Simulations show that the proposed methods outperform some existing methods.

2.2 Motivation

For model (2.2), we consider two scenarios: (i) sparse longitudinal data: $n_1, \ldots, n_n$ are independent and identically distributed positive-integer-valued random variables with $E(n_i) < \infty$; and (ii) dense longitudinal data: $n_i \geq M_n$ for some $M_n \to \infty$ as $n \to \infty$.

Throughout we let $f(\cdot)$ denote the density function of X_{ij} and let x be an interior point of the support of $f(\cdot)$. Li and Hsing (2010) proposed a sample-size weighted local linear estimator of $\mu(x)$. For technical convenience, we consider the weighted local constant estimator

$$\hat{\mu}_n(x) = \operatorname*{argmin}_{\theta} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} (Y_{ij} - \theta)^2 K\Big(\frac{X_{ij} - x}{b}\Big) = \frac{G_n}{H_n}, \tag{2.3}$$

where K is a kernel function satisfying $\int_{\mathbb{R}} K(u)du = 1$ and $b > 0$ is a bandwidth, with

$$H_n = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} K\Big(\frac{X_{ij} - x}{b}\Big), \quad G_n = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} K\Big(\frac{X_{ij} - x}{b}\Big). \tag{2.4}$$

The convergence rates and limiting variances are different for sparse and dense longitudinal data. To gain intuition about this, write

$$\hat{\mu}_n(x) - \mu(x) - \frac{1}{H_n} \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \{\mu(X_{ij}) - \mu(x)\} K\Big(\frac{X_{ij} - x}{b}\Big) = \frac{1}{H_n} \sum_{i=1}^{n} \xi_i, \tag{2.5}$$

where the right hand side determines the asymptotic distribution of $\hat{\mu}_n(x)$, with

$$\xi_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\xi_{ij}, \quad \xi_{ij} = \{v_i(X_{ij}) + \sigma(X_{ij})\epsilon_{ij}\}K\Big(\frac{X_{ij}-x}{b}\Big). \tag{2.6}$$

Recall $\gamma(x,x') = \mathrm{cov}\{v_i(x), v_i(x')\}$. For $j \neq j'$, by $E(\xi_{ij}\xi_{ij'}) = E\{E(\xi_{ij}\xi_{ij'} \mid X_{ij}, X_{ij'})\}$,

$$E(\xi_{ij}\xi_{ij'}) = E\Big\{\gamma(X_{ij}, X_{ij'})K\Big(\frac{X_{ij}-x}{b}\Big)K\Big(\frac{X_{ij'}-x}{b}\Big)\Big\} \approx b^2 f^2(x)\gamma(x,x). \tag{2.7}$$

Throughout, $c_n \approx d_n$ means that $c_n/d_n \to 1$. Similarly,

$$E(\xi_{ij}^2) = E\{E(\xi_{ij}^2 \mid X_{ij})\} \approx bf(x)\psi_K\{\gamma(x,x)+\sigma^2(x)\}, \quad \psi_K = \int_{\mathbb{R}} K^2(u)du. \tag{2.8}$$

Applying (2.7)–(2.8) to $\mathrm{var}(\xi_i \mid n_i) = n_i^{-2}\{\sum_{1\le j\neq j'\le n_i} E(\xi_{ij}\xi_{ij'}) + \sum_{j=1}^{n_i} E(\xi_{ij}^2)\}$, we obtain

$$\mathrm{var}(\xi_i \mid n_i) \approx (1-1/n_i)b^2 f^2(x)\gamma(x,x) + f(x)\psi_K\{\gamma(x,x)+\sigma^2(x)\}b/n_i. \tag{2.9}$$

For sparse case with $b \to 0$, $\mathrm{var}(\xi_i \mid n_i) \approx bf(x)\psi_K\{\gamma(x,x)+\sigma^2(x)\}/n_i$; for dense case with $n_i \geq M_n$ and $M_n b \to \infty$, $\mathrm{var}(\xi_i \mid n_i) \approx b^2 f^2(x)\gamma(x,x)$.

Theorem 1. *Assume Assumption in Section 2.5 on page 16. Let $f(x)$ be the density of X_{ij}. Write*

$$\psi_K = \int_{\mathbb{R}} K^2(u)du, \quad \rho(x) = \Big\{\frac{\mu''(x)}{2} + \frac{\mu'(x)f'(x)}{f(x)}\Big\}\int_{\mathbb{R}} u^2 K(u)du.$$

(i) Sparse data: Assume $nb \to \infty$ and $\sup_n nb^5 < \infty$. Then

$$\sqrt{(nb)}\{\hat{\mu}_n(x) - \mu(x) - b^2\rho(x)\} \to N\{0, s^2_{\mathrm{sparse}}(x)\}, \tag{2.10}$$

where $s^2_{\mathrm{sparse}}(x) = \tau\psi_K\{\gamma(x,x)+\sigma^2(x)\}/f(x)$ and $\tau = E(1/n_1)$.

(ii) Dense data: Assume $n_i \geq M_n, M_n b \to \infty, nb \to \infty$ and $\sup_n nb^4 < \infty$. Then

$$\sqrt{n}\{\hat{\mu}_n(x) - \mu(x) - b^2\rho(x)\} \to N\{0, s^2_{\mathrm{dense}}(x)\}, \quad s^2_{\mathrm{dense}}(x) = \gamma(x,x). \tag{2.11}$$

It is worth mentioning some related results. Li and Hsing (2010) established the uniform consistency of $\hat{\mu}_n(x)$ with different rates under the sparse and dense cases, but they did not obtain the asymptotic distribution. Wu and Zhang (2002) also showed that the local polynomial mixed-effects estimator has different convergence rates and limiting variances under the two scenarios. Under a Karhunen–Loève representation of longitudinal models, Yao (2007) studied the sparse case by allowing n_i to be dependent on n; see also Ma et al. (2012).

By Theorem 1, the confidence interval for $\mu(x)$ is different under the two cases. Let $z_{1-\alpha/2}$ be the $1-\alpha/2$ standard normal quantile. Then an asymptotic $1-\alpha$ confidence interval for $\mu(x)$ is

$$\hat{\mu}_n(x) - b^2\hat{\rho}(x) \pm z_{1-\alpha/2}(nb)^{-1/2}[\hat{\tau}\psi_K\{\hat{\gamma}(x,x) + \hat{\sigma}^2(x)\}/\hat{f}(x)]^{1/2} \tag{2.12}$$

for sparse data, or

$$\hat{\mu}_n(x) - b^2\hat{\rho}(x) \pm z_{1-\alpha/2}n^{-1/2}\{\hat{\gamma}(x,x)\}^{1/2} \tag{2.13}$$

for dense data. Here, $\hat{\tau} = n^{-1}\sum_{i=1}^{n} n_i^{-1}$, $\hat{\gamma}(x,x)$, $\hat{\sigma}^2(x)$, $\hat{f}(x)$ and $\hat{\rho}(x)$ are consistent estimates of τ, $\gamma(x,x)$, $\sigma^2(x)$, $f(x)$ and $\rho(x)$. The ratio of the lengths of the two confidence intervals is $R = [\psi_K\hat{\tau}\{1 + \hat{\sigma}^2(x)/\hat{\gamma}(x,x)\}/\{b\hat{f}(x)\}]^{1/2}$, which depends on the denseness parameter τ, the signal-to-noise ratio $\gamma(x,x)/\sigma^2(x)$, the bandwidth b and the design density $f(\cdot)$. The further away R is from one, the larger the discrepancy between the two constructed confidence intervals.

Remark 1. In the dense case, suppose n_i is proportional to $M_n \to \infty$. Theorem 1 (ii) studies the case $M_n b \to \infty$. If $M_n b \to 0$, then the leading term in (2.9) is $f(x)\psi_K\{\gamma(x,x) + \sigma^2(x)\}b/n_i$. If $M_n b$ is bounded away from 0 and ∞, then both terms in (2.9) are of the same order. If b is proportional to $(nM_n)^{-1/5}$, then a sufficient condition for $M_n b \to \infty$ is $M_n^4/n \to \infty$. In many practical problems, n is about 30–200, M_n is about 10–30, and M_n^4/n is sufficiently large.

2.3 Unified approaches for sparse and dense data

2.3.1 A unified self-normalized central limit theorem

The discussion in Section 2.2 suggests a need for a unified approach. For independent and identically distributed random variables $Z_1, \ldots, Z_n$, de la Peña et al. (2009) gave an extensive account of the asymptotic properties of the self-normalized statistic $\sum_{i=1}^n Z_i/\sqrt{(\sum_{i=1}^n Z_i^2)}$. In this section, we present a unified self-normalized central limit theorem for $\hat{\mu}_n(x)$. For H_n in (2.4), define

$$U_n^2(x) = \frac{1}{H_n^2} \sum_{i=1}^{n} \Big[\frac{1}{n_i} \sum_{j=1}^{n_i} \{Y_{ij} - \hat{\mu}(X_{ij})\} K\Big(\frac{x - X_{ij}}{b}\Big) \Big]^2.$$

Theorem 2. *Assume Assumption in Section 2.5 on page 16. Suppose* $nb/\log n \to \infty$, $\sup_n nb^5 < \infty$ *for sparse data or* $n_i \geq M_n$, $M_n b \to \infty$, $nb^2/\log n \to \infty$, $\sup_n nb^4 < \infty$ *for dense data. Then* $\{\hat{\mu}_n(x) - \mu(x) - b^2\rho(x)\}/U_n(x) \to N(0,1)$ *in both the sparse and dense settings.*

Many papers treat sparse and dense data separately. For example, Yao et al. (2005 a,b), Yao (2007) and Ma et al. (2012) studied sparse longitudinal data. For the local polynomial mixed-effects estimator, Wu and Zhang (2002) obtained different central limit theorems under the two scenarios. By contrast, Theorem 2 establishes a unified central limit theorem, which can be used to construct a unified asymptotic pointwise $1 - \alpha$ confidence interval for $\mu(x)$:

$$\hat{\mu}_n(x) - b^2\hat{\rho}(x) \pm z_{1-\alpha/2} U_n(x). \tag{2.14}$$

While the confidence intervals (2.12) and (2.13) require estimation of the within-subject covariance function $\gamma(x, x)$ and the overall noise variance function $\sigma^2(x)$, (2.14) avoids such extra smoothing steps and can adapt to the sparse or dense setting through the self-normalizer $U_n(x)$.

To select the bandwidth b, we adopt subject-based cross-validation (Rice and Silverman, 1991). The idea is to leave one subject out in model fitting, validate the fitted model using the left-out subject, and choose the optimal bandwidth by

minimizing the prediction error:

$$b^* = \underset{b}{\operatorname{argmin}}\, \mathrm{SJCV}(b), \quad \mathrm{SJCV}(b) = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \{Y_{ij} - \hat{\mu}^{(-i)}(X_{ij})\}^2, \qquad (2.15)$$

where $\hat{\mu}^{(-i)}(x)$ represents the estimator of $\mu(x)$ based on data from all but the ith subject.

In practice, it is difficult to estimate the bias $b^2\rho(x)$ due to the unknown derivatives f', μ', μ''. In our simulations, we use $K(u) = 2G(u) - G(u/\sqrt{2})/\sqrt{2}$ with $G(u)$ the standard normal density. Then $\int_{\mathbb{R}} u^2 K(u) du = 0$ and $\rho(x) = 0$. However, this does not solve the bias issue. For example, if f and μ are four times differentiable, then we have the higher order bias term $O(b^4)$. The bias issue is inherently difficult and there is no good solution so far.

2.3.2 Self-normalization based on recursive estimates

In this section we introduce another self-normalization method based on recursive estimates. For $m = 1, \ldots, n$, denote by $\hat{\mu}_m(x)$ the estimator in (2.3) based on observations from the first m subjects. Then $\hat{\mu}_1(x), \ldots, \hat{\mu}_n(x)$ are estimates of $\mu(x)$ with increasing accuracy. Moreover, $\hat{\mu}_m(x)$ has similar asymptotic normality as in (2.10)–(2.11). For example, for each $0 < t \leq 1$, the counterpart of (2.10) for sparse data is $\sqrt{(ntb)}\{\hat{\mu}_{\lfloor nt \rfloor}(x) - \mu(x) - b^2\rho(x)\} \to N\{0, s^2_{\mathrm{sparse}}(x)\}$. Throughout, $\lfloor z \rfloor$ is the integer part of z. Therefore, $\hat{\mu}_n(x)$ and $\hat{\mu}_{\lfloor nt \rfloor}(x)$ have proportional convergence rates and the same limiting variance, which motivates us to consider certain ratios between $\hat{\mu}_n(x)$ and $\hat{\mu}_{\lfloor nt \rfloor}(x)$ to cancel out the convergence rates and limiting variance.

Since the above analysis holds for all $0 < t \leq 1$, we consider an aggregated version

$$T_n(x) = \frac{\hat{\mu}_n(x) - \mu(x) - b^2\rho(x)}{V_n(x)}, \quad V_n(x) = n^{-3/2} \Big\{ \sum_{m=\lfloor cn \rfloor}^{n} m^2 |\hat{\mu}_m(x) - \hat{\mu}_n(x)|^2 \Big\}^{1/2}.$$

Throughout $c > 0$ is a small constant included to avoid unstable estimation at the boundary. By our simulations, $c = 0{\cdot}1$ works reasonably well. Intuitively, we may interpret $\hat{\mu}_m(x), m = 1, \ldots, n$, as observations from a population with mean

$\mu(x)$ and treat $\hat{\mu}_n(x)$ as sample average. Thus, $V_n(x)$ can be viewed as a weighted sample standard deviation with the weight m^2 reflecting the accuracy of $\hat{\mu}_m(x)$, and $V_n(x)$ mimics the usual normalizer in the Student-t distribution.

Theorem 3. *Assume the conditions in Theorem 1 on page 10. Let $\{B_t\}$ be a standard Brownian motion. Then $T_n(x) \to B_1/\{\int_c^1 (B_t - tB_1)^2 dt\}^{1/2}$ under either the sparse or the dense settings.*

By Theorem 3, an asymptotic pointwise $1-\alpha$ confidence interval for $\mu(x)$ is $\hat{\mu}_n(x) - b^2\hat{\rho}(x) \pm q_{1-\alpha/2} V_n(x)$, where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the limiting distribution. The latter confidence interval is the same for both scenarios, with the convergence rate and limiting variance being built into the self-normalizer $V_n(x)$ implicitly. Our method can be viewed as an extension of the parametric self-normalization methods in Lobato (2001), Kiefer and Vogelsang (2005) and Shao (2010) for time series data to the nonparametric longitudinal model (2.2).

In practice, however, subjects have no natural ordering, and we can use the average of multiple copies of $V_n^2(x)$ through permuting the subjects. For a large n, since it is computationally infeasible to enumerate all permutations, we consider only a fixed number, say T, of random permutations. Denote the corresponding $V_n(x)$ by $V_n^1(x), \ldots, V_n^T(x)$. Consider

$$\tilde{T}_n(x) = \frac{\hat{\mu}_n(x) - \mu(x) - b^2\rho(x)}{\tilde{V}_n(x)}, \quad \tilde{V}_n^2(x) = \frac{1}{T}\sum_{r=1}^{T} \{V_n^r(x)\}^2.$$

By the above analysis, the asymptotic distribution of $\tilde{T}_n(x)$ is the same under both the sparse and dense settings. However, it is not clear whether $\tilde{T}_n(x)$ is asymptotically normally distributed. Nevertheless, in light of the asymptotic normality of $\hat{\mu}_n(x)$, the proof of Theorem 3 and $E\{\int_c^1 (B_t - tB_1)^2 dt\} = (1 - 3c^2 + 2c^3)/6$, we propose the pointwise confidence interval

$$\hat{\mu}_n(x) - b^2\hat{\rho}(x) \pm z_{1-\alpha/2}\tilde{V}_n(x)\sqrt{c_1}, \quad c_1 = 6/(1 - 3c^2 + 2c^3). \tag{2.16}$$

Here $z_{1-\alpha/2}$ is defined in (2.12). We call it the rule-of-thumb self-normalization based confidence interval. Our quantile-quantile studies show that the empirical quantile of $\tilde{T}_n(x)$ with 200 permutations matches well with that of $N(0, c_1)$ under

the settings in Section 2.4.

2.4 Numerical results

Following Li and Hsing (2010), we consider the model

$$Y_{ij} = \mu(X_{ij}) + \sum_{k=1}^{3} \alpha_{ik}\Phi_k(X_{ij}) + \sigma\epsilon_{ij} \quad (i = 1, \ldots, n;\ j = 1, \ldots, n_i),$$

where $\alpha_{ik} \sim N(0, \omega_k)$ and $\epsilon_{ij} \sim N(0, 1)$. Let $\mu(x) = 5(x-0.6)^2, \Phi_1(x) = 1, \Phi_2(x) = \sqrt{2}\sin(2\pi x), \Phi_3(x) = \sqrt{2}\cos(2\pi x)$, $(\omega_1, \omega_2, \omega_3)$ =(0·6,0·3,0·1), and $n = 200$. Then the variance function $\gamma(x, x)$ =0·6+0·6sin$^2(2\pi x)$+0·2cos$^2(2\pi x)$. Two noise levels $\sigma = 1, 2$ are considered. The design points X_{ij} are uniformly distributed on $[0, 1]$. For the vector $N = (n_1, \ldots, n_n)$ of the number of measurements on individual subjects, we consider four cases

$$N_1: \quad n_i \sim U[\{2, 3, \ldots, 8\}]; \qquad N_2: \quad n_i \sim U[\{15, 16, \ldots, 35\}]; \tag{2.17}$$

$$N_3: \quad n_i \sim U[\{30, 31, \ldots, 70\}]; \qquad N_4: \quad n_i \sim U[\{150, 151, \ldots, 250\}] \tag{2.18}$$

Here $U[\mathcal{D}]$ stands for the discrete uniform distribution on a finite set $\mathcal{D}$.

We compare six confidence intervals: the two self-normalization based confidence intervals in (2.14) and (2.16) with 200 permutations, the asymptotic normality based confidence intervals (2.12)–(2.13) assuming sparse and dense data, respectively, the bootstrap confidence interval with 200 bootstrap replications from sampling subjects with replacement, and the confidence interval

$$\hat{\mu}_n(x) - b^2\hat{\rho}(x) \pm z_{1-\alpha/2}n^{-1/2}\Big\{(1-\hat{\tau})\gamma(x, x) + \hat{\tau}\psi_K\frac{\gamma(x, x) + \sigma^2(x)}{bf(x)}\Big\}^{1/2}. \tag{2.19}$$

The confidence interval (2.19) is practically infeasible as we need to estimate the unknown functions. Nevertheless, by using the true theoretical limiting variance function in (2.9), (2.19) serves as a standard against which we can measure the performance of other confidence intervals. When using the local linear method in Li and Hsing (2010) to estimate $\gamma(x, x)$, we found that negative estimates of $\gamma(x, x)$ occur frequently, especially when the noise level σ is high. For the purpose

of comparison, we use the true functions $\gamma(x,x), \sigma^2(x)$ and $f(x)$ to implement (2.12)–(2.13).

We consider two criteria: empirical coverage probabilities and lengths of confidence intervals. Let $x_1 < \cdots < x_{20}$ be 20 grid points evenly spaced on $[0{\cdot}1, 0{\cdot}9]$. For each x_j and a given nominal level, we construct confidence intervals for $\mu(x_j)$, and compute the empirical coverage probabilities based on 1000 replications. For each of the six confidence intervals, we average their empirical coverage probabilities and lengths at 20 grid points. To facilitate computations in bandwidth selection, instead of using (2.15) for each replication, we set b to be the average of 20 optimal bandwidths in (2.15) based on 20 replications from each set of parameter choices.

The results are presented in Table 2.1 on page 20. The performance of the confidence intervals (2.12)–(2.13) depends on whether the data are sparse or dense. As we increase the number of measurements on each subject from the sparse setting N_1 to the dense setting N_4, (2.12) under sparse assumption performs increasingly worse whereas (2.13) under dense assumption performs increasingly better. The simulation study further confirms the theoretical results in Theorem 1 on page 10 that the confidence intervals (2.12)–(2.13) perform well only under their corresponding sparse or dense assumption. By contrast, the self-normalization based confidence intervals (2.14) and (2.16) deliver robust and superior performance: (i) they have similar widths but slightly better coverage probabilities than the bootstrap confidence interval; and (ii) they perform similar to the infeasible confidence interval (2.19) with true functions. Finally, (2.14) and (2.16) have comparable performance.

2.5 Regularity conditions and proofs

Assumption. (i) $K(\cdot)$ is bounded, symmetric, and has bounded support and bounded derivative. (ii) $\{v_i(\cdot)\}_i, \{X_{ij}\}_{ij}, \{\epsilon_{ij}\}_{ij}$ are independent and identically distributed and mutually independent. Furthermore, the density function $f(\cdot)$ of X_{ij} is twice continuously differentiable in a neighborhood of x and $f(x) > 0$. (iii) In a neighborhood of x, $\mu(\cdot)$ is twice continuously differentiable, $\sigma^2(\cdot)$ is continuously differentiable; in a neighborhood of (x,x), $\gamma(x,x') = \mathrm{cov}\{v_i(x), v_i(x')\}$ is continuously differentiable. Moreover, $\gamma(x,x) > 0, \sigma^2(x) > 0$. (iv) $E\{|v_i(\cdot) + \sigma(\cdot)\epsilon_{ij}|^4\}$ is

continuous in a neighborhood of x and $E\{|v_i(x)+\sigma(x)\epsilon_{ij}|^4\}<\infty$.

Assumption (i)–(iii) are standard conditions. Almost the same regularity conditions are imposed in Wu and Zhang (2002), too. To deal with the tails of $v_i(x)+\sigma(x)\epsilon_{ij}$, Assumption (iv) is required. Wu and Zhang (2002) assume a slightly stronger condition that $\{v_i(x)+\sigma(x)\epsilon_{ij}\}$ are uniformly bounded.

2.5.1 Proof of Theorem 1

Proof. Let ξ_i be defined in (2.6). Recall the decomposition (2.5). Write

$$H_n = \sum_{i=1}^{n}\nu_i, \quad \nu_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\nu_{ij}, \quad \nu_{ij}=K\Big(\frac{X_{ij}-x}{b}\Big), \tag{2.20}$$

$$I_n = \sum_{i=1}^{n}\zeta_i, \quad \zeta_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\zeta_{ij}, \quad \zeta_{ij}=\{\mu(X_{ij})-\mu(x)\}K\Big(\frac{X_{ij}-x}{b}\Big). \tag{2.21}$$

By the symmetry of K and Taylor's expansion, $E(\nu_{ij})=\{1+O(b^2)\}bf(x)$, $\mathrm{var}(\nu_{ij})=O(b)$, $E(\zeta_{ij})=b^3f(x)\rho(x)+o(b^3)$, $\mathrm{var}(\zeta_{ij})=O(b^3)$. In either the sparse or the dense case, $E(\nu_i\mid n_i)=E(\nu_{ij})$ is non-random. Thus, $\mathrm{var}(\nu_i)=E\{\mathrm{var}(\nu_i\mid n_i)\}=\mathrm{var}(\nu_{ij})E(1/n_i)$ and $\mathrm{var}(H_n)=\sum_{i=1}^n\mathrm{var}(\nu_i)=O(b)\sum_{i=1}^nE(1/n_i)$. Write $\tau_n=n^{-1}\sum_{i=1}^nE(1/n_i)$. Then

$$H_n=E(H_n)+O_p\{\sqrt{\mathrm{var}(H_n)}\}=[1+O_p\{b^2+(nb/\tau_n)^{-1/2}\}]nbf(x). \tag{2.22}$$

Similarly, $I_n=nb^3f(x)\rho(x)+o(nb^3)+O_p\{\sqrt{(nb^3\tau_n)}\}$. Thus,

$$I_n/H_n=b^2\rho(x)+\delta_n, \quad \delta_n=o_p(b^2)+O_p\{\sqrt{(b\tau_n/n)}\}. \tag{2.23}$$

Dense case: Under given conditions, $\{nb^2f^2(x)\}^{-1}\mathrm{var}(\sum_{i=1}^n\xi_i)\to\gamma(x,x)$ and $\delta_n=o_p(n^{-1/2})$. . For distinct j,r,s,k, by the argument in (2.7), $E(\xi_{ij}\xi_{ir}\xi_{is}\xi_{ik})=O(b^4), E(\xi_{ij}^2\xi_{ir}\xi_{is})=O(b^3), E(\xi_{ij}^2\xi_{ir}^2)=O(b^2), E(\xi_{ij}^3\xi_{ir})=O(b^2), E(\xi_{ij}^4)=O(b)$. Thus, $\sum_{i=1}^nE(\xi_i^4)=O(nb^4)=o\{(b\sqrt{n})^4\}$. By the Lyapunov central limit theorem, $\sum_{i=1}^n\xi_i/\{b\sqrt{n}f(x)\}\to N\{0,\gamma(x,x)\}$.

Sparse case: In (2.5), $\xi_1,\ldots,\xi_n$ are independent and identically distributed. The result follows from $\delta_n=o_p\{(nb)^{-1/2}\}$ and $\mathrm{var}(\xi_i)=E\{\mathrm{var}(\xi_i\mid n_i)\}\approx b\tau\psi_Kf(x)\{\gamma(x,x)+\sigma^2(x)\}$. ♢

2.5.2 Proof of Theorem 2

Proof. By Theorem 1, it suffices to show $nU_n^2(x) \to s^2_{\text{dense}}(x)$ for dense data or $nbU_n^2(x) \to s^2_{\text{sparse}}(x)$ for sparse data. For convenience, write $K_{ij} = K\{(X_{ij}-x)/b\}$. Let

$$S_n = \sum_{i=1}^{n} \Big[\frac{1}{n_i}\sum_{j=1}^{n_i}\{Y_{ij} - \hat{\mu}_n(X_{ij})\}K_{ij}\Big]^2 = \sum_{i=1}^{n}(\xi_i^2 + \eta_i^2 + 2\xi_i\eta_i), \tag{2.24}$$

where ξ_i is defined in (2.6) and $\eta_i = {n_i}^{-1}\sum_{j=1}^{n_i}\{\mu(X_{ij}) - \hat{\mu}(X_{ij})\}K_{ij}$. By Theorem 3.1 in Li and Hsing (2010), $|\hat{\mu}(z) - \mu(z)| = O_p(\ell_n)$ uniformly for z in the neighborhood of x, where $\ell_n = b^2 + (n/\log n)^{-1/2}$ for dense data or $\ell_n = b^2 + (nb/\log n)^{-1/2}$ for sparse data. Then $\eta_i = O_p(\ell_n)n_i^{-1}\sum_{j=1}^{n_i}|K_{ij}|$. Using $\xi_i = n_i^{-1}\sum_{j=1}^{n_i}\xi_{ij}$, where ξ_{ij} is defined in (2.6), we obtain

$$\begin{aligned}\sum_{i=1}^{n}|\eta_i^2 + 2\xi_i\eta_i| &= O_p(\ell_n^2)\sum_{i=1}^{n}\Big(\frac{1}{n_i}\sum_{j=1}^{n_i}|K_{ij}|\Big)^2 + O_p(\ell_n)\sum_{i=1}^{n}\frac{1}{n_i^2}\sum_{j=1}^{n_i}|\xi_{ij}|\sum_{j=1}^{n_i}|K_{ij}| \\ &\le O_p(\ell_n)J_n, \quad J_n = \sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}K_{ij}^2 + \sum_{i=1}^{n}\frac{1}{n_i^2}\sum_{j=1}^{n_i}\sum_{j'=1}^{n_i}(K_{ij}^2 + \xi_{ij'}^2).\end{aligned}$$

Here we have used $\ell_n^2 = o(\ell_n)$, $(\sum_{j=1}^{n_i}|K_{ij}|)^2 \le n_i\sum_{j=1}^{n_i}K_{ij}^2$, $2|K_{ij}\xi_{ij'}| \le K_{ij}^2 + \xi_{ij'}^2$. By $E(K_{ij}^2) = O(b)$ and $E(\xi_{ij}^2) = O(b)$, $E(J_n) = O(nb)$. Thus, $\sum_{i=1}^{n}|\eta_i^2 + 2\xi_i\eta_i| = O_p(nb\ell_n)$. By (2.24) and the independence of $\xi_1, \ldots, \xi_n$,

$$S_n = \sum_{i=1}^{n}E(\xi_i^2) + O_p(\chi_n), \quad \chi_n = \Big\{\sum_{i=1}^{n}\text{var}(\xi_i^2)\Big\}^{1/2} + nb\ell_n.$$

From the proof of Theorem 1, $\{nb^2f^2(x)\}^{-1}\sum_{i=1}^{n}E(\xi_i^2) \to s^2_{\text{dense}}(x)$ for dense data or $\{nbf^2(x)\}^{-1}\sum_{i=1}^{n}E(\xi_i^2) \to s^2_{\text{sparse}}(x)$ for sparse data. By (2.22), $H_n = \{1 + o_p(1)\}nbf(x)$. Thus, it remains to show $\chi_n = o(nb^2)$ for dense data or $\chi_n = o(nb)$ for sparse data. In the dense case, by the proof of Theorem 1, $\sum_{i=1}^{n}\text{var}(\xi_i^2) \le \sum_{i=1}^{n}E(\xi_i^4) = O(nb^4)$, and consequently $\chi_n = O\{\sqrt{nb^2 + nb^3 + b}\sqrt{(n\log n)}\} = o(nb^2)$. In the sparse case, by the proof of the dense case in Theorem 1, $E(\xi_i^4 \mid n_i) = O(1)n_i^{-4}(n_i^4b^4 + n_i^3b^3 + n_i^2b^2 + n_ib)$, $E(\xi_i^4) = E\{E(\xi_i^4 \mid n_i)\} = O(b)$, and thus $\chi_n = O\{\sqrt{(nb)} + nb^3 + \sqrt{(nb\log n)}\} = o(nb)$. ♢

2.5.3 Proof of Theorem 3

Proof. Recall $s_{\text{sparse}}(x)$ and $s_{\text{dense}}(x)$ in Theorem 1. Let $\Gamma_n = nbf(x)/\Lambda_n$, $\Lambda_n = \sqrt{(nb)f(x)s_{\text{sparse}}(x)}$ for sparse data or $\Lambda_n = b\sqrt{nf(x)s_{\text{dense}}(x)}$ for dense data. Suppose we can show the weak convergence

$$\{\Gamma_n t\{\hat{\mu}_{\lfloor nt \rfloor}(x) - \mu(x) - b^2\rho(x)\}\}_{c \le t \le 1} \to \{B_t\}_{c \le t \le 1}. \tag{2.25}$$

For convenience, we write $\mathcal{L}_2(g) = \{\int_c^1 |g(t)|^2 dt\}^{1/2}$ and suppress the argument x. By (2.25) and the continuous mapping theorem, $(\hat{\mu}_n - \mu - b^2\rho)/\mathcal{L}_2\{t(\hat{\mu}_{\lfloor nt \rfloor} - \hat{\mu}_n)\} \to B_1/\mathcal{L}_2(B_t - tB_1)$. By $|n^{-1}\lfloor nt \rfloor - t| \le n^{-1}$ for $t \in [c, 1]$, $\mathcal{L}_2\{t(\hat{\mu}_{\lfloor nt \rfloor} - \hat{\mu}_n)\}$ is asymptotically equivalent to $\mathcal{L}_2\{n^{-1}\lfloor nt \rfloor(\hat{\mu}_{\lfloor nt \rfloor} - \hat{\mu}_n)\} = V_n(x)$, where $V_n(x)$ is defined in $T_n(x)$. This completes the proof.

It remains to show (2.25). Recall ν_i and ζ_i in (2.20)–(2.21). As in (2.3) and (2.5),

$$\hat{\mu}_{\lfloor nt \rfloor}(x) - \mu(x) - \frac{1}{H_{\lfloor nt \rfloor}} \sum_{i=1}^{\lfloor nt \rfloor} \zeta_i = \frac{W_n(t)}{H_{\lfloor nt \rfloor}}, \quad H_{\lfloor nt \rfloor} = \sum_{i=1}^{\lfloor nt \rfloor} \nu_i, \quad W_n(t) = \sum_{i=1}^{\lfloor nt \rfloor} \xi_i.$$

By Kolmogorov's maximal inequality for independent random variables,

$$\sup_{c \le t \le 1} |H_{\lfloor nt \rfloor} - E(H_{\lfloor nt \rfloor})| = \max_{\lfloor cn \rfloor \le m \le n} |H_m - E(H_m)| = O_p\Big[\Big\{\sum_{i=1}^{n} \text{var}(\nu_i)\Big\}^{1/2}\Big].$$

Thus, similar to (2.22), $H_{\lfloor nt \rfloor} = [1 + O_p\{b^2 + (nb/\tau_n)^{-1/2}\}]\lfloor nt \rfloor bf(x)$ uniformly in $c \le t \le 1$. Applying the same argument to (2.23) gives $\sum_{i=1}^{\lfloor nt \rfloor} \zeta_i / H_{\lfloor nt \rfloor} = b^2\rho(x) + \delta_n$ uniformly, where δ_n is defined in (2.23). Thus it suffices to show $\{W_n(t)/\Lambda_n\}_{c \le t \le 1} \to \{B_t\}_{c \le t \le 1}$. The finite-dimensional convergence follows from the same argument in Theorem 1 and the Cramér–Wold device. It remains to prove the tightness. Let $c \le t < t' \le 1$. By independence,

$$\Delta_n(t, t') = E\Big\{\frac{W_n(t)}{\Lambda_n} - \frac{W_n(t')}{\Lambda_n}\Big\}^4 = \frac{1}{\Lambda_n^4}\Big\{\sum_{i=\lfloor nt \rfloor + 1}^{\lfloor nt' \rfloor} E(\xi_i^4) + 6 \sum_{\lfloor nt \rfloor + 1 \le i < k}^{\lfloor nt' \rfloor} E(\xi_i^2)E(\xi_k^2)\Big\}.$$

By the argument in the proof of Theorem 1, in the dense case, $E(\xi_i^2) = O(b^2)$, $E(\xi_i^4) = O(b^4)$, and thus $\Delta_n(t, t') = O\{|t - t'|/n + |t - t'|^2\}$; in the sparse case,

$E(\xi_i^4) = O(b)$, $E(\xi_i^2) = O(b)$, and thus $\Delta_n(t,t') = O\{|t-t'|/(nb) + |t-t'|^2\}$. This proves the tightness. ♢

Table 2.1. Average empirical coverage percentages and lengths, in brackets, of six confidence intervals. SN1 and SN2: the self-normalized confidence intervals in (2.14) and (2.16) with 200 permutations, respectively; NS and ND: the asymptotic normality based confidence intervals (2.12)–(2.13) assuming sparse and dense data, respectively; NSD: the infeasible confidence interval in (2.19); BS: bootstrap confidence interval; N_1–N_4: the numbers of measurements on individual subjects in (2.17)–(2.18).

$1-\alpha$	σ	N	SN1	SN2	NS	ND	NSD	BS
90%	1	N_1	88·1 (0·381)	88·6 (0·386)	82·8 (0·331)	66·5 (0·236)	89·2 (0·389)	87·4 (0·377)
		N_2	88·9 (0·288)	89·2 (0·290)	68·0 (0·178)	81·0 (0·236)	89·4 (0·292)	88·1 (0·284)
		N_3	89·8 (0·262)	89·8 (0·263)	57·8 (0·126)	86·3 (0·236)	90·3 (0·265)	89·1 (0·258)
		N_4	88·4 (0·246)	88·5 (0·247)	37·3 (0·076)	86·9 (0·236)	88·6 (0·247)	87·5 (0·242)
	2	N_1	88·8 (0·528)	89·3 (0·534)	86·5 (0·497)	51·7 (0·236)	89·4 (0·537)	87·8 (0·523)
		N_2	88·6 (0·330)	88·7 (0·332)	75·8 (0·243)	74·3 (0·236)	89·3 (0·335)	87·9 (0·326)
		N_3	89·5 (0·293)	89·4 (0·294)	69·5 (0·183)	81·1 (0·236)	90·1 (0·297)	88·8 (0·289)
		N_4	88·4 (0·257)	88·6 (0·257)	48·6 (0·106)	85·1 (0·236)	88·7 (0·258)	87·6 (0·252)
95%	1	N_1	93·6 (0·454)	94·0 (0·460)	89·7 (0·394)	75·2 (0·281)	94·6 (0·464)	92·9 (0·446)
		N_2	94·1 (0·343)	94·3 (0·345)	76·4 (0·212)	88·1 (0·281)	94·7 (0·347)	93·4 (0·335)
		N_3	95·0 (0·312)	95·1 (0·314)	66·0 (0·150)	92·1 (0·281)	95·3 (0·316)	94·0 (0·305)
		N_4	94·2 (0·293)	94·3 (0·294)	43·7 (0·090)	92·9 (0·281)	94·3 (0·294)	93·1 (0·286)
	2	N_1	94·2 (0·629)	94·4 (0·636)	92·6 (0·592)	59·7 (0·281)	94·8 (0·640)	93·2 (0·618)
		N_2	93·9 (0·393)	94·0 (0·395)	83·6 (0·289)	82·3 (0·281)	94·2 (0·399)	93·0 (0·385)
		N_3	94·7 (0·349)	94·8 (0·351)	77·9 (0·219)	88·0 (0·281)	95·1 (0·354)	93·8 (0·341)
		N_4	94·1 (0·306)	94·1 (0·307)	56·4 (0·127)	91·6 (0·281)	94·2 (0·308)	93·0 (0·298)

Chapter 3

Efficient estimation for time-varying coefficient longitudinal models

3.1 Introduction

Driven by applications from medical, environmental, and social sciences, longitudinal data analysis has become one of the most active research areas; see, e.g., the monograph of Fitzmaurice, Laird and Ware (2004). Denote by $(t_{ij}, X_i(t_{ij}), Y_{ij}), j = 1, \ldots, m_i, i = 1, \ldots, n$, the measurement time t_{ij}, d-dimensional covariates $X_i(t_{ij})$, and one-dimensional response Y_{ij} from subject i, where $m_1, \ldots, m_n$ are the number of measurements from n subjects, we consider the time-varying coefficient longitudinal model [Hoover et al. (1998)]:

$$Y_{ij} = \alpha(t_{ij}) + X_i(t_{ij})^T \beta(t_{ij}) + \varepsilon_{ij}, \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n, \tag{3.1}$$

where $\alpha(\cdot)$ is the time-varying trend intercept, $\beta(\cdot)$ is the d-dimensional vector of coefficient functions of interest, and ε_{ij} is the noise. As a natural extension of the classical linear models, (3.1) allows the influence of covariates to vary over time, and thus entertains both the flexibility of nonparametric modeling and the interpretability of linear models. See Wu and Yu (2002) for a survey of related contributions.

For nonparametric estimation of the coefficient $\beta(\cdot)$, the widely used local least-squares (LS, hereafter) regression [Fan and Gijbels (1996)] minimizes a local square

loss, and thus the resultant estimator has a local sample average interpretation. See Hoover et al. (1998), Wu, Chiang and Hoover (1998) and Fan and Zhang (2000) for local LS estimation of (3.1), and Zeger and Diggle (1994) for a semiparametric approach. While they are efficient for Gaussian errors, LS methods may perform poorly in the presence of extreme outliers.

Recently there has been a growing interest in studying the conditional quantile regression (QR, hereafter) [e.g., Koenker (2005); Wei and He (2006)] for longitudinal data as an alternative approach to the conditional mean based LS regression. For linear models with longitudinal data, He, Fu and Fung (2003), Koenker (2004) and Wang and Fygenson (2009) applied QR for parameter estimation. In a semiparametric setting with constant coefficient $\beta(\cdot) \equiv \beta$, He, Zhu and Fung (2002) studied median QR estimation. Honda (2004), Kim (2007) and Cai and Xu (2008) studied QR for varying coefficient models but not under the longitudinal setting. Using a B-spline basis function approach, Wang, Zhu and Zhou (2009) studied QR inferences for a partially linear varying-coefficient longitudinal model.

In general, both the LS method and the traditional QR method use only partial information (unless the criterion function coincides with the likelihood function): (weighted) sample average for the LS method and single sample quantile for the QR method. Intuitively, more efficient estimators can be potentially constructed by exploiting more information from the data. Unlike the LS method which focuses on the conditional mean, QR provides a way of estimating the whole conditional distribution and thus offers a natural framework for combining information across multiple quantiles.

Our goal is to construct an asymptotically efficient estimator for $\beta(\cdot)$ through combining information across multiple quantiles. In particular, we propose the Optimally Weighted Local Quantile Average Estimator (OWLQAE, hereafter) for the time-varying coefficient longitudinal model (3.1). While the literature on estimation via combining quantile information has focused on simple linear models [Koenker (1984); Portnoy and Koenker (1989); Zou and Yuan (2008); Bradic, Fan and Wang (2011)] and nonparametric regression models with i.i.d. symmetric error [Kai, Li and Zou (2010)], our development for the time-varying coefficient longitudinal model (3.1) involves novel theory and methodology.

We first show that the proposed OWLQAE is asymptotically efficient under

working independence (i.e., ignoring the within-subject dependence) in the sense that its asymptotic variance approaches the Cramér-Rao bound of the density of the error ε_{ij}. In the case with within-subject dependence, it is well-known that estimators developed for independent data are no longer efficient, and we construct efficient estimation through combining information across multiple quantiles coupled with a prewhitening technique. We consider model (3.1) with AR errors, and apply a prewhitening transformation to transform errors into i.i.d. innovations. When applied to the prewhitened data, the proposed prewhitened OWLQAE is asymptotically efficient in the sense that its asymptotic variance attains the Cramér-Rao bound of the i.i.d. innovations, which is the optimal bound and cannot be further improved. To implement the proposed OWLQAE and prewhitened OWLQAE, we develop fully data-driven bandwidth selection and optimal weight estimation procedure.

In summary, our methodology has three components: (i) combine information across quantiles under working independence; (ii) use a prewhitening technique to remove dependence; and (iii) address practical implementation issues. By integrating the above three components together, our proposal provides an appealing alternative, both theoretically and methodologically, over the local LS method, the QR method using a single quantile, and the weighted local LS method in the literature. Simulation studies convincingly demonstrate the superior performance of the proposed method.

Section 3.2 reviews local LS and local QR methods. Section 3.3 introduces OWLQAE under working independence and study its properties. Section 3.4 presents the prewhitened OWLQAE for dependent data. Section 3.5 addresses bandwidth selection and optimal weight estimation. Simulation studies and a real data analysis are presented in Section 3.6.

3.2 Review of local kernel smoothing methods

We briefly review the local linear LS and QR methods for model (3.1). Throughout $K(\cdot)$ is a kernel function and $b > 0$ is a bandwidth.

Assumptions and notations:

(A1) For each i, $\{X_i(t)\}_t$ is an independent realization of a process $\{X(t)\}_t$ such that $X(t)$ is bounded, $\mathbb{E}[X(t)] = 0$, and $\Gamma_X(t) = \mathbb{E}[X(t)X(t)^T]$ is positive-definite and differentiable. In practice we need not to centralize the covariates since the mean value can be absorbed into $\alpha(t)$. The measurement times $t_{i1} \leq \cdots \leq t_{im_i}$ are ordered statistics of m_i uniform random samples on an interval $[T_l, T_u]$, and ε_{ij} is independent of $t_{ij}, X_i(t_{ij})$.

(A2) Denote by f_ε and F_ε the density and distribution functions of ε_{ij}, respectively. f_ε is bounded, positive, and twice continuously differentiable on $\{v : 0 < F_\varepsilon(v) < 1\}$.

(A3) $\alpha(\cdot)$ and $\beta(\cdot)$ are four times continuously differentiable in a neighborhood of t.

(A4) Let $N = m_1 + \cdots + m_n$, $Nb \to \infty, Nb^9 \to 0$ and $N^{-1}(1/\sqrt{Nb} + b)\sum_{i=1}^n m_i^2 \to 0$.

(A5) K is symmetric with bounded support and bounded derivative. Write $\mu_K = \int_{\mathbb{R}} u^2 K(u)du$ and $\varphi_K = \int_{\mathbb{R}} K^2(u)du$.

Consider the special case $m_i \equiv m$, then the condition $N^{-1}(1/\sqrt{Nb}+b)\sum_{i=1}^n m_i^2 \to 0$ in (A4) holds if $m[b + (nb)^{-1}] \to 0$, which includes both sparse ($m_i \leq M$ for a constant M) and some dense ($m_i \to \infty$) data. Here we do not impose any dependence structure on the error process $\{\varepsilon_{ij}\}_{j=1}^{m_i}$. This nice feature is achieved at the cost of the loose bounds in (3.35) and (3.41) [Section 3.7], where we use the bound $\text{var}(Z_1+\cdots+Z_m) \leq m[\mathbb{E}(Z_1^2)+\cdots+\mathbb{E}(Z_m^2)]$ for random variables $Z_1, \ldots, Z_m$. For many practical longitudinal survey data (e.g., the data considered in Section 3.6.2), measurements are taken annually or biannually and thus we have sparse or moderate dense observations. On the other hand, under some short-memory mixing conditions, we can substantially weaken the condition in (A4).

3.2.1 Local linear least-squares (LS) regression

Let t be any given point. In model (3.1), by Taylor's linear approximation, as $t_{ij} \approx t$,

$$\alpha(t_{ij}) + X_i(t_{ij})^T\beta(t_{ij}) \approx \alpha(t) + \alpha'(t)(t_{ij} - t) + X_i(t_{ij})^T[\beta(t) + \beta'(t)(t_{ij} - t)]. \quad (3.2)$$

Based on (3.2), the widely used local linear LS regression [Fan and Gijbels (1996)] is

$$\underset{\beta,\beta^*,\alpha,\alpha^*}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{m_i} \Big\{Y_{ij} - \alpha - \alpha^*(t_{ij}-t) - X_i(t_{ij})^T[\beta + \beta^*(t_{ij}-t)]\Big\}^2 K\Big(\frac{t_{ij}-t}{b}\Big) \quad (3.3)$$

for a kernel function $K(\cdot)$ and a bandwidth $b > 0$. Denote by $\hat{\beta}_{\mathrm{LS}}(t)$ the β component in the solution. Assume $\sigma_\varepsilon^2 := \mathrm{var}(\varepsilon_{ij}) < \infty$. Under the conditions and notations in (A1)–(A5),

$$\sqrt{Nb}\Big\{\hat{\beta}_{\mathrm{LS}}(t) - \beta(t) - \frac{\beta''(t)\mu_K}{2}b^2\Big\} \Rightarrow N\Big(0, (T_u - T_l)\varphi_K \Gamma_X^{-1}(t)\sigma_\varepsilon^2\Big). \quad (3.4)$$

For Gaussian errors ε_{ij}, $\hat{\beta}_{\mathrm{LS}}(t)$ is the local maximum likelihood estimator and thus asymptotically efficient. For heavy-tail errors (e.g., Cauchy errors), $\hat{\beta}_{\mathrm{LS}}(t)$ may perform poorly.

3.2.2 Local linear quantile regression (QR) regression

Given (3.1), denote by $Q_\varepsilon(\tau)$ the τ-th quantile of ε_{ij}, and by $Q_{Y_{ij}}(\tau|t_{ij}, X_i(t_{ij}))$ the conditional τ-th quantile of Y_{ij} given $t_{ij}, X_i(t_{ij})$. By the independence of ε_{ij} and $(t_{ij}, X_i(t_{ij}))$,

$$Q_{Y_{ij}}(\tau|t_{ij}, X_i(t_{ij})) = \alpha_\tau(t_{ij}) + X_i(t_{ij})^T\beta(t_{ij}), \quad \text{where} \quad \alpha_\tau(t_{ij}) = \alpha(t_{ij}) + Q_\varepsilon(\tau) \quad (3.5)$$

Using a similar linear approximation in (3.2) for $Q_{Y_{ij}}(\tau|t_{ij}, X_i(t_{ij}))$, we have the local linear QR [as a counterpart of the local LS regression in (3.3)]:

$$\begin{aligned} &\Big(\hat{\beta}(t|\tau), \hat{\beta}^*(t|\tau), \hat{\alpha}(t|\tau), \hat{\alpha}^*(t|\tau)\Big) \\ = \ &\underset{\beta,\beta^*,\alpha,\alpha^*}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j=1}^{m_i} \rho_\tau\Big\{Y_{ij} - \alpha - \alpha^*(t_{ij}-t) - X_i(t_{ij})^T[\beta + \beta^*(t_{ij}-t)]\Big\} K\Big(\frac{t_{ij}-t}{b}\Big), \end{aligned} \quad (3.6)$$

where $\rho_\tau(z) = z(\tau - \mathbf{1}_{z\le 0})$ is the quantile loss function at a quantile $\tau \in (0,1)$ and $\mathbf{1}$ is the indicator function. In particular, $\tau = 0.5$ corresponds to the local median QR.

Theorem 4. *Assume (A1)–(A5). Let* $\mathcal{I}(\tau) = \tau(1-\tau)/f_\varepsilon^2(Q_\varepsilon(\tau))$. *For* $\tau \in (0,1)$, *we have*

$$\sqrt{Nb}\Big\{\hat{\beta}(t|\tau) - \beta(t) - \frac{\beta''(t)\mu_K}{2}b^2\Big\} \Rightarrow N\Big(0, (T_u - T_l)\varphi_K\Gamma_X^{-1}(t)\mathcal{I}(\tau)\Big). \qquad (3.7)$$

By Theorem 4 and the result (3.4), the asymptotic bias $\beta''(t)\mu_K b^2/2$ of $\hat{\beta}(t|\tau)$ is the same as that of $\hat{\beta}_{\mathrm{LS}}(t)$ and does not depend on τ, but its asymptotic variance is proportional to $\mathcal{I}(\tau)$. Thus, depending on $\mathcal{I}(\tau)$, each quantile τ is of different importance to the estimation of $\beta(t)$. Figure 3.1 shows that, median works best for $N(0,1)$, Cauchy, Student-t and Laplace distributions; for normal mixture I, extreme quantiles are more efficient and median is the worst; for normal mixture II, quantiles within $[0.3, 0.7]$ have roughly the same performance.

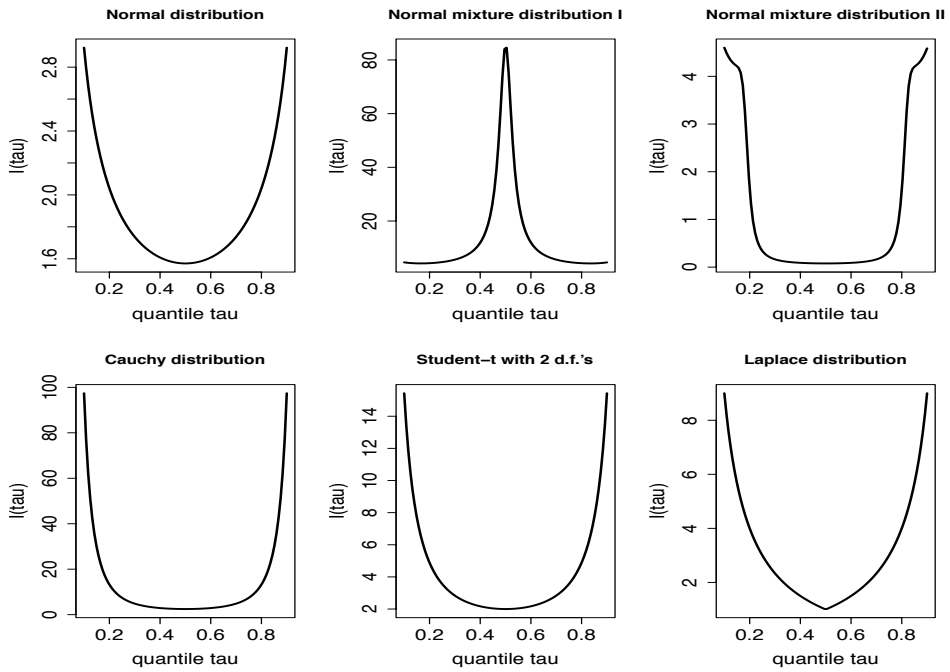

Figure 3.1. Plots of $\mathcal{I}(\tau)$ in Theorem 4 for six distributions: $N(0,1)$, normal mixture I: $0.5N(2,1) + 0.5N(-2,1)$, normal mixture II: $0.5N(0,1) + 0.5N(0, 0.5^6)$, Cauchy, Student-$t$ with 2 d.f.'s (t_2), Laplace.

Remark 2. *In this paper we focus on estimation of the slope* $\beta(t)$. *In (3.6), the component* α *is an estimator of* $\alpha_\tau(t)$ *defined in (3.5). Without further assumptions on* ε_{ij}, *the intercept* $\alpha(t)$ *is identifiable only up to a constant* $Q_\varepsilon(\tau)$. *If we assume*

ε_{ij} has median zero, i.e. $Q_\varepsilon(0.5) = 0$, then $\alpha(t)$ can be estimated by $\hat{\alpha}(t|0.5)$ in (3.6). In practice, the research interest usually focuses on the slope $\beta(t)$ as it measures the importance of different covariates.

3.3 Optimally weighted local quantile average estimator under working independence

The local LS in (3.3) uses the information of local sample average whereas the local QR in (3.6) uses the information of local sample quantile at a single quantile τ. Intuitively, more efficient estimators can be potentially constructed by exploiting more information from the data. Our proposed estimator improves upon the local LS and local QR in two important directions. First, in Section 3.3 , we propose a new estimator by optimally combining information across multiple quantiles and prove its asymptotic efficiency under working independence (i.e., we do not take into account the dependence structure). Second, in Section 3.4 on page 32, we propose a prewhitening transformation to transform dependent data into independent data to further improve the estimation efficiency.

3.3.1 The proposed OWLQAE under working independence

Let $0 < \tau_1 < \cdots < \tau_k$ be any given k quantiles, and we consider combining distributional information over these k quantiles. By Theorem 4 on page 26, $\hat{\beta}(t|\tau)$ in (3.6) is a consistent estimator of $\beta(t)$ for all $\tau \in (0, 1)$. To combine information across quantiles, we propose the following *weighted local quantile average estimator* (WLQAE) with weights $\omega = [\omega_1, \ldots, \omega_k]^T$:

$$\hat{\beta}_{\text{WLQAE}}(t|\omega) = \sum_{r=1}^{k} \omega_r \hat{\beta}(t|\tau_r), \quad \text{subject to} \quad \sum_{r=1}^{k} \omega_r = 1. \tag{3.8}$$

Since the local QR in (3.6) does not take into account the within-subject dependence structure, we call this estimator the WLQAE under *working independence*.

Theorem 5. *Assume the same conditions in Theorem 4 on page 26. Then*

$$\sqrt{Nb}\Big\{\hat{\beta}_{\mathrm{WLQAE}}(t|\omega) - \beta(t) - \frac{\beta''(t)\mu_K}{2}b^2\Big\} \Rightarrow N\Big(0, (T_u - T_l)\varphi_K\Gamma_X^{-1}(t)G(\omega)\Big),$$

where $G(\omega)$ is a quadratic form given by

$$G(\omega) = \omega^T H\omega \qquad \textit{and} \qquad H = \left\{\frac{\min(\tau_r, \tau_s) - \tau_r\tau_s}{f_\varepsilon(Q_\varepsilon(\tau_r))f_\varepsilon(Q_\varepsilon(\tau_s))}\right\}_{1\le r,s\le k}. \tag{3.9}$$

By Theorem 5, the asymptotic bias of $\hat{\beta}_{\mathrm{WLQAE}}(t|\omega)$ does not depend on ω. Thus, the optimal weight can be obtained by minimizing the asymptotic variance:

$$\omega^* = \underset{\omega}{\operatorname{argmin}}\Big\{G(\omega) : \omega_1 + \cdots + \omega_k = 1\Big\} = [\omega_1^*, \ldots, \omega_k^*]^T. \tag{3.10}$$

Note that H is symmetric and positive definite. By the Lagrange multiplier method,

$$\omega^* = \underset{\omega_1+\cdots+\omega_k=1}{\operatorname{argmin}} G(\omega) = \frac{H^{-1}e_k}{e_k^T H^{-1}e_k}, \quad G(\omega^*) = \frac{1}{e_k^T H^{-1}e_k}, \tag{3.11}$$

where $e_k = [1, \ldots, 1]^T$. With the optimal weight ω^*, we call $\hat{\beta}_{\mathrm{WLQAE}}(t|\omega^*)$ the optimal WLQAE (OWLQ AE).

3.3.2 Choices of quantiles

By Section 3.3.1 above, $G(\omega^*)$ in (3.11) is the smallest possible variance that one can achieve by using quantiles $\tau_1, \ldots, \tau_k$. A natural question is how to choose $\tau_1, \ldots, \tau_k$. As shown in Figure 3.1, different quantiles contain different information. For some distributions, median is the optimal quantile; for other distributions, small or large quantiles may work better.

Without prior information, a natural choice is the set of uniformly spaced quantiles

$$\mathcal{U}_k = \{\tau_r = r/(k+1), r = 1, \ldots, k\}. \tag{3.12}$$

Clearly, for any quantile $\tau \in (0, 1)$, there exists one quantile $\tau' \in \mathcal{U}_k$ such that

$|\tau' - \tau| \leq 1/(k+1)$. Thus, the distance between any quantile and $\mathcal{U}_k$ is at most $1/(k+1)$, which can be made sufficiently small for a reasonably large k. Moreover, Proposition 1 below shows that, for any given set of rational quantiles, there exists one $k^* \in \mathbb{N}$ such that OWLQAE with $\mathcal{U}_{k^*}$ outperforms OWLQAE with the given rational quantiles.

Proposition 1. *Recall $G(\omega^*)$ in (3.11). Let $\{\tau_1, \ldots, \tau_k\}$ be any set of quantiles such that $\tau_1, \ldots, \tau_k$ are rational numbers. Then there exists k^* such that $G(\omega^*|\mathcal{U}_{k^*}) \leq G(\omega^*|\{\tau_1, \ldots, \tau_k\})$.*

By Proposition 1, it is practically sufficient to consider uniform quantiles $\mathcal{U}_k$ for some k. Clearly, $G(\omega^*|\mathcal{U}_k)$ depends on the underlying distribution (which we have no control) and k. Large k leads to computational complexity while small k may result in efficiency loss. To understand how to choose k, Table 3.1 tabulates $G(\omega^*|\mathcal{U}_k)$ for different k for the six distributions in Figure 3.1 on page 26. Also, we tabulate the corresponding Cramér-Rao bound $1/\mathcal{F}(f_\varepsilon)$, where $\mathcal{F}(f_\varepsilon)$ is the Fisher information of the error density f_ε; see Section 3.3.3 below.

Table 3.1. $G(\omega^*|\mathcal{U}_k)$ and the theoretical limit $1/\mathcal{F}(f_\varepsilon)$ for the six distributions in Figure 3.1 on page 26.

	$G(\omega^*\|\mathcal{U}_k)$ with $k =$						
Distribution of ε	9	19	29	39	49	99	$1/\mathcal{F}(f_\varepsilon)$
$N(0,1)$	1.04	1.02	1.01	1.01	1.01	1.00	1.00
Mixture I	1.52	1.44	1.42	1.40	1.40	1.38	1.38
Mixture II	0.050	0.049	0.048	0.048	0.047	0.047	0.047
Cauchy	2.07	2.02	2.01	2.00	2.00	2.00	2.00
t_2	1.71	1.68	1.67	1.67	1.67	1.67	1.67
Laplace	1.00	1.00	1.00	1.00	1.00	1.00	1.00

From Table 3.1, as k increases, $G(\omega^*|\mathcal{U}_k)$ stabilizes quickly to the Cramér-Rao bound $1/\mathcal{F}(f_\varepsilon)$. With $k = 9$ uniformly spaced quantiles, $G(\omega^*|\mathcal{U}_k)$ is already very close to $1/\mathcal{F}(f_\varepsilon)$. Thus, in practice, we recommend using $k = 9$ uniformly spaced quantiles $\mathcal{U}_9$.

3.3.3 Asymptotic efficiency as $k \to \infty$

The result in Table 3.1 suggests the conjecture $\lim_{k\to\infty} G(\omega^*|\mathcal{U}_k) = 1/\mathcal{F}(f_\varepsilon)$. Theorem 6 below confirms it under the condition (A6), which holds for the distributions in Table 3.1,

(A6) For $g(\tau) = f_\varepsilon(Q_\varepsilon(\tau))$, $[g^2(\tau) + g^2(1-\tau)]/\tau + \tau^2 \int_\tau^{1-\tau} |g''(t)|^2 dt \to 0$ as $\tau \to 0$.

Theorem 6. *(i) Let $q_r = f_\varepsilon(Q_\varepsilon(\tau_r)), r = 1, \ldots, k$, and $q_0 = q_{k+1} = 0$. If we use $\mathcal{U}_k$, then*

$$\omega_r^* = \frac{(2q_r - q_{r-1} - q_{r+1})q_r}{\sum_{r=1}^k (2q_r - q_{r-1} - q_{r+1})q_r}, \quad r = 1, \ldots, k. \tag{3.13}$$

(ii) Furthermore, under (A6), we have the convergence to the inverse of Fisher information:

$$\lim_{k\to\infty} G(\omega^*|\mathcal{U}_k) = \frac{1}{\mathcal{F}(f_\varepsilon)}, \quad \textit{where} \quad \mathcal{F}(f_\varepsilon) = \int_{\mathbb{R}} \frac{[f_\varepsilon'(u)]^2}{f_\varepsilon(u)} du. \tag{3.14}$$

By Theorem 6(ii), as the number of quantiles $k \to \infty$, $G(\omega^*|\mathcal{U}_k)$ converges to the inverse of the Fisher information of f_ε, which is the well-known optimal Cramér-Rao bound. In this sense, by optimally combining information across quantiles $\mathcal{U}_k$, the OWLQAE $\hat{\beta}_{\text{WLQAE}}(t|\omega^*)$ is asymptotically efficient as we use more and more quantiles. This confirms the empirical result in Table 3.1 above. Also, as shown in Table 3.1, the convergence is very fast and it is practically sufficient to use $\mathcal{U}_9$. We point out that, the explicit expression in (3.13) is valid only when we use the set of uniform quantiles $\mathcal{U}_k$. If we use other choices of non-uniformly spaced quantiles, we need to use (3.11) to compute ω^*. This is another computational advantage of using $\mathcal{U}_k$ in addition to the nice feature in Proposition 1.

To combine information across $\hat{\beta}(t|\tau_1), \ldots, \hat{\beta}(t|\tau_k)$, the simplest method is to use uniform weights $[1/k, \ldots, 1/k]^T$ in (3.8). For nonparametric regression with i.i.d. symmetric errors, Kai, Li and Zou (2010) considered uniformly weighting quantile information. This approach ignores different importance among quantiles (as shown in Figure 3.1 on page 26) and thus may not be efficient. Clearly, $G(\omega^*|\mathcal{U}_k) \le G([1/k, \ldots, 1/k]^T|\mathcal{U}_k)$. In fact, by Kai, Li and Zou (2010), using uniform weight is asymptotically equivalent to LS regression as $k \to \infty$. Intuitively, as we take the simple average of all sample quantiles, we are using the sample average (LS). This also shows the importance of properly weighting quantile information.

3.3.4 Relative efficiency of LS, QR, and OWLQAE

In this section we compare the local LS estimator $\hat{\beta}_{\rm LS}(t)$ in Section 3.2.1 on page 24, the local QR estimator $\hat{\beta}(t|\tau)$ in Section 3.2.2 on page 25, and the proposed OWLQAE $\hat{\beta}_{\rm WLQAE}(t|\omega^*)$ in Section 3.3.1 on page 28. All the involved estimators have asymptotic normality of the form with different s^2:

$$\sqrt{Nb}\Big\{\hat{\beta}(t)-\beta(t)-\frac{\beta''(t)\mu_K}{2}b^2\Big\}\Rightarrow N\Big(0,(T_u-T_l)\varphi_K\Gamma_X^{-1}(t)s^2\Big). \tag{3.15}$$

For the generic estimator $\hat{\beta}$ in (3.15), its mean integrated squared error on interval $[\ell_1,\ell_2]$ is

$$\text{MISE}(\hat{\beta}|b)=\int_{\ell_1}^{\ell_2}\frac{1}{4}\|\beta''(t)\|^2\mu_K^2b^4dt+\int_{\ell_1}^{\ell_2}\frac{T_u-T_l}{Nb}\varphi_K\text{trace}\{\Gamma_X^{-1}(t)\}s^2dt,$$

where $\|z\|^2=z^Tz$ and trace$(\cdot)$ is the matrix trace operation. The optimal bandwidth is:

$$b^*=\underset{b}{\text{argmin}}\,\text{MISE}(\hat{\beta}|b)=\left[\frac{(T_u-T_l)\varphi_K\int_{\ell_1}^{\ell_2}\text{trace}\{\Gamma_X^{-1}(t)\}dt}{\mu_K^2\int_{\ell_1}^{\ell_2}\|\beta''(t)\|^2dt}\right]^{1/5}\Big(\frac{s^2}{N}\Big)^{1/5}. \tag{3.16}$$

The corresponding minimum mean integrated squared error is

$$\text{MISE}(\hat{\beta}|b^*)=\frac{5}{4}\Big[\mu_K^2\int_{\ell_1}^{\ell_2}|\beta''(t)|^2dt\Big]^{1/5}\Big[(T_u-T_l)\varphi_K\int_{\ell_1}^{\ell_2}\text{trace}\{\Gamma_X^{-1}(t)\}dt\Big]^{4/5}\Big(\frac{s^2}{N}\Big)^{4/5}.$$

For the local LS estimator $\hat{\beta}_{\rm LS}$, $s^2=\sigma_\varepsilon^2$. Thus, it is natural to define the asymptotic relative efficiency (ARE) of the estimator $\hat{\beta}$ in (3.15) relative to the benchmark $\hat{\beta}_{\rm LS}$ as

$$\text{ARE}(\hat{\beta})=\Big(\frac{\sigma_\varepsilon^2}{s^2}\Big)^{4/5},\qquad \hat{\beta}=\hat{\beta}(t|\tau),\ \hat{\beta}_{\rm WLQAE}(t|\omega^*). \tag{3.17}$$

A value of ARE$(\hat{\beta})>1$ indicates better performance of $\hat{\beta}$ relative to $\hat{\beta}_{\rm LS}$.

For the six distributions in Figure 3.1 on page 26, Table 3.2 tabulates ARE of the local median QR estimator $\hat{\beta}(t|0.5)$ in (3.6) and WLQAE $\hat{\beta}_{\rm WLQAE}(t|\omega)$ in (3.8) with the uniform weights $[1/k,\ldots,1/k]^T$ and the optimal weights ω^* in (3.11). For

all non-normal distributions considered, substantial efficiency gains over the local LS can be achieved by using $\hat{\beta}_{\mathrm{WLQAE}}(t|\omega)$ with optimal weights ω^*; for $N(0,1)$, they have almost equivalent performance.

Table 3.2. Asymptotic relative efficiency in (3.17) of $\hat{\beta}(t|0.5)$ and $\hat{\beta}_{\mathrm{WLQAE}}(t|\omega)$ with quantiles $\mathcal{U}_9$. UW and OW stand for the uniform and optimal weights, respectively. Since Cauchy and Student-t_2 have infinite variance, we use their truncated versions on $[-10, 10]$.

Distribution of ε	$\hat{\beta}_{\mathrm{LS}}(t)$	$\hat{\beta}(t\|0.5)$	$\hat{\beta}_{\mathrm{WLQAE}}(t\|\omega)$ UW	OW
$N(0,1)$	1	0.70	0.97	0.97
Mixture I	1	0.10	0.80	2.59
Mixture II	1	4.50	1.14	6.39
Cauchy on $[-10,10]$	1	2.20	0.87	2.53
t_2 on $[-10,10]$	1	1.54	1.32	1.74
Laplace	1	1.74	1.16	1.74

3.4 Prewhitened OWLQAE for dependent data

The OWLQAE in Section 3.3.1 on page 28 does not take into account the possible within-subject dependence, i.e., the method works the same regardless of the within-subject dependence. Under this working independence framework, Theorem 6 on page 30 shows that OWLQAE can asymptotically achieve the Cramér-Rao bound $1/\mathcal{F}(f_\varepsilon)$ of the density f_ε, which is the optimal bound if we ignore the dependence structure of the error process $\{\varepsilon_{ij}\}_{j=1}^{m_i}$. However, the latter bound is no longer optimal if we take into account possible dependence in the error process. In this section, we study the dependent case and show that substantial efficiency gain can be achieved through a prewhitening transformation.

3.4.1 Prewhitened OWLQAE: A general theoretical framework

To take into account dependence, different approaches have been proposed. For nonparametric regression with linear process errors, Xiao et al. (2003) approximated the process by an AR(p) and then applied a prewhitening transformation to obtain a new regression with independent error. For nonparametric regression with AR(p) errors, Li and Li (2009) proposed a penalized profile LS method. For

nonparametric regression with clustered or longitudinal data, Chen and Jin (2005) studied a covariance-weighted LS local polynomial method; see Zhu, Fung and He (2008) for a related LS regression spline approach.

Here we propose a prewhitening approach to transform dependent data into independent data. To illustrate the idea, suppose $\{\varepsilon_{ij}\}_{j=1}^{m_i}$ for subject i follows an AR(p) process

$$\varepsilon_{ij} = R(\varepsilon_{ij-1}, \ldots, \varepsilon_{ij-p}) + e_{ij}, \quad j \in \mathbb{N}, \tag{3.18}$$

for a function $R(\cdot)$ and i.i.d. innovations $e_{ij}, j \in \mathbb{N}$. Assume that e_{ij} is independent of the past $\varepsilon_{ij-1}, \varepsilon_{ij-2}, \ldots$. Consider the transformed response

$$Y_{ij}^* := Y_{ij} - R(\varepsilon_{ij-1}, \ldots, \varepsilon_{ij-p}), \tag{3.19}$$

then, given (3.1) and (3.18), we have

$$Y_{ij}^* = \alpha(t_{ij}) + X_i(t_{ij})^T \beta(t_{ij}) + e_{ij}. \tag{3.20}$$

Through the prewhitening transformation, the original model (3.1) with correlated errors ε_{ij} is transformed to the new model (3.20) with i.i.d. errors e_{ij}. We then apply the OWLQAE methodology in Section 3.3.1 on page 28 to the prewhitened data $(t_{ij}, X_i(t_{ij}), Y_{ij}^*)$, and we call the resultant estimator the prewhitened OWLQAE (POWLQAE). By Theorem 6 on page 30, we have

Theorem 7. *The prewhitened OWLQAE asymptotically attains the Cramér-Rao bound* $1/\mathcal{F}(f_e)$ *of* f_e*, the density of the i.i.d. innovations* e_{ij}*. Moreover, the optimal weights are*

$$\omega_r^+ = \frac{(2q_r^* - q_{r-1}^* - q_{r+1}^*)q_r^*}{\sum_{r=1}^k (2q_r^* - q_{r-1}^* - q_{r+1}^*)q_r^*}, \quad r = 1, \ldots, k, \tag{3.21}$$

where $q_r^* = f_e(Q_e(\tau_r)), r = 1, \ldots, k$*,* $q_0^* = q_{k+1}^* = 0$*, and* Q_e *is the quantile function of* e_{ij}*.*

The working independence OWLQAE in Section 3.3.1 on page 28 uses only (3.1) but ignores the dependence structure (3.18), and thus it can only attain the

Cramér-Rao bound $1/\mathcal{F}(f_\varepsilon)$ of ε_{ij}. Theorem 8 below gives a theoretical justification for the prewhitened OWLQAE over the working independence OWLQAE.

Theorem 8. *Suppose ε_{ij} follows (3.18). Then $1/\mathcal{F}(f_e) \leq 1/\mathcal{F}(f_\varepsilon)$.*

To appreciate the efficiency gain from prewhitening, assume that the error process $\{\varepsilon_{ij}\}_{j\in\mathbb{N}}$ in (3.1) follows the AR(2) process

$$\varepsilon_{ij} = \frac{2}{3}\varepsilon_{ij-1} - \frac{1}{3}\varepsilon_{ij-2} + e_{ij}, \tag{3.22}$$

where e_{ij} are i.i.d. variables from the six distributions in Figure 3.1 on page 26. Consider three cases

Case 1: Without prewhitening, i.e., the original model (3.1) with errors ε_{ij}.

Case 2: Partial prewhitening, i.e., model (3.1) with errors $\varepsilon_{ij} - 2/3\varepsilon_{ij-1}$.

Case 3: Full prewhitening, i.e., model (3.1) with errors $\varepsilon_{ij} - 2/3\varepsilon_{ij-1} + 1/3\varepsilon_{ij-2}$.

By Theorem 6(ii) on page 30, for the three cases above, the OWLQAE can asymptotically achieve the inverse of the Fisher information of ε_j, $\varepsilon_j - 2/3\varepsilon_{j-1}$, and $\varepsilon_j - 2/3\varepsilon_{j-1} + 1/3\varepsilon_{j-2}$, respectively. We tabulate the result in Table 3.3 below. Clearly, substantial efficiency gain can be achieved through prewhitening or partial prewhitening.

Table 3.3. Theoretical efficiency of the OWLQAE under the three settings above (without, partial and full prewhitening) for the six distributions in Figure 3.1 on page 26.

Distribution of e_j	Without prewhitening	Partial prewhitening	Full prewhitening
$N(0,1)$	1.49	1.16	1.00
Mixture I	5.68	3.43	1.38
Mixture II	0.356	0.230	0.047
Cauchy	9.61	6.00	2.00
t_2	3.80	2.64	1.67
Laplace	2.51	1.76	1.00

3.4.2 Prewhitened OWLQAE: A practical procedure

To implement the POWLQAE in Section 3.4.1 above, a key step is to obtain the prewhitened data (3.19). We propose the following prewhitened OWLQAE procedure:

(i) Use some preliminary estimates $\hat{\alpha}_0(\cdot)$ and $\hat{\beta}_0(\cdot)$ [e.g., median QR] to obtain $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\alpha}_0(t_{ij}) - X_i(t_{ij})^T\hat{\beta}_0(t_{ij})$. See Step I in Section 3.5.2 on page 38.

(ii) Based on $\hat{\varepsilon}_{ij}$, apply regression methods to (3.18) and obtain an estimate $\hat{R}$ of R.

(iii) In light of (3.19), compute the prewhitened data through $\hat{Y}_{ij} = Y_{ij} - \hat{R}(\hat{\varepsilon}_{ij-1}, \ldots, \hat{\varepsilon}_{ij-p})$.

(iv) Apply the OWLQAE in Section 3.3.1 on page 28 to $(t_{ij}, X_i(t_{ij}), \hat{Y}_{ij})$.

To implement the regression in step (ii), it is necessary to impose some structure on R to avoid the "curse of dimensionality". In this paper, we consider AR(p) model in (3.18):

$$\text{AR}(p): \quad \varepsilon_{ij} = \theta_1 \varepsilon_{ij-1} + \cdots + \theta_p \varepsilon_{ij-p} + e_{ij}, \quad j \in \mathbb{N}. \tag{3.23}$$

In their profile LS nonparametric regression, Li and Li (2009) imposed the same AR model. Furthermore, by Wold's decomposition, any covariance-stationary process has an MA(∞) representation, which can be approximated by an AR under the invertibility condition. This practice is also adopted in Xiao et al. (2003) for nonparametric regression with linear process errors. Thus, the AR structure (3.23) is not very restrictive. In the simulation studies of Section 3.6.1 on page 40, we also examine the performance of the POWLQAE when the errors follow nonlinear models. With a given p and using median QR, we implement step (ii) above as:

$$\hat{\theta} = \underset{\theta_1,\ldots,\theta_p}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=p+1}^{m_i} \left| \hat{\varepsilon}_{ij} - \theta_1 \hat{\varepsilon}_{ij-1} - \cdots - \hat{\theta}_p \hat{\varepsilon}_{ij-p} \right|. \tag{3.24}$$

Now we discuss the selection of p in (3.24). The main purpose of prewhitening is to "soak up" some of the temporal dependence in ε_{ij} and to obtain a prewhitened longitudinal model with residuals that are close to white noises. Note that only those prewhitened data with $p + 1 \leq j \leq m_i$ will be used in the OWLQAE. Consider the simple case $m_1 = \cdots = m_n = m$ for some m. For the POWLQAE with AR(p) fitting, the ratio of the effective sample size to the overall sample size is $[n(m-p)]/(nm) = 1 - p/m$, which decreases as p increases. In the dense setting $m \to \infty$, $1 - p/m \to 1$ and thus this is not an issue (at least asymptotically). In

such case, we can use the BIC criterion $\hat{p} = \text{argmin}_p \, \text{BIC}(p)$,

$$\text{BIC}(p) = 2 \log \left\{ \sum_{i=1}^{n} \sum_{j=p+1}^{m_i} \left| \hat{\varepsilon}_{ij} - \hat{\theta}_1 \hat{\varepsilon}_{ij-1} - \cdots - \hat{\theta}_p \hat{\varepsilon}_{ij-p} \right| \right\} + \frac{p \log N}{N}. \quad (3.25)$$

However, in practice, especially in the sparse setting with a small m, the factor $1 - p/m$ may play a non-negligible role. For example, if $m = 10$, then the effective sample size ratio is $90\%, 80\%, 70\%$ for $p = 1, 2, 3$, respectively, and thus a small p is preferred.

(**Rule-of-thumb choice:**) For practical longitudinal data, it is reasonable to assume that observations are more strongly affected by their nearby observations than by distant observations. In particular, one may even assume a Markovian structure. Our simulation studies show that the POWLQAE with a smaller p tends to outperform that with a larger p. Intuitively, for causal models with exponentially decaying dependence (such as causal AR models), much of the dependence can be captured by models with a small lag. In our simulations, using a smaller p works well even when the AR model is mis-specified. As a rule-of-thumb choice, we recommend $p = 1$ for very sparse data ($m \leq 10$) and $p = \lfloor m/10 \rfloor$ for $m > 10$ so that the effective sample size ratio is about $1 - p/m \approx 90\%$. This simple rule-of-thumb choice works quite well in our simulations.

3.4.3 Comparison with covariance-weighted local LS regression

For comparison, we briefly review Chen and Jin (2005)'s covariance-weighted LS (WLS) local polynomial method in our longitudinal setting. For subject i, denote by $V_i = \{\text{cov}(\varepsilon_{ij}, \varepsilon_{ij'})\}_{1 \leq j, j' \leq m_i}$ the $m_i \times m_i$ covariance matrix of the error process, and write

$$Y_i = \begin{bmatrix} Y_{i1} \\ \dots \\ Y_{im_i} \end{bmatrix}, \quad X_i = \begin{bmatrix} X_i(t_{i1})^T & X_i(t_{i1})^T(t_{i1} - t) & 1 & (t_{i1} - t) \\ \dots & & & \\ X_i(t_{im_i})^T & X_i(t_{im_i})^T(t_{im_i} - t) & 1 & (t_{im_i} - t) \end{bmatrix}, \quad \Theta = \begin{bmatrix} \beta \\ \beta^* \\ \alpha \\ \alpha^* \end{bmatrix}.$$

Define the diagonal matrix $W_i = \text{diag}\{\sqrt{K\{(t_{i1}-t)/b\}}, \ldots, \sqrt{K\{(t_{im_i}-t)/b\}}\}$. For comparison purpose, we use Gaussian kernel so that the I_i matrix in Equation (2) of Chen and Jin (2005) is the identity matrix. The local linear WLS regression is

$$\hat{\Theta} \;=\; \underset{\Theta}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i\Theta)^T W_i V_i^{-1} W_i (Y_i - X_i\Theta). \tag{3.26}$$

The β component of $\hat{\Theta}$ is the estimate of $\beta(t)$.

For multivariate normal errors, WLS is equivalent to the local MLE. If the error has infinite variance, WLS is not well-defined. By contrast, our prewhitened OWLQAE is always well-defined and asymptotically efficient. Another issue of WLS is that, for uncorrelated errors, V_i is the identity matrix and (3.26) becomes the local LS in (3.3). For example, consider the autoregressive conditional heteroscedastic (ARCH) model $\varepsilon_{ij} = e_{ij}\sqrt{a_0 + a_1\varepsilon_{ij-1}^2}, a_0 > 0, 0 \le a_1 < 1, \mathbb{E}(e_{ij}) = 0, \mathbb{E}(e_{ij}^2) = 1$, then $\{\varepsilon_{ij}\}_{j\in\mathbb{N}}$ are uncorrelated and WLS becomes LS.

3.5 Implementation

In this section we address the bandwidth selection and optimal weight estimation issues. As argued in Section 3.3.2 on page 28, throughout we use the uniform quantiles $\mathcal{U}_k$ so that we can use the short-cut expressions (3.13) and (3.21) to compute optimal weights.

3.5.1 Bandwidth selection

Bandwidth selection is an important issue in nonparametric regression. Denote by b_{LS} the optimal bandwidth for the local LS estimator in (3.3). To choose b_{LS}, we use the "leave-one-subject-out" cross-validation method [Rice and Silverman (1991)]:

$$b_{\text{LS}} = \underset{b}{\text{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left\{ Y_{ij} - \hat{\alpha}^{(-i)}(t_{ij}; b) - X_i(t_{ij})^T \hat{\beta}^{(-i)}(t_{ij}; b) \right\}^2. \tag{3.27}$$

Here $\hat{\alpha}^{(-i)}(\cdot;b)$ and $\hat{\beta}^{(-i)}(\cdot;b)$ are estimates of $\alpha(\cdot)$ and $\beta(\cdot)$ based on data from all but subject i using bandwidth b. Similarly, we choose bandwidth for WLS in Section 3.4.3 on page 37 by

$$b_{\text{WLS}} = \underset{b}{\text{argmin}} \sum_{i=1}^{n} (Y_i - X_i \hat{\Theta}^{(-i)})^T V_i^{-1} (Y_i - X_i \hat{\Theta}^{(-i)}). \tag{3.28}$$

Here $\hat{\Theta}^{(-i)}$ is the estimator based on data from all but subject i using bandwidth b.

Now we consider bandwidth selection for a generic estimator $\hat{\beta}$ satisfying (3.15). By (3.16), the optimal bandwidth b^* for $\hat{\beta}$ is related to b_{LS} through $b^*/b_{\text{LS}} = (s^2/\sigma_\varepsilon^2)^{1/5}$. Thus, with b_{LS} from (3.27), we estimate b^* through $b^* = b_{\text{LS}}(\hat{s}^2/\hat{\sigma}_\varepsilon^2)^{1/5}$, where $\hat{s}^2$ and $\hat{\sigma}_\varepsilon^2$ are estimates of s^2 and σ_ε^2, respectively, via the two-step procedure in Section 3.5.2 below. See Section 3.5.3 below for details on calculating bandwidths for OWLQAE and prewhitened OWLQAE.

3.5.2 Optimal weight estimation: Two-step procedure

To implement the OWLQAE in Section 3.3.1 on page 28 and the prewhitened OWLQAE in Section 3.4 on page 33, we need to estimate the weights ω^* in (3.13) and ω^+ in (3.21), respectively. Equivalently, we need to estimate $f_\varepsilon(Q_\varepsilon(\tau))$ and $f_e(Q_e(\tau))$. Note that $f_\varepsilon(Q_\varepsilon(\tau))$ and $f_e(Q_e(\tau))$ are invariant under the shift $\varepsilon + c$ and $e + c$ for any c, thus the inconsistent estimation (up to a constant shift; see Remark 1) of $\alpha(\cdot)$ is not an issue. We propose a two-step procedure below.

Step I: Use QR with $\tau = 0.5$ in (3.6) to obtain preliminary estimates $\hat{\alpha}_0(t)$ and $\hat{\beta}_0(t)$. Following Yu and Jones (1998), we use bandwidth $b = b_{\text{LS}}(\pi/2)^{1/5}$, where b_{LS} is the LS bandwidth in (3.27). Then obtain residuals $\hat{\varepsilon}_{ij} = Y_{ij} - \hat{\alpha}_0(t_{ij}) - X_i(t_{ij})^T \hat{\beta}_0(t_{ij})$.

To estimate $f_\varepsilon(Q_\varepsilon(\tau))$, we use the following Step II:

Step II: [Estimate $f_\varepsilon(Q_\varepsilon(\tau))$]. First, we estimate f_ε by the kernel density estimator:

$$\hat{f}_\varepsilon(u) = \frac{1}{N h_f} \sum_{i=1}^{n} \sum_{j=1}^{m_i} K\Big(\frac{u - \hat{\varepsilon}_{ij}}{h_f}\Big), \tag{3.29}$$

where $h_f = 0.9N^{-1/5}\min\{\text{sd}(\hat{\varepsilon}_{ij}), \text{IQR}(\hat{\varepsilon}_{ij})/1.34\}$ is the rule-of-thumb bandwidth in Silverman (1986), with $\text{sd}(\hat{\varepsilon}_{ij})$ and $\text{IQR}(\hat{\varepsilon}_{ij})$ being, respectively, the sample standard deviation and the sample interquartile of the residuals $\hat{\varepsilon}_{ij}$. Then we estimate $f_\varepsilon(Q_\varepsilon(\tau))$ by $\hat{f}_\varepsilon(\hat{Q}_\varepsilon(\tau))$, where $\hat{Q}_\varepsilon(\tau)$ is the sample τ-th quantile of $\hat{\varepsilon}_{ij}, j = 1, \ldots, m_i, i = 1, \ldots, n$.

To estimate $f_e(Q_e(\tau))$, we use the following Step II*:

Step II*: [Estimate $f_e(Q_e(\tau))$]. With $\hat{\varepsilon}_{ij}$ from Step I above, we use (3.24) to obtain $\hat{\theta}$. Then compute $\hat{e}_{ij} = \hat{\varepsilon}_{ij} - \hat{\theta}_1\hat{\varepsilon}_{ij-1} - \cdots - \hat{\theta}_p\hat{\varepsilon}_{ij-p}$. Finally, apply the same method in Step II above to $\hat{e}_{ij}$ to obtain $\hat{f}_e(\hat{Q}_e(\tau))$.

3.5.3 Implementation of OWLQAE and prewhitened OWLQAE

To implement OWLQAE in (3.8) with optimal weight ω^* in (3.13), we follow:

(i) Use the method in Section 3.5.2 to obtain the estimate $\hat{f}_\varepsilon(\hat{Q}_\varepsilon(\tau))$.

(ii) Use (3.13) to obtain the estimated optimal weights $\hat{\omega}^*$ and compute $G(\hat{\omega}^*)$ in (3.9).

(iii) By Section 3.5.1, compute bandwidth $b_{\text{WLQAE}} = b_{\text{LS}}\{G(\hat{\omega}^*)/\hat{\sigma}_\varepsilon^2\}^{1/5}$, where $\hat{\sigma}_\varepsilon^2$ is the sample variance of $\hat{\varepsilon}_{ij}$ [see Step I in Section 3.5.2], and b_{LS} is the LS bandwidth in (3.27).

(iv) In (3.6), use bandwidth b_{WLQAE} to obtain estimates $\hat{\beta}(t|\tau), \tau = \tau_1, \ldots, \tau_k$.

(v) Plug $\hat{\omega}^*$ and the estimates $\hat{\beta}(t|\tau), \tau = \tau_1, \ldots, \tau_k$, into (3.8) to obtain OWLQAE.

To implement the prewhitened OWLQAE with optimal weight ω^+ in (3.21), we use

(i*) Use the method in Section 3.5.2 to obtain the estimate $\hat{f}_e(\hat{Q}_e(\tau))$.

(ii*) Use (3.21) to obtain the estimated optimal weights $\hat{\omega}^+$ and compute $G^*(\hat{\omega}^+)$, where G^* is defined similarly to G in (3.9) with $f_\varepsilon(Q_\varepsilon(\tau))$ being replaced by $f_e(Q_e(\tau))$.

(iii*) As in Step (iii) above, compute $b_{\text{WLQAE}} = b_{\text{LS}}\{G^*(\hat{\omega}^+)/\hat{\sigma}_\varepsilon^2\}^{1/5}$, where $\hat{\sigma}_\varepsilon^2$ and b_{LS} are the same as in Step (iii) above [Note: still use $\hat{\varepsilon}_{ij}$ to compute $\hat{\sigma}_\varepsilon^2$].

(iv*) With $\hat{\varepsilon}_{ij}$ from Step I in Section 3.5.2, use (3.24) to obtain $\hat{\theta}$, and then compute the prewhitened data $\hat{Y}_{ij} = Y_{ij} - \hat{\theta}_1\hat{\varepsilon}_{ij-1} - \cdots - \hat{\theta}_p\hat{\varepsilon}_{ij-p}$.

(v*) In (3.6), use bandwidth b_{WLQAE} and prewhitened data $\hat{Y}_{ij}$ to obtain prewhitened estimates $\hat{\beta}^*(t|\tau), \tau = \tau_1, \ldots, \tau_k$.

(vi*) Plug $\hat{\omega}^+$ and $\hat{\beta}^*(t|\tau), \tau = \tau_1, \ldots, \tau_k$, into (3.8) to compute prewhitened OWLQAE.

3.6 Numerical results

3.6.1 Simulation studies

In (3.1), following Wu, Chiang and Hoover (1998), we consider the coefficient curves:

$$\alpha(t) = 15 + 20\sin\Big(\frac{t\pi}{60}\Big), \quad \beta_1(t) = 4 - \Big(\frac{t-20}{10}\Big)^2,$$
$$\beta_2(t) = 2 - 3\cos^2\Big\{\frac{(t-25)\pi}{15}\Big\}, \quad \beta_3(t) = -5 + \frac{(30-t)^3}{5000},$$

and the covariates $X_i = [X_{i1}, X_{i2}, X_{i3}]^T$ have time-independent i.i.d. $N(0, 0.25)$ components. Let $n = 100$ and the measurement times t_{ij} be uniformly distributed on $[0, 30]$. For the error process $\{\varepsilon_{ij}\}_{1\le j\le m_i}$, we consider two models

$$\text{(Model I) AR(2):} \quad \varepsilon_{ij} = \frac{5}{6}\varepsilon_{ij-1} - \frac{1}{6}\varepsilon_{ij-2} + e_{ij}, \quad j = 1, \ldots, m_i, \tag{3.30}$$
$$\text{(Model II) ARCH:} \quad \varepsilon_{ij} = \sqrt{1 + 0.2\varepsilon_{ij-1}^2}\, e_{ij}, \quad j = 1, \ldots, m_i, \tag{3.31}$$

where e_{ij} follows the six distributions in Table 3.2 on page 32. The nonlinear model (3.31) is used to examine the effect of model mis-specification of the prewhitened OWLQAE. To examine the effect of the sparseness or denseness of the data, we consider

$$\text{Sparse case :} \quad m_1 = \cdots = m_n = 10,$$
$$\text{Dense case :} \quad m_1 = \cdots = m_n = 20.$$

We compare the following five methods:

(1°) LS method in (3.3) with cross-validation bandwidth in (3.27).

(2°) Median quantile regression (MQR) in (3.6) with $\tau = 0.5$ and bandwidth selected as in Section 3.5.1 on page 38.

(3°) Chen and Jin (2005)'s WLS in (3.26) with cross-validation bandwidth in (3.28). While the covariance matrix V_i in (3.26) must be estimated in practice, we use their true value, which will favor WLS. For Model I, V_i is the $m_i \times m_i$ matrix (up to a constant factor) with (j, j') entry $4.8/2^{|j-j'|} - 2.7/3^{|j-j'|}, j, j' = 1, \ldots, m_i$. for Model II, V_i is the $m_i \times m_i$ identity matrix (up to a constant factor), and WLS reduces to LS.

(4°) Working independence OWLQAE with bandwidth selected as in Sections 3.5.1 on page 37 and 3.5.3 on page 39.

(5°) Prewhitened OWLQAE with bandwidth selected as in Sections 3.5.1 and 3.5.3. To examine the effect of the order p, we consider $p = 1, 2, 3$ in (3.24). As argued in Section 3.3.2 on page 29, we use uniform quantiles $\mathcal{U}_9$ [cf. equation (3.12)] for the two OWLQAE methods.

For an estimator $\hat{\beta}$, we consider its mean integrated squared error with 200 realizations:

$$\mathrm{MISE}(\hat{\beta}) = \frac{1}{12000} \sum_{i=1}^{3} \sum_{j=1}^{200} \sum_{s=1}^{20} [\hat{\beta}_{i,j}(t_s) - \beta_{i,j}(t_s)]^2,$$

where $t_1, \ldots, t_{20}$ are 20 evenly spaced points on $[5, 25]$, and $\hat{\beta}_{i,j}(t)$ is the estimator of $\beta_i(t)$ from the jth realization at time t. The empirical relative efficiency (ERE) of $\hat{\beta}$ relative to the benchmark $\hat{\beta}_{\mathrm{LS}}$ is defined as $\mathrm{ERE}(\hat{\beta}) = \mathrm{MISE}(\hat{\beta}_{\mathrm{LS}})/\mathrm{MISE}(\hat{\beta})$. A value of ERE ≥ 1 indicates that $\hat{\beta}$ outperforms $\hat{\beta}_{\mathrm{LS}}$.

The result is presented in Table 3.4 below. Overall, the proposed prewhitened OWLQAE delivers more robust and superior performance than other methods. As discussed in Section 3.4.2 on page 35, the effective sample size ratio $1 - p/m$ plays a non-negligible role in the sparse case $m = 10$, which explains the decreasing performance of prewhitened OWLQAE with $p = 1, 2, 3$ in the sparse case. On the other hand, in the dense case, the choices of $p = 1, 2, 3$ have comparable performance. It is interesting to see that, for the nonlinear ARCH model errors (Model II), the working independence OWLQAE and prewhitened OWLQAE with $p = 1$ are comparable and delivers much better overall performance than other methods. In both the sparse and dense cases, the rule-of-thumb choice of p in Section 3.4.2 works reasonably well. In summary, we conclude that the prewhitened OWLQAE

significantly outperforms existing methods, and in practice we recommend it with uniform quantiles $\mathcal{U}_9$ and the rule-of-thumb choice of p in Section 3.4.2.

Table 3.4. Empirical relative efficiency (relative to the benchmark $\hat{\beta}_{\text{LS}}$) for Model I–II in (3.30)–(3.31) with e_{ij} from the six distributions in Table 3.2 on page 32. LS, MQR, WLS, OWLQAE, and POWLQAE(p) stand for the LS, median QR with $\tau = 0.5$ in (3.6), Chen and Jin (2005)'s weighted LS, working independence OWLQAE, and prewhitened OWLQAE(p) with AR(p) fitting in (3.24), respectively.

Table 3.4 (a): Sparse case $m_1 = \cdots = m_n = 10$

						POWLQAE(p)		
	e_{ij} distribution	LS	MQR	WLS	OWLQAE	$p=1$	$p=2$	$p=3$
Model I	N(0,1)	1.00	0.71	1.51	0.94	1.48	1.27	0.76
	Mixture I	1.00	0.63	1.82	0.90	2.45	2.11	1.00
	Mixture II	1.00	1.12	1.54	1.17	4.85	5.07	1.49
	Cauchy on $[-10, 10]$	1.00	1.02	2.05	1.16	2.85	2.52	1.37
	t_2 on $[-10, 10]$	1.00	0.95	1.59	1.10	2.26	1.95	1.14
	Laplace	1.00	0.88	1.65	1.04	2.05	1.77	1.04
Model II	N(0,1)	1.00	0.72	1.00	0.94	0.91	0.77	0.46
	Mixture I	1.00	0.96	1.00	3.10	2.72	1.96	1.07
	Mixture II	1.00	3.64	1.00	3.72	3.48	2.64	0.80
	Cauchy on $[-10, 10]$	1.00	9.82	1.00	10.61	9.82	7.80	3.76
	t_2 on $[-10, 10]$	1.00	2.42	1.00	2.58	2.47	2.04	1.14
	Laplace	1.00	1.66	1.00	1.69	1.60	1.35	0.71

Table 3.4 (b): Dense case $m_1 = \cdots = m_n = 20$

						POWLQAE(p)		
	e_{ij} distribution	LS	MQR	WLS	OWLQAE	$p=1$	$p=2$	$p=3$
Model I	N(0,1)	1.00	0.69	1.89	0.93	1.59	1.59	1.50
	Mixture I	1.00	0.69	2.02	0.92	2.88	3.04	2.80
	Mixture II	1.00	1.21	1.77	1.22	7.24	12.77	11.96
	Cauchy on $[-10, 10]$	1.00	1.04	2.02	1.14	3.13	3.44	3.21
	t_2 on $[-10, 10]$	1.00	0.98	1.86	1.13	2.54	2.67	2.53
	Laplace	1.00	0.86	1.85	1.03	2.15	2.25	2.13
Model II	N(0,1)	1.00	0.71	1.00	0.95	0.94	0.92	0.87
	Mixture I	1.00	0.99	1.00	4.28	3.98	3.60	3.15
	Mixture II	1.00	3.89	1.00	4.33	4.30	4.18	3.91
	Cauchy on $[-10, 10]$	1.00	11.20	1.00	12.02	11.91	11.46	10.41
	t_2 on $[-10, 10]$	1.00	2.44	1.00	2.66	2.64	2.55	2.39
	Laplace	1.00	1.73	1.00	1.74	1.73	1.68	1.58

3.6.2 The Six Cities Study of Air Pollution and Health

The Six Cities Study of Air Pollution and Health was conducted to investigate pulmonary function (PF) in children born in 1967 or later from six cities across the United States; see Dockery et al. (1983). To measure PF, a basic spirometry procedure is to measure the forced expiratory volume (FEV) in one second. From

the first or second grade, spirometry was performed and other information, such as height and age, were collected roughly annually, until graduation from high school or loss to follow-up. See Section 8.8 in Fitzmaurice, Laird and Ware (2004) for detailed discussions.

Here we analyze a subset of the data from the study. The data, available at *http://www.biostat.harvard.edu/~fitzmaur/ala/fev1.txt*, consist of measurements of FEV, height, and age from 300 female participants in Topeka, Kansas. Each girl has different number of observations, ranging from 1 to 12, measured at different times. According to Gould (1966), PF and height (Height) has the approximate relationship $\text{PF} = \alpha \times \text{Height}^{\beta}$, where β is an allometric constant. Taking logarithm, we consider (FEV is a proxy of PF):

$$\log\{\text{FVE}(t)\} = \alpha(t) + \beta(t) \log\{\text{Height}(t)\} + \varepsilon(t), \tag{3.32}$$

where t represents the age. The above model allows the effect of the logarithm height to vary with age. In Wang et al. (1993), they fitted separate linear regression models for each year of ages $t = 6, 7, \ldots, 18$. By contrast, the framework (3.32) allows us to estimate the time-varying coefficient $\beta(\cdot)$ as a nonparametric function of age.

Examining the residuals from the median QR, we find one clear outlier of residuals: the measurement at age 9.1 years from subject 197 (-8.4 standard deviation). Figure 3.2 below plots the residuals' density estimator in (3.29), which clearly shows an asymmetric feature. Furthermore, the Kolmogorov-Smirnov test for normality gives a p-value 0.036. Therefore, we conclude that the residuals are non-normally distributed and have extreme outliers.

To reduce the effect of outlier on the bandwidth selection, we drop the observation in (3.27) to obtain the optimal cross-validation bandwidth $b_{\text{LS}} = 0.38$ for the local LS estimator. We then follow Sections 3.5.1–3.5.3 on page 37–40 to select the optimal bandwidth and estimate the optimal weights for the prewhitened OWLQAE (POWLQAE) estimator. By the nature of annual sparse longitudinal measurements as well as the empirical evidence in the simulation studies in Section 3.6.1 above, we use $p = 1$ in (3.24). The thick solid curve in Figure 3.3 below is the POWLQAE of $\beta(\cdot)$. The plot clearly shows that the allometric constant $\beta(\cdot)$

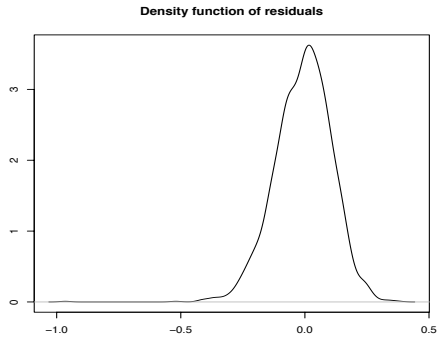

Figure 3.2. Plot of the residuals' nonparametric kernel density estimator in (3.29).

in (3.32) varies with age. From age 8 years, $\beta(\cdot)$ decreases with a local trough 2.2 at about age 9.5 years, which is then followed by a steady increase until the peak about 2.9 at age 13 years. After age 13 years, $\beta(\cdot)$ gradually decreases to about 1.8 at age 17 years. This complicated non-linear pattern can hardly be captured by simple parametric forms (e.g., linear or quadratic trends).

For comparison, we also plot the local LS estimator of $\beta(\cdot)$ in Figure 3.3 (the thin solid curve). Overall, the LS estimator exhibits similar pattern as the POWLQAE. However, there is clear discrepancy between the two estimators on $[8, 10]$, and in particular, the local LS estimator fails to capture the "V" shape as shown in the POWLQAE. As mentioned above, subject 197 has an unusually low FEV at age 9.1 and height 1.22 meters. Since the height is well below the average, the unusually low response will cause an increase in the local LS estimator of the coefficient $\beta(\cdot)$, which explains the relatively larger local LS estimator on the region $[8, 10]$ compared to the more robust QR based POWLQAE.

To appreciate the impact of the outlier, we consider perturbing the data as follows:

(i) Scenario I: remove the outlier. Based on the remaining observations, the corresponding POWLQAE and local LS estimators of $\beta(\cdot)$ are plotted in Figure 3.3 by thick dashed curve and thin dashed curve, respectively.

(ii) Scenario II: make the outlier even more extreme by further shifting its values 0.3 units down. See the thick dotted curve and thin dotted curve for the POWLQAE and local LS estimator, respectively.

As shown in Figure 3.3, compared with the estimator based on the original data, the local LS estimators under the two perturbation scenarios differ significantly on the region $[8, 10]$. By contrast, the POWLQAEs under the three cases almost remain the same, indicating the robustness in the presence of outliers. As argued above, the residuals are non-normally distributed and have outliers, therefore the proposed POWLQAE offers an attractive robust and efficient alternative over the local LS estimator.

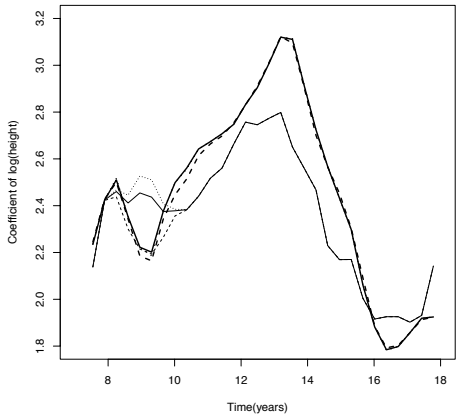

Figure 3.3. Estimates of $\beta(\cdot)$ in (3.32) under different settings. Thin solid, dashed, and dotted curves are the local LS estimators based on the original data, perturbation scenario I (remove the outlier), and perturbation scenario II (shift the outlier down by 0.3 units), respectively. Similarly, thick solid, dashed, and dotted curves are the POWLQAEs based on the original, perturbation scenario I and II data.

3.7 Proofs

Write $x_n \asymp y_n$ if $x_n/y_n \to 1$, $x_n = O(y_n)$ if $\sup_n |x_n/y_n| < \infty$, and $x_n = o(y_n)$ if $x_n/y_n \to 0$.

3.7.1 Proof of Theorem 4

Proof. Write $u_{ij} = (t_{ij} - t)/b$, $K_{ij} = K(u_{ij})$, $X_{ij} = X_i(t_{ij})$,

$$\Delta = \begin{bmatrix} \Delta_\beta \\ \Delta_\beta^* \\ \Delta_\alpha \\ \Delta_\alpha^* \end{bmatrix} = \sqrt{Nb} \begin{bmatrix} \beta - \beta(t) \\ b\{\beta^* - \beta'(t)\} \\ \alpha - \alpha_\tau(t) \\ b\{\alpha^* - \alpha'_\tau(t)\} \end{bmatrix} \quad \text{and} \quad Z_{ij} = \begin{bmatrix} X_{ij} \\ X_{ij}u_{ij} \\ 1 \\ u_{ij} \end{bmatrix}.$$

Also, let $\xi_{ij} = \varepsilon_{ij} - Q_\varepsilon(\tau)$ and recall $\alpha_\tau(t) = \alpha(t) + Q_\varepsilon(\tau)$ in (3.5). Then we can write

$$Y_{ij} - \alpha - \alpha^*(t_{ij} - t) - X_i(t_{ij})^T[\beta + \beta^*(t_{ij} - t)] = d_{ij} + \xi_{ij} - Z_{ij}^T\Delta/\sqrt{Nb}, \quad (3.33)$$

where $d_{ij} = [\alpha_\tau(t_{ij}) - \alpha_\tau(t) - \alpha'_\tau(t)(t_{ij} - t)] + X_{ij}^T[\beta(t_{ij}) - \beta(t) - \beta'(t)(t_{ij} - t)]$. By (3.33), since $[\hat\beta(t|\tau)^T, \hat\beta^*(t|\tau)^T, \hat\alpha(t|\tau), \hat\alpha^*(t|\tau)]^T$ minimizes (3.6), the re-scaled vector $\hat\Delta = \sqrt{Nb}[\{\hat\beta(t|\tau) - \beta(t)\}^T, b\{\hat\beta^*(t|\tau) - \beta'(t)\}^T, \hat\alpha(t|\tau) - \alpha_\tau(t), b\{\hat\alpha^*(t|\tau) - \alpha'_\tau(t)\}]^T$ minimizes the re-parameterized function of Δ:

$$\mathcal{L}(\Delta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \Big\{\rho_\tau\Big(d_{ij} + \xi_{ij} - Z_{ij}^T\Delta/\sqrt{Nb}\Big) - \rho_\tau\Big(d_{ij} + \xi_{ij}\Big)\Big\}K_{ij}. \quad (3.34)$$

Write $\delta_{ij} = Z_{ij}^T\Delta/\sqrt{Nb}$. Applying Knight's identity $\rho_\tau(u - \theta) - \rho_\tau(u) = -\theta(\tau - \mathbf{1}_{u<0}) + \int_0^\theta(\mathbf{1}_{u\le s} - \mathbf{1}_{u\le 0})ds$, we can write $\mathcal{L}(\Delta) = -A_n\Delta + I_n$, where

$$\begin{aligned} A_n &= \frac{1}{\sqrt{Nb}}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(\tau - \mathbf{1}_{d_{ij}+\xi_{ij}<0})K_{ij}Z_{ij}^T, \\ I_n &= \sum_{i=1}^{n}\sum_{j=1}^{m_i} K_{ij}\eta_{ij}, \quad \eta_{ij} = \int_0^{\delta_{ij}}(\mathbf{1}_{d_{ij}+\xi_{ij}\le s} - \mathbf{1}_{d_{ij}+\xi_{ij}\le 0})ds. \end{aligned}$$

Since $\{t_{ij}\}_{j=1,\dots,m_i}$ are ordered statistics of uniform random variables and the summations in A_n and I_n are taken over all $j = 1, \dots, m_i$, without loss of generality we can treat $\{t_{ij}\}_{j=1,\dots,m_i}$ as i.i.d. uniform random variables (instead of their ordered statistics).

Consider I_n. Since K has bounded support, it suffices to consider $|t_{ij} - t| = O(b)$. By the boundedness of X_{ij}, $|\delta_{ij}| \le c_1/\sqrt{Nb}$ and $|d_{ij}| \le c_1 b^2$ for some constant

c_1. Thus,

$$|\eta_{ij}| \le |\delta_{ij}|\mathbf{1}_{-|\delta_{ij}|\le\xi_{ij}+d_{ij}\le|\delta_{ij}|} \le \frac{c_1}{\sqrt{Nb}}\mathbf{1}_{-c_1\rho_n\le\xi_{ij}\le c_1\rho_n}, \quad \rho_n = \frac{1}{\sqrt{Nb}} + b^2.$$

Then we have $\mathbb{E}(K_{ij}^2\eta_{ij}^2) = O(\rho_n/N)$. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \mathrm{var}(I_n) = \sum_{i=1}^{n} \mathrm{var}\left(\sum_{j=1}^{m_i} K_{ij}\eta_{ij}\right) &\le \sum_{i=1}^{n}\left[m_i \sum_{j=1}^{m_i} \mathbb{E}(K_{ij}^2\eta_{ij}^2)\right] \\ &= O\left(\sum_{i=1}^{n} m_i^2\rho_n/N\right) \to 0, \end{aligned} \tag{3.35}$$

in view of Assumption (A4). By $d_{ij} = O(b^2)$ and simple Taylor's expansion,

$$\mathbb{E}(\eta_{ij}|X_{ij}, t_{ij}) = \int_0^{\delta_{ij}} \left[F_\varepsilon\Big\{Q_\varepsilon(\tau) + s - d_{ij}\Big\} - F_\varepsilon\Big\{Q_\varepsilon(\tau) - d_{ij}\Big\}\right] ds \asymp \delta_{ij}^2 \frac{f_\varepsilon(Q_\varepsilon(\tau))}{2}, \tag{3.36}$$

uniformly for all (i, j). Recall that $\mathbb{E}[X(t)] = 0$ and $\Gamma_X(t) = \mathbb{E}[X(t)X(t)^T]$. Note that

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} \mathbb{E}(K_{ij}\delta_{ij}^2) \to \frac{\Delta^T\Omega(t)\Delta}{T_u - T_l}, \quad \text{where} \quad \Omega(t) = \mathrm{diag}\{\Gamma_X(t), \mu_K\Gamma_X(t), 1, \mu_K\} \tag{3.37}$$

is a block diagonal matrix. By (3.35)–(3.37), we have the convergence in probability:

$$I_n = \mathbb{E}(I_n) + o_p(1) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \mathbb{E}[K_{ij}\mathbb{E}(\eta_{ij}|X_{ij}, t_{ij})] + o_p(1) \to \frac{f_\varepsilon(Q_\varepsilon(\tau))}{2(T_u - T_l)}\Delta^T\Omega(t)\Delta.$$

Recall $\hat{\Delta} = \mathrm{argmin}_\Delta\, \mathcal{L}(\Delta)$. By the quadratic approximation and convexity lemma,

$$\hat{\Delta} = \underset{\Delta}{\mathrm{argmin}}\Big\{ - A_n\Delta + \frac{f_\varepsilon(Q_\varepsilon(\tau))}{2(T_u - T_l)}\Delta^T\Omega(t)\Delta\Big\} + o_p(1) = \frac{T_u - T_l}{f_\varepsilon(Q_\varepsilon(\tau))}\Omega(t)^{-1}A_n^T + o_p(1).$$

For the $\hat{\beta}$ components of $\hat{\Delta}$, we have

$$\hat{\beta}(t|\tau) - \beta(t) = \frac{(T_u - T_l)\Gamma_X^{-1}(t)}{f_\varepsilon(Q_\varepsilon(\tau))} \frac{1}{Nb} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (\tau - \mathbf{1}_{\xi_{ij}<0} + \zeta_{ij}) K_{ij} X_{ij} + o_p[(Nb)^{-1/2}], \quad (3.38)$$

where $\zeta_{ij} = \mathbf{1}_{\xi_{ij}<0} - \mathbf{1}_{d_{ij}+\xi_{ij}<0}$. Let $R_n = (Nb)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \zeta_{ij} K_{ij} X_{ij}$. By the arguments in (3.35)–(3.36) and Taylor's expansion $d_{ij} = \sum_{s=2}^{3} b^s u_{ij}^s \{\partial^s \alpha_\tau(t)/\partial t + X_{ij}^T \partial^s \beta(t)/\partial t\}/s! + O(b^4)$,

$$\begin{aligned} \mathbb{E}(R_n) &= \frac{1}{Nb} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathbb{E}\{K_{ij} X_{ij} \mathbb{E}(\zeta_{ij}|X_{ij}, t_{ij})\} \\ &= \frac{f_\varepsilon(Q_\varepsilon(\tau))\mu_K}{2(T_u - T_l)} \Gamma_X(t) \beta''(t) b^2 + O(b^4), \end{aligned} \quad (3.39)$$

and $\mathrm{var}(R_n) = o[(Nb)^{-1/2}]$. Note that $b^4 = o[(Nb)^{-1/2}]$. Thus, by (3.38)–(3.39),

$$\sqrt{Nb}\Big[\hat{\beta}(t|\tau) - \beta(t) - \frac{\beta''(t)\mu_K}{2} b^2\Big] = \frac{(T_u - T_l)\Gamma_X^{-1}(t)}{f_\varepsilon(Q_\varepsilon(\tau))} \frac{1}{\sqrt{Nb}} \sum_{i=1}^{n} \varrho_i + o_p(1), \quad (3.40)$$

where $\varrho_i = \sum_{j=1}^{m_i} \varrho_{ij}$ with $\varrho_{ij} = [\tau - \mathbf{1}_{\varepsilon_{ij}<Q_\varepsilon(\tau)}] K_{ij} X_{ij}$. For $j \neq j'$, $\mathbb{E}(\varrho_{ij}\varrho_{ij'}^T) = O(b^2)$. Thus,

$$\begin{aligned} \mathrm{var}\Big(\frac{1}{\sqrt{Nb}} \sum_{i=1}^{n} \varrho_i\Big) &= \frac{1}{Nb} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathbb{E}\{[\tau - \mathbf{1}_{\varepsilon_{ij}<Q_\varepsilon(\tau)}]^2 K_{ij}^2 X_{ij} X_{ij}^T\} + \frac{1}{Nb} \sum_{i=1}^{n} O(m_i^2 b^2) \\ &\to \frac{\tau(1-\tau)\varphi_K \Gamma_X(t)}{T_u - T_l}. \end{aligned} \quad (3.41)$$

The desired result then easily follows from (3.40) and the independence of $\varrho_1, \ldots, \varrho_n$. ♢

3.7.2 Proof of Theorem 5

Proof. By (3.40), we have

$$\sqrt{Nb}\Big\{\hat{\beta}_{\mathrm{WLQAE}}(t|\omega) - \beta(t) - \frac{\beta''(t)\mu_K}{2} b^2\Big\} = \frac{(T_u - T_l)\Gamma_X^{-1}(t)}{\sqrt{Nb}} \sum_{i=1}^{n} \nu_i + o_p(1),$$

$\nu_i = \sum_{j=1}^{m_i} \sum_{r=1}^{k} \omega_r / f_\varepsilon(Q_\varepsilon(\tau_r))[\tau_r - \mathbf{1}_{\varepsilon_{ij} < Q_\varepsilon(\tau_r)}] K_{ij} X_{ij}$. The result then follows from the same argument in (3.41) and $\text{cov}\{\tau_r - \mathbf{1}_{\varepsilon_{ij} < Q_\varepsilon(\tau_r)}, \tau_s - \mathbf{1}_{\varepsilon_{ij} < Q_\varepsilon(\tau_s)}\} = \min(\tau_r, \tau_s) - \tau_r \tau_s$. ♢

3.7.3 Proof of Theorem 6

Proof. Recall H in (3.9) and e_k in (3.11). Define matrix Γ and vector q:

$$\Gamma = \Big[\min(\tau_r, \tau_s) - \tau_r \tau_s \Big]_{1 \le r,s \le k} \quad \text{and} \quad q = [f_\varepsilon(Q_\varepsilon(\tau_1)), \ldots, f_\varepsilon(Q_\varepsilon(\tau_k))]^T.$$

Also, define the diagonal matrix $P = \text{diag}\{f_\varepsilon(Q_\varepsilon(\tau_1)), \ldots, f_\varepsilon(Q_\varepsilon(\tau_k))\}$. Then $H = P^{-1}\Gamma P^{-1}$, $H^{-1} = P\Gamma^{-1}P$, and $Pe_k = q$. From (3.11), $\omega^* = P\Gamma^{-1}q/(q^T\Gamma^{-1}q)$ and $G(\omega^*) = 1/(q^T\Gamma^{-1}q)$.

(i) By direct matrix multiplications, we can verify that Γ^{-1} has $2(k+1)$ on the principal diagonal, $-(k+1)$ on the super- and sub-diagonals, and 0 elsewhere. Then we can easily show that $\omega^* = P\Gamma^{-1}q/(q^T\Gamma^{-1}q)$ has the explicit form in (3.13).

(ii) It suffices to prove $q^T\Gamma^{-1}q \to \mathcal{F}(f_\varepsilon)$. Recall $g(\tau) = f_\varepsilon(Q_\varepsilon(\tau))$ in (A6). Using the explicit expression of Γ^{-1} in (i) above, we can obtain

$$q^T\Gamma^{-1}q = (k+1)[g^2(\tau_1) + g^2(\tau_k)] + W_k + \int_\Delta^{1-\Delta} [g'(t)]^2 dt, \qquad \Delta = \frac{1}{k+1}, \quad (3.42)$$

where $W_k = (k+1)\sum_{r=2}^{k}[g(\tau_r) - g(\tau_{r-1})]^2 - \int_\Delta^{1-\Delta}[g'(t)]^2 dt$. Since $\tau_1 = \Delta, \tau_k = 1-\Delta$, and $\tau_r - \tau_{r-1} = \Delta$, we can rewrite W_k as

$$\begin{aligned} W_k &= (k+1)\sum_{r=2}^{k}\left\{ \left[\int_{\tau_{r-1}}^{\tau_r} g'(t)dt\right]^2 - (\tau_r - \tau_{r-1})\int_{\tau_{r-1}}^{\tau_r}[g'(t)]^2 dt \right\} \\ &= -\frac{k+1}{2}\sum_{r=2}^{k}\int_{\tau_{r-1}}^{\tau_r}\int_{\tau_{r-1}}^{\tau_r}[g'(t) - g'(s)]^2 dtds. \end{aligned} \quad (3.43)$$

For all $t, s \in [\tau_{r-1}, \tau_r]$, we have $|g'(t) - g'(s)| = |\int_s^t g''(v)dv| \le \int_{\tau_{r-1}}^{\tau_r}|g''(v)|dv$, uniformly. Thus, by the Cauchy-Schwarz inequality, $\max_{t,s\in[\tau_{r-1},\tau_r]}|g'(t) - g'(s)|^2 \le [\int_{\tau_{r-1}}^{\tau_r}|g''(v)|dv]^2 \le \Delta\int_{\tau_{r-1}}^{\tau_r}[g''(v)]^2 dv$. Applying the latter inequality to (3.43) and

under (A6), we have

$$|W_k| \leq \frac{(k+1)\Delta^2}{2} \sum_{r=2}^{k} \max_{t,s \in [\tau_{r-1}, \tau_r]} |g'(t) - g'(s)|^2 \leq \frac{\Delta^2}{2} \int_{\Delta}^{1-\Delta} |g''(t)|^2 dt \to 0.$$

Define $u = Q_\varepsilon(\tau)$. By $\partial\tau/\partial u = f_\varepsilon(u)$ and the chain rule $g'(\tau) = (\partial g/\partial u)(\partial u/\partial \tau) = f'_\varepsilon(u)/f_\varepsilon(u)$, we have $\int_0^1 [g'(\tau)]^2 d\tau = \mathcal{F}(f_\varepsilon)$. Thus, in (3.43), $\int_\Delta^{1-\Delta} [g'(t)]^2 dt \to \mathcal{F}(f_\varepsilon)$ as $\Delta \to 0$; also, by assumption (A6), $(k+1)[g^2(\tau_1) + g^2(\tau_k)] \to 0$, completing the proof. ♢

Proof of Proposition 1. Let τ and τ be two sets of quantiles. Clearly, if $\tau \subset \tau'$, then $G(\omega^*|\tau') \leq G(\omega^*|\tau)$ since we can let the quantiles $\tau^* \in \tau'$ and $\tau^* \notin \tau$ have weights zero. For rational quantiles $\tau_1, \ldots, \tau_k \in (0,1)$, there exist integers $1 \leq s_1, \ldots, s_r < k^* + 1$ such that $\tau_r = s_r/(k^*+1), r = 1, \ldots, k$. The result follows from $\{\tau_1, \ldots, \tau_k\} \subset \mathcal{U}_{k^*}$. ♢

3.7.4 Proof of Theorem 7

Proof. For the prewhitened model (3.20), the errors e_{ij} are i.i.d. Thus, the result follows from Theorem 6. ♢

3.7.5 Proof of Theorem 8

Proof. With a slight abuse of notation, write $\mathcal{F}(Z)$ as the Fisher information of the density of a random variable Z. Let Z_1 and Z_2 be independent random variables. By the Fisher information inequality [Equation (2.9) in Stam (1959)], $(\alpha_1+\alpha_2)^2 \mathcal{F}(Z_1+Z_2) \leq \alpha_1^2 \mathcal{F}(Z_1) + \alpha_2^2 \mathcal{F}(Z_2)$ for $\alpha_1, \alpha_2 > 0$. Letting $\alpha_1 = 1/\mathcal{F}(Z_1)$ and $\alpha_2 = 1/\mathcal{F}(Z_2)$ gives $1/\mathcal{F}(Z_1+Z_2) \geq 1/\mathcal{F}(Z_1) + 1/\mathcal{F}(Z_2)$. The desired result follows by letting $Z_1 = R(\varepsilon_{ij-1}, \ldots, \varepsilon_{ij-p})$ and $Z_2 = e_{ij}$, and using the imposed causal assumption. ♢

Chapter 4

Specification test for Markov models with measurement errors

4.1 Introduction

Let $\{X_i\}_{i\in\mathbb{N}}$ be a real-valued stationary time series of interest. In some applications, $\{X_i\}$ may not be directly observable and instead we observe a contaminated version $\{Y_i\}$:

$$Y_i = X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{4.1}$$

where $\{\varepsilon_i\}$ are independent and identically distributed (i.i.d.) measurement errors. For example, (4.1) has been proposed to explain the microstructure noise phenomenon observed in high-frequency financial data [Aït-Sahalia et al. (2005); Zhang et al. (2005)]. In the vast literature on errors-in-variables or measurement errors models, the central goal has been to study parameter estimation and inference of parametric regressions in the presence of measurement errors on the covariates; see the monographs Fuller (1987) and Carroll et al. (2006) and the recent survey paper Chen et al. (2011) for an extensive account of related contributions. Unlike the aforementioned works, our focus is on inferring the model dynamics of the unobservable process $\{X_i\}$.

The main purpose of this article is to address specification testing regarding the underlying data-generating mechanism, denoted by $\mathcal{Q}$, that generates $\{X_i\}$ based

on the contaminated observations $\{Y_i\}$. Specifically, we are interested in testing

$$H_0 : \mathcal{Q} = \mathcal{Q}_\theta, \text{ for a parametric specification } \mathcal{Q}_\theta \text{ with unknown parameter } \theta. \quad (4.2)$$

Parametric models can provide a parsimonious interpretation of the model dynamics, but a mis-specification of the underlying model may result in wrong conclusions. Therefore, it is necessary to validate the adequacy of the parametric model before employing it.

There is an extensive literature on specification testing but most existing works are concentrated on the case that data of interest are directly observable. Some representative works include pseudo-likelihood ratio test [Azzalini and Bowman (1993)], square distance between parametric and nonparametric estimate [Härdle and Mammen (1993)], residuals-based tests [Fan and Li (1996); Hong and White (1995)], generalized likelihood ratio test [Fan et al. (2001)], and density based approaches [Aït-Sahalia (1996); Gao and King (2004); Hong and Li (2005); Aït-Sahalia et al. (2009)]. In the above works, direct observations from the model of interest are available, a feature unfortunately not shared by (4.1).

Due to the unobservability and dependence of $\{X_i\}$, the aforementioned methods are not applicable and it is a difficult task to address specification testing regarding $\mathcal{Q}$. To alleviate this, we impose a Markovian assumption on $\{X_i\}$. Markov chains are used in a wide range of fields, ranging from quantitative fields such as econometrics and statistics to more applied fields such as biology and engineering. In econometrics, one important example is the nonlinear autoregressive conditional heteroscedastic model

$$X_i = \mu(X_{i-1}) + s(X_{i-1})\eta_i, \quad (4.3)$$

for i.i.d. errors $\{\eta_i\}_{i\in\mathbb{Z}}$. Given different specifications of (μ, s), (4.3) includes many popular models, such as threshold autoregressive models and autoregressive conditional heteroscedastic models. Another example is discrete samples from the diffusion model

$$dX_t = \mu(X_t)dt + s(X_t)dW_t, \quad t \geq 0, \quad (4.4)$$

where $\{W_t\}_{t\geq 0}$ is a standard Brownian motion. Special examples of (4.4) include the well-known Black-Scholes model, Vasicek model, Cox-Ingersoll-Ross model, and Chan-Karoly-Longstaff-Sanders model among others. See Zhao (2008) for a review.

In this article, we propose a conditional expectation generator based approach to address the specification testing problem (4.2). Our approach is motivated by three facts: (i) the evolving dynamics of the unobservable Markov chain $\{X_i\}$ is characterized by its transition density, denoted by $q_X(x'|x)$; (ii) the transition density $q_X(x'|x)$ of $\{X_i\}$ is implicitly coded into the conditional density, denoted by $q_Y(y'|y)$, of Y_i given Y_{i-1}; and (iii) furthermore, $q_Y(y'|y)$ is coded into the conditional expectation

$$\mathcal{G}_g(y) = \mathbb{E}[g(Y_i)|Y_{i-1} = y], \tag{4.5}$$

for proper transformations $g(\cdot)$. To address specification testing for hidden Markov models, Zhao (2011) compared the parametric estimate of $q_Y(y'|y)$ to its nonparametric estimate. Using $\mathcal{G}_g(y)$ instead of $q_Y(y'|y)$ has several practical advantages. In order to estimate the two-dimensional function $q_Y(y'|y)$ nonparametrically, a large sample size is required, and moreover it is a challenging issue to choose two bandwidths. By contrast, estimation of $\mathcal{G}_g(y)$ is a well-studied standard nonparametric regression problem.

The main component of our methodology is the construction of a nonparametric simultaneous confidence band (SCB) for $\mathcal{G}_g(y)$. The constructed nonparametric SCB does not depend on any specific model structure and hence can serve as a true reference. To test (4.2), we then check whether the parametric estimate of $\mathcal{G}_g(y)$ under H_0 is contained within the nonparametric SCB. The problem of SCB construction has been studied previously for marginal density of independent data [Bickel and Rosenblatt (1973)], nonparametric regression function for both independent data [Knafl et al. (1985); Eubank and Speckman (1993); Fan and Zhang (2000)] and time series data [Zhao and Wu (2008)]. For hidden Markov models, Zhao (2011) studied SCB for conditional density function. Our development on SCB for $\mathcal{G}_g(y)$ under the Markov-chain measurement-error model involves novel technical developments. The main argument is to decompose summation of depen-

dent variables into a leading summation of martingale differences and a negligible error term. Unlike the nonparametric kernel density estimation case where the summands are uniformly bounded, nonparametric kernel smoothing estimate of the regression function $\mathcal{G}_g(y)$ involves unbounded terms and is significantly more challenging to deal with.

Throughout, for a random variable Z, we write $Z \in \mathcal{L}^q, q > 0$, if $\|Z\|_q := [\mathbb{E}(|Z|^q)]^{1/q} < \infty$; for $z \in \mathbb{R}$, write $\lfloor z \rfloor$ as the integer part of z. Section 2 presents the main methodology. Section 3 contains simulation studies. Technical proofs are provided in Section 4.4.

4.2 Methodology

For the Markov chain $\{X_i\}$, its evolving dynamics is fully characterized by the joint density function, denoted by $p_X(x, x')$, of the pair (X_{i-1}, X_i). Since $\{X_i\}$ is unobservable, we propose extracting information about $p_X(x, x')$ from the observed chain $\{Y_i\}$. In (4.1), we assume that $\{\varepsilon_i\}_{i\in\mathbb{Z}}$ are i.i.d. and independent of the Markov chain $\{X_i\}_{i\in\mathbb{Z}}$.

To illustrate the idea, denote by $q_Y(y'|y)$ the conditional density function of Y_i given $Y_{i-1} = y$, by $p_Y(y, y')$ the joint density function of (Y_{i-1}, Y_i), and by $q_\varepsilon(\cdot)$ the density function of ε_i. Conditioning on (X_{i-1}, X_i), $Y_{i-1} \sim q_\varepsilon(y - X_{i-1})$ and $Y_i \sim q_\varepsilon(y' - X_i)$ are independent. Thus, $p_Y(y, y') = \mathbb{E}[q_\varepsilon(y - X_{i-1})q_\varepsilon(y' - X_i)]$. Similarly, conditioning on X_{i-1}, we obtain $f_Y(y) = \mathbb{E}[q_\varepsilon(y - X_{i-1})]$. Therefore,

$$q_Y(y'|y) = \frac{p_Y(y, y')}{f_Y(y)} = \frac{\mathbb{E}[q_\varepsilon(y - X_{i-1})q_\varepsilon(y' - X_i)]}{\mathbb{E}[q_\varepsilon(y - X_{i-1})]}. \tag{4.6}$$

The identity (4.6) shows how the density information $p_X(x, x')$ of the unobservable pair (X_{i-1}, X_i) is coded into the transition density $q_Y(y'|y)$ of the observed chain $\{Y_i\}$.

In practice, due to the two-dimensional feature, it is generally difficult to work with $q_Y(y'|y)$. Our proposed method is based on the conditional expectation operator $\mathcal{G}_g$ defined in (4.5). Different choices of $g(\cdot)$ can extract different information from the conditional density $q_Y(y'|y)$ and hence from $p_X(x, x')$ through the identity (4.6). For example, $g_1(Y_i) = Y_i$ and $g_2(Y_i) = Y_i^2$ extract information from the first

two conditional moments, and $g_t(Y_i) = \mathbf{1}_{Y_i \le t}, t \in \mathbb{R}$, extracts information from the conditional distribution; see Section 4.2.3 on page 60 for more discussions. The conditional expectation structure (4.5) can be nicely fitted into the nonparametric regression problem

$$Y_i^* = \mathcal{G}_g(Y_{i-1}) + \text{error}, \quad \text{where} \quad Y_i^* = g(Y_i). \tag{4.7}$$

Now we briefly introduce our proposed conditional expectation based approach to address the specification testing problem (4.2). First, we apply nonparametric kernel smoothing methods to (4.7) to construct a nonparametric estimate of $\mathcal{G}_g(y)$, denoted by $\hat{\mathcal{G}}_g(y)$. Without imposing any specific model structure, $\hat{\mathcal{G}}_g(y)$ is always a consistent estimate of $\mathcal{G}_g(y)$ and hence can be used as a reference quantity. Under H_0, we use the right hand side of (4.6) to construct a parametric estimate of $\mathcal{G}_g(y)$, denoted by $\mathcal{G}_g(y|\mathcal{Q}_{\hat{\theta}})$, where $\hat{\theta}$ is a consistent estimate of θ. To test H_0, we examine the distance between the parametric estiamte $\mathcal{G}_g(y|\mathcal{Q}_{\hat{\theta}})$ and the nonparametric reference $\hat{\mathcal{G}}_g(y)$, with a large discrepancy indicating rejection of H_0.

To determine the critical value, we use the idea of simultaneous confidence band (SCB). For a significance level $\alpha \in (0,1)$, we say that $[l_n(\cdot), u_n(\cdot)]$ is an asymptotic $(1-\alpha)$ nonparametric SCB for $\mathcal{G}(y)$ on a given compact set $\mathcal{Y} \subset \mathbb{R}$ if

$$\lim_{n\to\infty} \mathbb{P}\{l_n(y) \le \mathcal{G}_g(y) \le u_n(y), \quad \text{for all } y \in \mathcal{Y}\} = 1-\alpha. \tag{4.8}$$

Intuitively, the function $\mathcal{G}_g(\cdot)$ is contained within the nonparametric band $[l_n(\cdot), u_n(\cdot)]$ with asymptotic probability $(1-\alpha)$. As will be illustrated in Section 4.2.1, nonparametric SCB of $\mathcal{G}_g(\cdot)$ usually centers at a nonparametric estimate $\hat{\mathcal{G}}_g(y)$. Therefore, the band $[l_n(\cdot), u_n(\cdot)]$ with center $\hat{\mathcal{G}}_g(y)$ provides an acceptance region for H_0. If the parametric estimate $\mathcal{G}_g(y|\mathcal{Q}_{\hat{\theta}})$ under H_0 falls outside the band, then the deviation between $\hat{\mathcal{G}}_g(y)$ and $\mathcal{G}_g(y|\mathcal{Q}_{\hat{\theta}})$ is too large to be in favor of H_0. Clearly, the concept of SCB is an extension of the classical confidence interval for a one-dimensional parameter (e.g., the population mean) to a function.

We now summarize our nonparametric SCB based specification testing procedure:

(i) Apply nonparametric methods to the nonparametric regression problem (4.7)

to construct a nonparametric estimate $\hat{\mathcal{G}}_g(y)$ of $\mathcal{G}_g(y)$, and then use $\hat{\mathcal{G}}_g(y)$ to build a $(1-\alpha)$ nonparametric SCB for $\mathcal{G}_g(y)$, denoted by $[\ell_n(\cdot), u_n(\cdot)]$.

(ii) Under H_0, apply parametric methods to obtain an estimate $\hat{\theta}$ of θ, and further obtain a parametric estimate $\mathcal{G}_g(y|\mathcal{Q}_{\hat{\theta}})$ of $\mathcal{G}_g(y)$; see the right hand side of (4.6).

(iii) Check whether $l_n(y) \leq \mathcal{G}_g(y|\mathcal{Q}_{\hat{\theta}}) \leq u_n(y)$ holds for all $y \in \mathcal{Y}$, or equivalently, whether $\mathcal{G}_g(y|\mathcal{Q}_{\hat{\theta}})$ is contained within the constructed SCB. If no, we reject H_0 at level α.

In Sections 4.2.1 and 4.2.2, we construct nonparametric SCB and parametric estimate of $\mathcal{G}_g(y)$, respectively; Section 4.2.3 discusses choices of $g(\cdot)$ and the Bonferroni correction.

4.2.1 Nonparametric simultaneous confidence band

In the nonparametric regression problem (4.7), consider the Nadaraya-Watson kernel smoothing estimate of $\mathcal{G}_g(y)$:

$$\hat{\mathcal{G}}_g(y) = \frac{\sum_{i=1}^n g(Y_i) K_{b_n}(y - Y_{i-1})}{\sum_{i=1}^n K_{b_n}(y - Y_{i-1})}, \tag{4.9}$$

where and hereafter $K_{b_n}(u) = K(u/b_n)$ for a kernel function K satisfying $\int_{\mathbb{R}} K(u)du = 1$ and bandwidth $b_n > 0$. To study asymptotic properties of $\hat{\mathcal{G}}_g(y)$, we define the conditional variance function $\sigma_g^2(y)$ and impose Conditions 1–3 as follows:

$$\sigma_g^2(y) = \mathbb{E}\{[g(Y_i) - \mathcal{G}_g(Y_{i-1})]^2 | Y_{i-1} = y\}. \tag{4.10}$$

Condition 1 (Kernel assumption)**.** The kernel K is bounded, symmetric, and has bounded derivative and support $[-\omega, \omega]$. Write $\varphi_K = \int_{-\omega}^{\omega} K^2(u)du$ and $\psi_K = \int_{-\omega}^{\omega} u^2 K(u)du$.

Condition 2 (Regularity assumption)**.** Without loss of generality, let $\mathcal{Y} = [-T, T]$ for some $T > 0$. There exists some small $\epsilon > 0$ such that $f_Y(y) > 0$ and $\mathcal{G}_g(y)$ have bounded fourth order derivative on $\mathcal{Y}_\epsilon := [-T - \epsilon, T + \epsilon]$, and that $\sigma_g^2(y) > 0$

has bounded derivative on $\mathcal{Y}_\epsilon$. The density function q_ε of ε_i is bounded and has bounded derivative on $\mathbb{R}$.

Condition 3 (Dependence assumption)**.** The unobservable process $\{X_i\}$ is an α-mixing stationary Markov chain with α-mixing coefficients $\alpha_k, k \in \mathbb{N}$. Assume that $g(Y_i) \in \mathcal{L}^\delta$ for some $\delta \geq 4$ and $\sum_{k=1}^{\infty} \alpha_k^{1-2/\delta} < \infty$.

Condition 1 is a typical assumption about kernel function in nonparametric inference problems. Some smoothness assumptions are imposed in Condition 2. Condition 3 is frequently used to control dependence structures in time series. It is reasonably weak and is fulfilled for many time series models.

Definition 1. *Let $\tau_n \to 0$ and $m_n \to \infty$. We say that $\mathcal{Y}_n \subset \mathcal{Y}$ is a (τ_n, m_n) approximation of $\mathcal{Y}$ if: (i) $\mathcal{Y}_n$ contains m_n distinct points from $\mathcal{Y}$; (ii) the distance between any two points from $\mathcal{Y}_n$ is at least τ_n; and (iii) the distance between $\mathcal{Y}_n$ and $\mathcal{Y}$ goes to zero as $n \to \infty$.*

Theorem 9 below establishes a maximal deviation result for $\hat{\mathcal{G}}_g(y)$, which can be used to construct a nonparametric SCB for $\hat{\mathcal{G}}_g(y)$.

Theorem 9. *Assume that Conditions 1–3 hold and $nb_n^9 \log n + (nb_n^3)^{-1} \log n \to 0$. Then for any (τ_n, m_n) approximation $\mathcal{Y}_n$ of $\mathcal{Y}$ such that $b_n = o(\tau_n)$ and $(\log n)^3[(nb_n)^{-1} + b_n^2]m_n^2 \to 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{ \sup_{y\in\mathcal{Y}_n} \left[\frac{nb_n f_Y(y)}{\varphi_K \sigma_g^2(y)}\right]^{1/2} \left|\hat{\mathcal{G}}_g(y) - \mathcal{G}_g(y) - \rho_g(y) b_n^2\right| \leq B_{m_n}(z) \right\} = e^{-2e^{-z}} \quad (4.11)$$

for $z \in \mathbb{R}$, where $\sigma_g^2(y)$ is defined in (4.10), $\rho_g(y) = [f_Y'(y)\mathcal{G}_g'(y)/f_Y(y) + \mathcal{G}_g''(y)/2]\psi_K$, and

$$B_{m_n}(z) = \sqrt{2\log m_n} - \frac{1}{\sqrt{2\log m_n}}\left[\frac{1}{2}\log\log m_n + \log(2\sqrt{\pi}) - z\right].$$

In (4.11), the limiting distribution $\exp[-2\exp(-z)], z \in \mathbb{R}$, is the well-known Gumbel or type I extreme value distribution. Due to the unknown derivatives $f_Y'(y)$, $\mathcal{G}_g'(y)$ and $\mathcal{G}_g''(y)$, it is generally difficult to estimate the bias $\rho_g(y)b_n^2$ in

Theorem 9. To address this issue, we adopt a bias-correction procedure so that it is not necessary to estimate the second-order bias; see Section 4.3 on page 62 for more details. Using this bias-correction, we can use Theorem 9 to construct an asymptotic $(1-\alpha)$ SCB for $\mathcal{G}_g(y)$ on the region $\mathcal{Y}_n$:

$$\hat{\mathcal{G}}_g(y) \pm \left[\frac{\varphi_K \hat{\sigma}_g^2(y)}{n b_n \hat{f}_Y(y)}\right]^{1/2} B_{m_n}(z_\alpha), \quad y \in \mathcal{Y}_n, \tag{4.12}$$

where $z_\alpha = -\log\log[(1-\alpha)^{-1/2}]$ is the $(1-\alpha)$ quantile of the limiting distribution in (4.11), $\hat{\sigma}_g^2(y)$ and $\hat{f}_Y(y)$ are estimates of $\sigma_g^2(y)$ and $f_Y(y)$, respectively. By Definition 1, $\mathcal{Y}_n$ becomes denser and denser in $\mathcal{Y}$ as $n \to \infty$. Thus, the constructed SCB on $\mathcal{Y}_n$ provides a good approximation to (4.8) for sufficiently large n.

For any fixed $c > 0$, let $m_n = \lfloor 2T/[c(\log n)^2 b_n] \rfloor$ and $\mathcal{Y}_n = \{y_j = -T + c(\log n)^2 b_n j, j = 0, 1, \ldots, m_n - 1\}$. Then $\mathcal{Y}_n$ is a (τ_n, m_n) approximation of $\mathcal{Y}$ with $\tau_n = c(\log n)^2 b_n$. It is easily seen that, under $n b_n^9 \log n + (n b^3)^{-1} \log n \to 0$, the conditions $b_n = o(\tau_n)$ and $(\log n)^3[(n b_n)^{-1} + b_n^2] m_n^2 \to 0$ in Theorem 9 automatically hold.

Now we discuss estimation of $\sigma_g^2(y)$ and $f_Y(y)$. To estimate $f_Y(y)$, we use the nonparametric kernel density estimator:

$$\hat{f}_Y(y) = \frac{1}{n l_n} \sum_{i=1}^{n} K_{l_n}(y - Y_i), \tag{4.13}$$

for a bandwidth $l_n > 0$. Based on residuals $g(Y_i) - \hat{\mathcal{G}}_g(Y_{i-1})$, we propose the Nadaraya-Watson kernel smoothing estimate of $\sigma_g^2(y)$ in (4.10):

$$\hat{\sigma}_g^2(y) = \frac{\sum_{i=1}^{n} [g(Y_i) - \hat{\mathcal{G}}_g(Y_{i-1})]^2 K_{h_n}(y - Y_{i-1})}{\sum_{i=1}^{n} K_{h_n}(y - Y_{i-1})}, \tag{4.14}$$

where $K_{h_n}(u) = K(u/h_n)$ for another bandwidth $h_n > 0$.

Theorem 10. *(i) Under Conditions 1–3 and* $l_n^4 \log n + (n l_n)^{-1} (\log n)^2 \to 0$*, we have*

$$\sup_{y \in \mathcal{Y}} |\hat{f}_Y(y) - f_Y(y)| = o_p[(\log n)^{-1/2}].$$

(ii) Assume that Conditions 1–3 hold. Further assume $nb_n^8(\log n)^2+(nb_n^4)^{-1}(\log n)^6 \to 0$ *and* $h_n^4 \log n + (nh_n^2)^{-1}(\log n)^3 \to 0$. *Then we have*

$$\sup_{y\in\mathcal{Y}} |\hat{\sigma}_g^2(y) - \sigma_g^2(y)| = o_p[(\log n)^{-1/2}].$$

By Theorem 10 and Slutsky's theorem, the convergence in Theorem 9 still holds if we replace $\sigma_g^2(y)$ and $f_Y(y)$ by their estimates $\hat{\sigma}_g^2(y)$ and $\hat{f}_Y(y)$. We point out that, from the proof of Theorem 10, the bound $o_p[(\log n)^{-1/2}]$ can be substantially improved. For brevity, we present the loose bound $o_p[(\log n)^{-1/2}]$ since it is enough for our asymptotic results.

4.2.2 Parametric estimate under $H_0 : \mathcal{Q} = \mathcal{Q}_\theta$

In this section we develop a general procedure to construct parametric estimate of $\mathcal{G}_g(y)$ under $H_0 : \mathcal{Q} = \mathcal{Q}_\theta$. Recall that q_ε is the density function of ε_i. Theorem 11 below presents a theoretical representation for $\mathcal{G}_g(y)$.

Theorem 11. *Let* ε *be a random sample from* $q_\varepsilon(\cdot)$ *independent of* (X_{i-1}, X_i). *Then*

$$\mathcal{G}_g(y) = \frac{\mathbb{E}[g(X_i+\varepsilon)q_\varepsilon(y-X_{i-1})]}{\mathbb{E}[q_\varepsilon(y-X_{i-1})]}. \tag{4.15}$$

Without further assumptions, it is generally impossible to use (4.15) to construct a parametric estimate of $\mathcal{G}_g(y)$. For example, Fan (1991) assumed that ε_i has an exactly known density function in order to study the nonparametric deconvolution problem of estimating the density of X_i based on noisy observations Y_i from (4.1). Here we assume that ε_i has the normal distribution $N(0, \sigma^2)$ for some unknown variance $\sigma^2 > 0$. Denote by $\phi(z)$ the standard normal density, and write $\phi_\sigma(z) = \sigma^{-1}\phi(z/\sigma)$. Then $q_\varepsilon(z) = \phi_\sigma(z)$.

In practice, there is generally no closed-form expression for the joint density of (X_{i-1}, X_i) for most linear and nonlinear time series models, and thus it is infeasible to evaluate the expectations on the right hand side of (4.15) directly. For example, even for the simplest threshold autoregressive model $X_i = a_1 X_{i-1}\mathbf{1}_{X_{i-1}\le 0} + a_2 X_{i-1}\mathbf{1}_{X_{i-1}>0} + \eta_i$ with i.i.d. errors $\eta_i \sim N(0,1)$, the stationary

joint density remains unknown. To solve this issue, we propose a Markov Chain Monte Carlo (MCMC) simulation based method below.

(i) Under $H_0 : \mathcal{Q} = \mathcal{Q}_\theta$, obtain consistent estimate of (θ, σ), denoted by $(\hat{\theta}, \hat{\sigma})$. Under the parametric specification, a natural parameter estimation method is the maximum likelihood estimator, which may be computationally expensive. In some cases, it is computationally appealing to use, for example, moments based methods.

(ii) Simulate sample path $\{X_i^*\}_{0 \le i \le m}$ from the estimated null model $\mathcal{Q}_{\hat{\theta}}$.

(iii) Let ε_i be i.i.d. $N(0, \hat{\sigma}^2)$ variables. Using empirical version of (4.15), we propose

$$\hat{\mathcal{G}}_g(y|\mathcal{Q}_{\hat{\theta}}) = \frac{m^{-1}\sum_{i=1}^m g(X_i^* + \varepsilon_i)\phi_{\hat{\sigma}}(y - X_{i-1}^*)}{m^{-1}\sum_{i=1}^m \phi_{\hat{\sigma}}(y - X_{i-1}^*)}.$$

For large m, the numerator and denominator of $\hat{\mathcal{G}}(y|\mathcal{Q}_{\hat{\theta}})$ approach their expectations.

As an illustration, we consider (4.3) with $\eta_i \sim N(0,1)$. Let $\mathcal{Q} = (\mu, s)$ and $\mathcal{Q}_\theta = (\mu_\theta, s_\theta)$ for some parametric specification (μ_θ, s_θ). Then step (ii) above is implemented using

$$X_i^* = \mu_{\hat{\theta}}(X_{i-1}^*) + s_{\hat{\theta}}(X_{i-1}^*)\eta_i, \quad \eta_i \sim N(0,1).$$

Clearly, the above proposed procedure can be readily extended to the case of non-Gaussian errors. We simply replace $\phi_{\hat{\sigma}}$ with another given parametric density with estimated parameters and draw ε_i from the latter density.

4.2.3 Choices of the transformation $g(\cdot)$ and Bonferroni correction

In (4.5), different choices of $g(\cdot)$ can extract different information about the underlying distribution. In many practical problems, the conditional mean and conditional variance are the two most important pieces of information researchers are interested in. For example, in model (4.3) with i.i.d. $\eta_i \sim N(0,1)$, the conditional

mean function $\mu(\cdot)$ and the conditional variance function $s^2(\cdot)$ fully determine the model structure. Similarly, in model (4.4), $\mu(\cdot)$ and $s^2(\cdot)$ represent the conditional drift (mean) function and conditional volatility (variance) function, respectively, and they fully characterize the underlying model. To study the underlying model dynamics in (4.3) and (4.4), we let $\mathcal{Q} = (\mu, s)$ and $H_0 : \mathcal{Q} = \mathcal{Q}_\theta = (\mu_\theta, s_\theta)$ for parametric specifications $(\mu_\theta, \sigma_\theta)$.

Motivated by the above discussion, we propose using two simple transformations $g_1(Y_i) = Y_i$ and $g_2(Y_i) = Y_i^2$. By combining the two transformations together, the test can detect deviations from the conditional mean and/or conditional variance in the underlying model. To combine the two corresponding tests together, we adopt the following procedure:

(Bonferroni correction):

Suppose the pre-specified significance level is α, then we construct $(1 - \alpha/2)$ SCBs, denoted by SCB_1 *and* SCB_2*, for $\mathcal{G}_{g_1}(y)$ and $\mathcal{G}_{g_2}(y)$ separately, and reject H_0 if either* SCB_1 *or* SCB_2 *cannot cover the corresponding parametric estimates for $\mathcal{G}_{g_1}(y)$ or $\mathcal{G}_{g_2}(y)$.*

Theoretically speaking, we can combine tests across multiple transformations. For example, one natural choice is to combine multiple conditional moments, i.e., $g_k(Y_i) = Y_i^k, k = 1, \ldots, J$, for some $J \in \mathbb{N}$. However, we do not recommend this approach based on three considerations. First, the Bonferroni correction is well-known to be very conservative for multiple tests. Second, as discussed above, for the two most popular Markov models (4.3) and (4.4), the conditional mean and the conditional variance contain all important information and higher-order moments do not provide extra information. Third, using high-order moments would require high-order finite-moment assumptions, which may be too restrictive in practice.

4.2.4 Alternative approaches: conditional distribution or conditional characteristic function

Some alternative approaches are to use a class of transformations $g(\cdot)$ indexed by a continuous parameter. For example, $g_t(Y_i) = \mathbf{1}_{Y_i \leq t}$ for $t \in \mathbb{R}$ corresponds to the conditional distribution function, and $g_t(Y_i) = \exp(\sqrt{-1} Y_i t)$ for $t \in \mathbb{R}$ corresponds

to the conditional characteristic function. Under different contexts, Hong (1999) and Pinkse (1998) used empirical characteristic functions to test for serial dependence. In our SCB setting, using such choices of transformations involves studying maximum deviations of $\hat{\mathcal{G}}_{g_t}(y)$ over both $t \in \mathbb{R}$ and $y \in \mathbb{R}$. With $g_t(Y_i) = \mathbf{1}_{Y_i \leq t}$, we expect that, after proper normalization, the process $\{\hat{\mathcal{G}}_{g_t}(y)\}_{t\in\mathbb{R}}$ with any fixed y converges in distribution in the Skorokhod space $D[-\infty, +\infty]$ to the process $\{\mathbb{B}(Q_Y(t|y))\}_{t\in\mathbb{R}}$, where $\mathbb{B}$ is the standard Brownian bridge and $Q_Y(t|y)$ is the conditional distribution function of Y_i given $Y_{i-1} = y$. Therefore, by the continuous mapping theorem, we can handle the supremum over $t \in \mathbb{R}$. Unfortunately, it is unclear how to deal with the supremum over $y \in \mathbb{R}$. Furthermore, in order to establish the latter functional convergence, we need to prove the tightness of the process. It seems that substantial theoretical developments are necessary and we leave them for future research.

4.3 Monte Carlo simulation study

In this section, we conduct a small simulation study to examine the empirical performance of the proposed specification test. First, we address some practical implementation issues.

(Bias-correction): We adopt a higher-order kernel to remove the second-order bias term $\rho_g(y)b_n^2$ in Theorem 9 on page 57. Let $\phi(u)$ be the standard normal density function. In our numerical analysis, we use the kernel function $K(u) = 2\phi(u) - \phi(u/\sqrt{2})/\sqrt{2}$, which is symmetric and satisfies $\int_{\mathbb{R}} K(u)du = 1$ and $\psi_K = \int_{\mathbb{R}} u^2 K(u)du = 0$ so that $\rho_g(y) = 0$.

(Bandwidth selection): To select the bandwidth b_n in (4.9), we apply the plug-in method in Ruppert et al. (1995) to the nonparametric regression (4.7), which is implemented using the command `dpill` in the software `R`. Similarly, we choose h_n in (4.14) based on the nonparametric regression of $[g(Y_i) - \hat{\mathcal{G}}_g(Y_{i-1})]^2$ on Y_{i-1}. For l_n in (4.13), we use the rule-of-thumb nonparametric kernel density bandwidth selector in Silverman (1986), which is implemented using the command `bw.nrd0` in `R`.

We compare the empirical performance of the proposed specification tests based on different transformations $g(\cdot)$:

(i) Test 1: using the single transformation $g_1(Y_i) = Y_i$;

(ii) Test 2: using the single transformation $g_2(Y_i) = Y_i^2$;

(iii) Test 3: combining the two transformations g_1 and g_2 with the Bonferroni correction.

In (4.12), we need to select a set $\mathcal{Y}_n$ of grid points. For a realization $\{Y_i\}$, let $l_{0.15}$ and $l_{0.85}$ be their 15 and 85 percentiles, respectively. We take $\mathcal{Y}_n$ to be the set of 11 evenly spaced grid points $y_i = l_{0.15} + i(l_{0.85} - l_{0.15})/10$, $i = 0 \ldots, 10$.

In (4.1), we generate $\{X_i\}$ from the following true models:

$$\begin{aligned} &\text{(Model 1)} \quad X_i = 0.6[(1-\lambda)X_{i-1} + \lambda|X_{i-1}|] + \eta_i, \quad \lambda \in [0,1], \\ &\text{(Model 2)} \quad X_i = 0.6X_{i-1} + \eta_i\sqrt{1 + 0.3\lambda X_{i-1}^2}, \quad \lambda \in [0,1], \end{aligned}$$

for i.i.d. noises $\eta_i \sim N(0,1)$. We wish to test the null hypothesis $H_0 : X_i = \theta_1 X_{i-1} + \eta_i$ based on contaminated observations $\{Y_i\}$ from (4.1) with i.i.d. measurement errors $\varepsilon_i \sim N(0,1)$. The parameter λ regulates the deviation from the null model. The case $\lambda = 0$ leads to the null model; as λ increases, Model 1 and Model 2 move further away from the null model. In particular, for $\lambda \neq 0$, Model 1 becomes the threshold autoregressive model, and Model 2 becomes the autoregressive conditional heteroscedastic model with a linear term. Model 1 and Model 2 are used to examine the sensitivity of the test to deviations in the conditional mean function and conditional variance function, respectively.

Under H_0, we need to estimate θ_1 and the variances of η_i and ε_i, denoted by θ_2^2 and θ_3^2, respectively. Then elementary calculations show that

$$\mathbb{E}(Y_i^2) = \frac{\theta_2^2}{1-\theta_1^2} + \theta_3^2, \quad \operatorname{cov}(Y_{i-1}, Y_i) = \frac{\theta_1\theta_2^2}{1-\theta_1^2}, \quad \operatorname{cov}(Y_{i-2}, Y_i) = \frac{\theta_1^2\theta_2^2}{1-\theta_1^2}.$$

Thus, we can estimate the parameters $(\theta_1, \theta_2, \theta_3)$ by the empirical versions of moments. We simulate 1000 realizations with sample size $n = 2000$ and significance level $\alpha = 0.05$.

The result is presented in Table 4.1. We see that, Test 1 based on the transformation g_1 is much more powerful in detecting deviations in the conditional mean (Model 1) than in detecting deviations in the conditional variance (Model 2). Similarly, Test 2 based on g_2 is more powerful in detecting deviations in the conditional variance. By contrast, Test 3 based on combining g_1 and g_2 through the Bonferroni correction can detect both deviations well. Moreover, the empirical size of Test 3 is quite close to the nominal size 0.05, and the power rises dramatically as the deviation parameter λ increases. This small simulation study demonstrates that the proposed test using the two transformations g_1 and g_2 with the Bonferroni correction works quite well.

Table 4.1. Empirical power: Test 1, Test 2, and Test 3 stand for the proposed specification tests based on SCB with $g_1(Y_i) = Y_i$, $g_2(Y_i) = Y_i^2$, and combining the two transformations together with the Bonferroni correction, respectively. Nominal size is 0.05.

	$\lambda =$	0.0	0.2	0.4	0.6	0.8	1.0
Model 1	Test 1	0.063	0.637	0.989	1.000	1.000	1.000
	Test 2	0.018	0.086	0.255	0.358	0.428	0.532
	Test 3	0.042	0.503	0.974	1.000	1.000	1.000
Model 2	Test 1	0.063	0.085	0.110	0.152	0.235	0.321
	Test 2	0.018	0.049	0.170	0.447	0.784	0.946
	Test 3	0.042	0.069	0.147	0.367	0.716	0.917

4.4 Proofs

4.4.1 Some preliminary facts of projection operator

For convenience, we recall some basic properties of conditional expectations. Let $Z \in \mathcal{L}^1$ be any integrable random variable and $\mathcal{F}$ a σ-algebra on the same probability space. Then

(C1) $\mathbb{E}(Z) = \mathbb{E}[\mathbb{E}(Z|\mathcal{F})]$.

(C2) Let $\mathcal{G}$ be another σ-algebra such that $\mathcal{F} \subset \mathcal{G}$. Then $\mathbb{E}(Z|\mathcal{F}) = \mathbb{E}[\mathbb{E}(Z|\mathcal{G})|\mathcal{F}]$.

(C3) If $Z \in \mathcal{L}^p$ for some $p \geq 1$, then $\|\mathbb{E}(|Z||\mathcal{F})\|_p \leq \|Z\|_p$ and $(\mathbb{E}|Z|)^p \leq \mathbb{E}(|Z|^p)$.

Recall that, in (4.1), $\{\varepsilon_i\}$ are i.i.d. and independent of the unobservable Markov chain $\{X_i\}$. Let $\mathcal{F}_i = \sigma(\varepsilon_j, X_{j+1} : j \leq i)$ be the σ-algebra generated by $\varepsilon_j, X_{j+1}, j \leq i$. Then $\{\mathcal{F}_i\}_{i\in\mathbb{Z}}$ is an increasing filtration. For $i \in \mathbb{Z}$, define the projection operator $\mathcal{P}_i$ by

$$\mathcal{P}_i Z = \mathbb{E}(Z|\mathcal{F}_i) - \mathbb{E}(Z|\mathcal{F}_{i-1}), \quad Z \in \mathcal{L}^1.$$

The projection operator $\mathcal{P}_i$ satisfies the following properties (In the statements below, $\{Z_i\}_{i\in\mathbb{Z}}$ is any sequence of random variables):

(C4) For any $\{Z_i \in \mathcal{L}^1\}_{i\in\mathbb{Z}}$, $\{\mathcal{P}_i Z_i\}_{i\in\mathbb{Z}}$ are martingale differences with respect to the increasing filtration $\{\mathcal{F}_i\}_{i\in\mathbb{Z}}$. Thus, $\sum_{i=1}^n \mathcal{P}_i Z_i$ is a martingale with respect to $\mathcal{F}_n$.

(C5) For any $\{Z_i \in \mathcal{L}^2\}_{i\in\mathbb{Z}}$, $\|\sum_{i=1}^n \mathcal{P}_i Z_i\|_2^2 = \sum_{i=1}^n \|\mathcal{P}_i Z_i\|_2^2$.

(C6) For any $Z \in \mathcal{L}^2$, $\|\mathcal{P}_i Z\|_2^2 \leq \|\mathbb{E}(Z|\mathcal{F}_i)\|_2^2 \leq \|Z\|_2^2$.

(C7) For any $Z \in \mathcal{L}^2$, $\mathbb{E}[(\mathcal{P}_i Z)^2|\mathcal{F}_{i-1}] = \mathbb{E}\{[\mathbb{E}(Z|\mathcal{F}_i)]^2|\mathcal{F}_{i-1}\} - [\mathbb{E}(Z|\mathcal{F}_{i-1})]^2$.

(C8) For any $Z \in \mathcal{L}^2$, $\mathbb{E}[(\mathcal{P}_i Z)^2|\mathcal{F}_{i-1}] \leq \mathbb{E}(Z^2|\mathcal{F}_{i-1})$.

Proof. By definition, $\mathcal{P}_i Z_i$ is $\mathcal{F}_i$-measurable. Furthermore, by property (C2), $\mathbb{E}(\mathcal{P}_i Z_i|\mathcal{F}_{i-1}) = \mathbb{E}[\mathbb{E}(Z_i|\mathcal{F}_i)|\mathcal{F}_{i-1}] - \mathbb{E}(Z_i|\mathcal{F}_{i-1}) = 0$. Thus, (C4) holds. By (C4), (C5) follows from the orthogonality of martingale differences. To see (C6), let $Z^* = \mathbb{E}(Z|\mathcal{F}_i)\mathbb{E}(Z|\mathcal{F}_{i-1})$, by property (C2), $\mathbb{E}(Z^*|\mathcal{F}_{i-1}) = [\mathbb{E}(Z|\mathcal{F}_{i-1})]^2$. Thus, $\mathbb{E}(Z^*) = \mathbb{E}[\mathbb{E}(Z^*|\mathcal{F}_{i-1})] = \mathbb{E}\{[\mathbb{E}(Z|\mathcal{F}_{i-1})]^2\}$. Using the latter identity, we can show $\|\mathcal{P}_i Z\|_2^2 = \mathbb{E}\{[\mathbb{E}(Z|\mathcal{F}_i)]^2\} - \mathbb{E}\{[\mathbb{E}(Z|\mathcal{F}_{i-1})]^2\} \leq \|\mathbb{E}(Z|\mathcal{F}_i)\|_2^2 \leq \|Z\|_2^2$, where the last inequality follows from property (C3). By simple calculations, (C7) follows from the definition of $\mathcal{P}_i$ and property (C2). Finally, by (C7), (C8) follows from $\mathbb{E}[(\mathcal{P}_i Z)^2|\mathcal{F}_{i-1}] \leq \mathbb{E}\{[\mathbb{E}(Z|\mathcal{F}_i)]^2|\mathcal{F}_{i-1}\} \leq \mathbb{E}\{\mathbb{E}(Z^2|\mathcal{F}_i)|\mathcal{F}_{i-1}\} = \mathbb{E}(Z^2|\mathcal{F}_{i-1})$. ♢

By (C4), $\{\mathcal{P}_i\}_{i\in\mathbb{Z}}$ are martingale difference operators with respect to $\{\mathcal{F}_i\}_{i\in\mathbb{Z}}$. This idea of martingale construction serves as the building block for our technical arguments. See Wu (2005) for more discussions.

4.4.2 Some preliminary results on mixing processes

In Condition 3, we impose α-mixing conditions on $\{X_i\}$. Lemmas 1–2 below present some useful results for α-mixing processes.

Lemma 1 (Proposition 2.5 in Fan and Yao (2003))**.** *Let U and V be two random variables such that $U \in \mathcal{L}^p$ and $V \in \mathcal{L}^q$ for some $p > 1, q > 1$, and $1/p + 1/q < 1$. Then*

$$|\mathrm{cov}(U, V)| \leq 8\alpha(U, V)^{1-1/p-1/q}\|U\|_p\|V\|_q.$$

Here $\alpha(U, V)$ is the α-mixing coefficient between the two σ-algebras generated by U and V.

Next, we present an important inequality regarding the supremum of any differentiable function $f(\cdot)$ on a given bounded interval $[a, b]$. Note that $|f(y)| = |f(a) + \int_a^y f'(z)dz| \leq |f(a)| + \int_a^b |f'(z)|dz$ for all $y \in [a, b]$. Thus, using $(u+v)^2 \leq 2(u^2+v^2)$ and the Cauchy-Schwarz inequality $[\int_a^b |f'(z)|dz]^2 \leq (b-a)\int_a^b |f'(z)|^2dz$, we have the uniform bound:

$$\sup_{y\in[a,b]} |f(y)|^2 \leq 2\left\{|f(a)|^2 + \Big[\int_a^b |f'(z)|dz\Big]^2\right\} \leq 2\left\{|f(a)|^2 + (b-a)\int_a^b |f'(z)|^2dz\right\}. \quad (4.16)$$

Clearly, if $f(\cdot)$ is a random function, then taking expectation on both sides of (4.16) gives

$$\begin{aligned}\left\|\sup_{y\in[a,b]} |f(y)|\right\|_2^2 &\leq 2\left\{\|f(a)\|_2^2 + (b-a)\int_a^b \|f'(z)\|_2^2dz\right\} \\ &\leq 2\left\{\|f(a)\|_2^2 + (b-a)^2 \sup_{y\in[a,b]} \|f'(y)\|_2^2\right\}. \quad (4.17)\end{aligned}$$

In (4.17), while it is generally difficult to study the left hand side with "sup" inside $\|\cdot\|_2$, it is much easier to handle the right hand side with "sup" outside $\|\cdot\|_2$. Thus, (4.17) provides a useful inequality in bounding the supremum of random processes indexed by a continuous parameter. In particular, we can obtain the following useful result:

Lemma 2. *Let $\{X_i\}_{i\in\mathbb{N}}$ be an α-mixing stationary process with mixing coefficient $\alpha_k, k \in \mathbb{N}$. For a bivariate measurable and differentiable function h, define*

$$H(y) = \sum_{i=1}^{n} \Big\{ h(y, X_i) - \mathbb{E}[h(y, X_i)] \Big\}.$$

Suppose there exists some $\delta > 2$ such that $\sum_{k=1}^{\infty} \alpha_k^{1-2/\delta} < \infty$ and $c := \sup_{y\in\mathbb{R}}[\|h(y, X_1)\|_\delta + \|\partial h(y, X_1)/\partial y\|_\delta] < \infty$. Let $[a, b]$ be any given bounded interval. Then

$$\mathbb{E}\left[\sup_{y\in[a,b]} |H(y)|^2 \right] = O(n). \tag{4.18}$$

Furthermore, if $b_n \to 0$ and $w(\cdot)$ is an integrable function with bounded support, then

$$\mathbb{E}\left[\sup_{y\in[a,b]} \Big| \int_{\mathbb{R}} w(u) H(y - ub_n) du \Big| \right]^2 = O(n). \tag{4.19}$$

Proof. Let $\gamma_k = \mathrm{cov}\{h(y, X_1), h(y, X_{k+1})\}$. Then $\gamma_0 \le \|h(y, X_1)\|_2^2 \le \|h(y, X_1)\|_\delta^2 \le c^2$. For $k \ge 1$, by Lemma 1, $|\gamma_k| \le 8\alpha_k^{1-2/\delta} \|h(y, X_1)\|_\delta^2 \le 8\alpha_k^{1-2/\delta} c^2$. Thus,

$$\|H(y)\|_2^2 = n\gamma_0 + 2\sum_{k=1}^{n-1}(n-k)\gamma_k \le n\left(\gamma_0 + 2\sum_{k=1}^{n} |\gamma_k| \right) \le nc^2 \left(1 + 16 \sum_{k=1}^{\infty} \alpha_k^{1-2/\delta} \right). \tag{4.20}$$

Similarly, using $H'(y) = \sum_{i=1}^{n} \{\partial h(y, X_i)/\partial y - \mathbb{E}[\partial h(y, X_i)/\partial y]\}$, we have

$$\|H'(y)\|_2^2 \le nc^2 \left(1 + 16 \sum_{k=1}^{\infty} \alpha_k^{1-2/\delta} \right). \tag{4.21}$$

The assertion (4.18) then follows by applying (4.20) and (4.21) to (4.17). To prove (4.19), by the bounded support of $w(\cdot)$ and $b_n \to 0$, for $y \in [a, b]$, we have $y - ub_n \in [-a - 1, b + 1]$ for sufficiently large n. Thus,

$$\sup_{y\in[a,b]} \Big| \int_{\mathbb{R}} w(u) H(y - ub_n) du \Big| \le \sup_{z\in[-a-1,b+1]} |H(z)| \int_{\mathbb{R}} |w(u)| du. \tag{4.22}$$

Taking square first and then taking expectation in (4.22), we can obtain (4.19) from (4.18). ♢

4.4.3 Proof of Theorem 9

Throughout our proofs, $c, c_1, c_2, \ldots$ are constants that may vary from places to places.

Proof of Theorem 9. Recall $\hat{\mathcal{G}}_g(y)$ in (4.9). Define

$$\tilde{f}_Y(y) = \frac{1}{nb_n}\sum_{i=1}^{n} K_{b_n}(y - Y_{i-1}), \tag{4.23}$$

$$\xi_i(y) = [g(Y_i) - \mathcal{G}_g(Y_{i-1})]K_{b_n}(y - Y_{i-1}). \tag{4.24}$$

By the definition of $\mathcal{G}_g$, we have $\mathbb{E}[\xi_i(y)] = \mathbb{E}\{\mathbb{E}[\xi_i(y)|Y_{i-1}]\} = 0$. Therefore, we can write

$$\begin{aligned}\hat{\mathcal{G}}_g(y) - \mathcal{G}_g(y) &= \frac{\sum_{i=1}^{n}\{\xi_i(y) - \mathbb{E}[\xi_i(y)]\}}{nb_n\tilde{f}_Y(y)} + \frac{\sum_{i=1}^{n}[\mathcal{G}_g(Y_{i-1}) - \mathcal{G}_g(y)]K_{b_n}(y - Y_{i-1})}{nb_n\tilde{f}_Y(y)} \\ &:= T_n(y) + U_n(y). \end{aligned} \tag{4.25}$$

In (4.25), $T_n(y)$ is the stochastic component determining the asymptotic distribution of $\hat{\mathcal{G}}_g(y)$, and $U_n(y)$ is the bias component. By Lemma 4 below, $\tilde{f}_Y(y) = f_Y(y) + O[b_n^2 + (nb_n/\log n)^{-1/2}]$ uniformly in $y \in \mathcal{Y}$. Furthermore, by Lemma 5 below, $U_n(y) = \rho_g(y)b_n^2 + O_p[b_n^4 + (b_n \log n/n)^{1/2}] = \rho_g(y)b_n^2 + o_p[(nb_n \log n)^{-1/2}]$ uniformly in $y \in \mathcal{Y}$. By Slutsky's theorem, it suffices to establish a maximal deviation result for $\sum_{i=1}^{n}\{\xi_i(y) - \mathbb{E}[\xi_i(y)]\}$.

We use the projection operator $\mathcal{P}_i$ in Section 4.4.1 on page 65 to write the decomposition

$$\begin{aligned} & \sum_{i=1}^{n}\{\xi_i(y) - \mathbb{E}[\xi_i(y)]\} \\ = & \sum_{i=1}^{n}\{\mathcal{P}_i\xi_i(y) + \mathcal{P}_{i-1}\xi_i(y) + \mathbb{E}[\xi_i(y)|\mathcal{F}_{i-2}] - \mathbb{E}[\xi_i(y)]\} \\ = & \sum_{i=1}^{n}[\mathcal{P}_i\xi_i(y) + \mathcal{P}_i\xi_{i+1}(y)] + \sum_{i=1}^{n}\{\mathbb{E}[\xi_i(y)|\mathcal{F}_{i-2}] - \mathbb{E}[\xi_i(y)]\} + [\mathcal{P}_0\xi_1(y) - \mathcal{P}_n\xi_{n+1}(y)] \end{aligned}$$

$$:= \quad S_n(y) + R_n(y) + M_n(y). \tag{4.26}$$

The decomposition (4.26) provides a convenient tool to study asymptotic properties. First, by property (C4) in Section 4.4.1, $S_n(y)$ is a martingale for which classical martingale asymptotic tools are available. Second, by Lemma 3 below, $\sup_{y\in\mathcal{Y}} |R_n(y)| = O_p(b_n\sqrt{n})$. Finally, it is easy to observe that $\sup_y |M_n(y)| = O_p(1)$. To see this, by the boundedness of $K(\cdot)$,

$$\sup_y |\mathcal{P}_0\xi_1(y)| \le \sup_u |K(u)| \Big\{ \mathbb{E}[|g(Y_1) - \mathcal{G}_g(Y_0)||\mathcal{F}_0] + \mathbb{E}[|g(Y_1) - \mathcal{G}_g(Y_0)||\mathcal{F}_{-1}] \Big\}.$$

Thus, by property (C1) in Section 4.4.1,

$$\mathbb{E}\Big[\sup_y |\mathcal{P}_0\xi_1(y)| \Big] \le 2 \sup_u |K(u)| [\mathbb{E}|g(Y_1)| + \mathbb{E}|\mathcal{G}_g(Y_0)|] = O(1),$$

where we have $\mathbb{E}|\mathcal{G}_g(Y_0)| = \mathbb{E}|\mathbb{E}[g(Y_1)|Y_0]| \le \mathbb{E}\{\mathbb{E}[|g(Y_1)||Y_0]\} = \mathbb{E}|g(Y_1)| = O(1)$. Similarly, $\mathbb{E}[\sup_y |\mathcal{P}_n\xi_{n+1}(y)|] = O(1)$. This proves $\sup_y |M_n(y)| = O_p(1)$. Finally, the desired result then follows from the maximal deviation of $S_n(y)$ in Lemma 7 on page 76. ♢

Lemma 3. *For $R_n(y)$ in (4.26), we have $\sup_{y\in\mathcal{Y}} |R_n(y)| = O_p(b_n\sqrt{n})$.*

Proof. Recall $\xi_i(y)$ in (4.24). Write $\xi_i(y) = \xi_{i,1}(y) - \xi_{i,2}(y)$, where

$$\xi_{i,1}(y) = g(Y_i)K_{b_n}(y - Y_{i-1}) \quad \text{and} \quad \xi_{i,2}(y) = \mathcal{G}_g(Y_{i-1})K_{b_n}(y - Y_{i-1}). \tag{4.27}$$

Then it suffices to prove $\sup_{y\in\mathcal{Y}} |J_r(y)| = O_p(b_n\sqrt{n})$, where

$$J_r(y) = \sum_{i=1}^{n} \Big\{ \mathbb{E}[\xi_{i,r}(y)|\mathcal{F}_{i-2}] - \mathbb{E}[\xi_{i,r}(y)] \Big\}, \quad r = 1, 2.$$

First, we consider $J_1(y)$. Since $\{\varepsilon_i\}_{i\in\mathbb{N}}$ are i.i.d. and independent of $\{X_i\}_{i\in\mathbb{N}}$, by writing $Y_{i-1} = X_{i-1} + \varepsilon_{i-1}$, we have

$$\begin{aligned} \mathbb{E}[\xi_{i,1}(y)|\mathcal{F}_{i-2}, X_i] &= \mathbb{E}[g(Y_i)K_{b_n}(y - X_{i-1} - \varepsilon_{i-1})|X_{i-1}, X_i] \\ &= \mathbb{E}\Big[g(Y_i) \int_{\mathbb{R}} K_{b_n}(y - X_{i-1} - v) q_\varepsilon(v) dv | X_{i-1}, X_i \Big] \end{aligned}$$

$$= \quad b_n \int_{-\omega}^{\omega} K(u) e_i(u) du, \quad e_i(u) = \mathbb{E}[g(Y_i) q_\varepsilon(y - X_{i-1} - u b_n) | X_{i-1}, X_i].$$

Here, the last equality follows from the transformation $u = (y - X_{i-1} - v)/b_n$. Note that the conditional expectation $e_i(u)$ is a function of X_{i-1}, X_i. Thus,

$$\begin{aligned}
\mathbb{E}[\xi_{i,1}(y)|\mathcal{F}_{i-2}] &= \mathbb{E}\{\mathbb{E}[\xi_{i,1}(y)|\mathcal{F}_{i-2}, X_i]|\mathcal{F}_{i-2}\} \\
&= b_n \int_{-\omega}^{\omega} K(u) \mathbb{E}[e_i(u)|X_{i-1}] du \\
&= b_n \int_{-\omega}^{\omega} K(u) \mathbb{E}[g(Y_i) q_\varepsilon(y - X_{i-1} - u b_n)|X_{i-1}] du, \quad (4.28)
\end{aligned}$$

where the first equality follows from property (C2) in Section 4.4.1, the second equality follows from the independence between $e_i(u)$ (which is a function of X_{i-1}, X_i) and $\{\varepsilon_i\}_{i\in\mathbb{N}}$ as well as the Markovian assumption on $\{X_i\}_{i\in\mathbb{N}}$, and the third equality follows from $\mathbb{E}[e_i(u)|X_{i-1}] = \mathbb{E}[g(Y_i) q_\varepsilon(y - X_{i-1} - u b_n)|X_{i-1}]$ (property (C2) in Section 4.4.1). Define $h(z, X_{i-1}) = \mathbb{E}[g(Y_i) q_\varepsilon(z - X_{i-1})|X_{i-1}]$ (After taking conditional expectation, it is a function of X_{i-1}). Then, using $\mathbb{E}[\xi_{i,1}(y)] = \mathbb{E}\{\mathbb{E}[\xi_{i,1}(y)|\mathcal{F}_{i-2}]\}$ and by (4.28), we obtain

$$J_1(y) = b_n \int_{-\omega}^{\omega} K(u) H(y - u b_n) du, \quad \text{where} \quad H(z) = \sum_{i=1}^{n} [h(z, X_{i-1}) - \mathbb{E}h(z, X_{i-1})].$$

Since $q_\varepsilon(\cdot)$ is bounded, by property (C3) in Section 4.4.1, we can easily see that $\|h(z, X_0)\|_\delta = O(1)\|\mathbb{E}[|g(Y_1)||X_0]\|_\delta \le O(1)\|g(Y_1)\|_\delta$ for all $z \in \mathbb{R}$. Similarly, using $\partial h(z, X_0)/\partial z = \mathbb{E}[g(Y_1) q'_\varepsilon(z - X_0)|X_0]$ and the boundedness of $q'_\varepsilon(\cdot)$, we have $\|\partial h(z, X_0)/\partial z\|_\delta \le O(1)\|g(Y_1)\|_\delta$ for all $z \in \mathbb{R}$. Thus, by (4.19) in Lemma 2, we conclude that $\sup_{y\in\mathcal{Y}} |J_1(y)| = O_p(b_n\sqrt{n})$.

Next, we consider $J_2(y)$. Using $Y_{i-1} = X_{i-1} + \varepsilon_{i-1}$, we obtain

$$\begin{aligned}
\mathbb{E}[\xi_{i,2}(y)|\mathcal{F}_{i-2}] &= \int_{\mathbb{R}} \mathcal{G}_g(X_{i-1} + v) K_{b_n}(y - X_{i-1} - v) q_\varepsilon(v) dv \\
&= b_n \int_{-\omega}^{\omega} K(u) \mathcal{G}_g(y - u b_n) q_\varepsilon(y - X_{i-1} - u b_n) du. \quad (4.29)
\end{aligned}$$

Thus, using $\mathbb{E}[\xi_{i,2}(y)] = \mathbb{E}\{\mathbb{E}[\xi_{i,2}(y)|\mathcal{F}_{i-2}]\}$, we have

$$J_2(y) = b_n \int_{-\omega}^{\omega} K(u)\mathcal{G}_g(y - ub_n)L(y - ub_n)du, \tag{4.30}$$

where $L(z) = \sum_{i=0}^{n-1}\{q_\varepsilon(z - X_i) - \mathbb{E}[q_\varepsilon(z - X_i)]\}$. Since $\mathcal{G}_g(\cdot)$ is bounded in the neighborhood $\mathcal{Y}_\epsilon$, the claim then follows from (4.19) in Lemma 2. ♢

Lemma 4. *For $\tilde{f}_Y(y)$ in (4.23), we have*

$$\sup_{y\in\mathcal{Y}} |\tilde{f}_Y(y) - f_Y(y)| = O_p[b_n^2 + (nb_n/\log n)^{-1/2}].$$

Proof. Let $\gamma_i(y) = K_{b_n}(y - Y_i)$. Observe the decomposition

$$\begin{aligned}\tilde{f}_Y(y) &= \frac{1}{nb_n}\sum_{i=0}^{n-1}\mathcal{P}_i\gamma_i(y) + \frac{1}{nb_n}\sum_{i=0}^{n-1}\{\mathbb{E}[\gamma_i(y)|\mathcal{F}_{i-1}] - \mathbb{E}[\gamma_i(y)]\} + \frac{\mathbb{E}\gamma_1(y)}{b_n} \\ &:= \frac{1}{nb_n}H_1(y) + \frac{1}{nb_n}H_2(y) + \frac{\mathbb{E}[\gamma_1(y)]}{b_n}. \end{aligned}\tag{4.31}$$

By the symmetry of $K(\cdot)$, we can show $b_n^{-1}\mathbb{E}[\gamma_1(y)] = f_Y(y) + O(b_n^2)$. To prove the desired result, it suffices to show $\sup_{y\in\mathcal{Y}}|H_1(y)| = O_p(\sqrt{nb_n\log n})$ and $\sup_{y\in\mathcal{Y}}|H_2(y)| = O_p(b_n\sqrt{n})$.

First, we consider $H_2(y)$. By the same argument in (4.29)–(4.30), $H_2(y) = b_n\int_{-\omega}^{\omega}K(u)L(y - ub_n)du$ with $L(\cdot)$ defined in (4.30). Thus, by (4.19) in Lemma 2, $\sup_{y\in\mathcal{Y}}|H_2(y)| = O_p(b_n\sqrt{n})$.

Next, we consider the martingale part $H_1(y)$. We shall adopt a chain argument to approximate $H_1(y), y \in \mathcal{Y}$, on increasingly denser grid points. Let $N = n^2$ and $y_j = jT/N, j = -N, 1-N, \ldots, N-1, N$. Then $y_{-N}, \ldots, y_N$ partition $\mathcal{Y} = [-T, T]$ into $2N$ equally spaced intervals with length T/N. By the bounded derivative of $K(\cdot)$, there exists some constant c_1 such that, for all $y \in [y_j, y_{j+1}]$,

$$|\gamma_i(y) - \gamma_i(y_j)| + |\mathbb{E}[\gamma_i(y)|\mathcal{F}_{i-1}] - \mathbb{E}[\gamma_i(y_j)|\mathcal{F}_{i-1}]| \le c_1|y - y_j|/b_n \le c_1T/(Nb_n).$$

Thus, $\sup_{y\in[y_j,y_{j+1}]}|H_1(y)-H_1(y_j)|\le nc_1T/(Nb_n)=O[(nb_n)^{-1}]$, and consequently,

$$\sup_{y\in\mathcal{Y}}|H_1(y)|\le\max_{j=-N,\ldots,N}|H_1(y_j)|+O[(nb_n)^{-1}]. \tag{4.32}$$

By property (C8) in Section 4.4.1, for some constant c_2,

$$\sum_{i=0}^{n-1}\mathbb{E}\{[\mathcal{P}_i\gamma_i(y_j)]^2|\mathcal{F}_{i-1}\}\le\sum_{i=0}^{n-1}\mathbb{E}[\gamma_i^2(y_j)|\mathcal{F}_{i-1}]\le c_2nb_n. \tag{4.33}$$

Let $c_3=\sup_u|K(u)|$. Then $|\mathcal{P}_i\gamma_i(y_j)|\le 2c_3$. Thus, by Freedman's exponential inequality for bounded martingale differences [Freedman (1975)], for any $c>0$,

$$\begin{aligned} p_j &:= \mathbb{P}\Big\{|H_1(y_j)|\ge c\sqrt{nb_n\log n}\Big\}\\ &\le 2\exp\left[-\frac{c^2nb_n\log n}{2(2c_3c\sqrt{nb_n\log n}+c_2nb_n)}\right]\\ &= 2\exp(-\lambda\log n),\quad \lambda=\frac{c^2}{4c_3c\sqrt{\frac{\log n}{nb_n}}+2c_2}. \end{aligned} \tag{4.34}$$

Since $(nb_n^3)^{-1}\log n\to 0$, $(nb_n)^{-1}\log n<1$ for sufficiently large n. Thus, $\lambda>c^2/(4c_3c+2c_2)\ge 3$ by choosing a large enough c (for example, we may take $c=12c_3+\sqrt{6c_2}$). Then

$$\mathbb{P}\Big\{\max_{j=-N,\ldots,N}|H_1(y_j)|\ge c\sqrt{nb_n\log n}\Big\}\le\sum_{j=-N}^{N}p_j=O(Nn^{-\lambda})=O(1/n)\to 0.$$

Therefore, $\max_{j=-N,\ldots,N}|H_1(y_j)|=O_p(\sqrt{nb_n\log n})$. The result then follows from (4.32). ♢

Lemma 5. *Recall $\rho_g(y)$ in Theorem 9 on page 57. Then*

$$\sum_{i=1}^{n}[\mathcal{G}_g(Y_{i-1})-\mathcal{G}_g(y)]K_{b_n}(y-Y_{i-1})=nb_n^3f_Y(y)\{\rho_g(y)+O_p[b_n^2+(nb_n^3/\log n)^{-1/2}]\}.$$

Proof. We adopt the same argument in Lemma 4. Let $\eta_i(y)=[\mathcal{G}_g(Y_i)-\mathcal{G}_g(y)]K_{b_n}(y$

$-Y_i$). As in (4.31), we use the decomposition

$$\sum_{i=1}^{n}[\mathcal{G}_g(Y_{i-1})-\mathcal{G}_g(y)]K_{b_n}(y-Y_{i-1})=\sum_{i=0}^{n-1}\eta_i(y)=N_1(y)+N_2(y)+n\mathbb{E}[\eta_1(y)],$$

where

$$\begin{aligned} N_1(y) &= \sum_{i=0}^{n-1}\{\eta_i(y)-\mathbb{E}[\eta_i(y)|\mathcal{F}_{i-1}]\}=\sum_{i=0}^{n-1}\mathcal{P}_i\eta_i(y), \\ N_2(y) &= \sum_{i=0}^{n-1}\{\mathbb{E}[\eta_i(y)|\mathcal{F}_{i-1}]-\mathbb{E}[\eta_i(y)]\}. \end{aligned}$$

By the symmetry of $K(\cdot)$ and Taylor's expansion, we can show $\mathbb{E}[\eta_1(y)]=b^3f_Y(y)[\rho_g(y)+O(b_n^2)]$. For $N_2(y)$, by the same argument in (4.29)–(4.30), we can obtain

$$N_2(y)=b_n\int_{-\omega}^{\omega}K(u)[\mathcal{G}_g(y-ub_n)-\mathcal{G}_g(y)]L(y-ub_n)du,$$

where $L(\cdot)$ is defined in (4.30). Note that $|\mathcal{G}_g(y-ub_n)-\mathcal{G}_g(y)|=O(b_n)$. Thus, by (4.19) in Lemma 2, $\sup_{y\in\mathcal{Y}}|N_2(y)|=O_p(b_n^2\sqrt{n})$. For the martingale part $N_1(y)$, using $\mathcal{G}_g(Y_i)-\mathcal{G}_g(y)=O(b_n)$ for $y-Y_i=O(b_n)$, we have $\mathbb{E}[\eta_i^2(y)|\mathcal{F}_{i-1}]=O(b_n^2)\mathbb{E}[K_{b_n}^2(y-Y_i)|\mathcal{F}_{i-1}]=O(b_n^3)$. Thus, by property (C8) in Section 4.4.1, the conditional variance satisfies $\sum_{i=0}^{n-1}\mathbb{E}\{[\mathcal{P}_i\eta_i(y)]^2|\mathcal{F}_{i-1}\}\le\sum_{i=0}^{n-1}\mathbb{E}[\eta_i^2(y)|\mathcal{F}_{i-1}]=O(nb_n^3)$. By the same chain argument in the proof of $H_1(y)$ in Lemma 4, we can show $\sup_{y\in\mathcal{Y}}|N_1(y)|=O_p(\sqrt{nb_n^3\log n})$. ♢

Recall $S_n(y)$ in (4.26). By property (C4) in Section 4.4.1, $S_n(y)$ is a martingale with respect to $\mathcal{F}_n$. To study asymptotic properties of $S_n(y)$, Lemma 6 below studies its conditional variance. In Lemma 7, we use the obtained result to study the quadratic characteristic matrix of the multivariate martingale $[S_n(y_1),\ldots,S_n(y_k)]^T$ for distinct $y_1,\ldots,y_k$.

Lemma 6. *Recall $\xi_i(y)$ in (4.24). Define*

$$d_i(y)=\frac{\mathcal{P}_i[\xi_i(y)+\xi_{i+1}(y)]}{\sigma_g(y)\sqrt{nb_n\varphi_K f_Y(y)}}. \tag{4.35}$$

Then

$$\sup_{y\in\mathcal{Y}}\left\|\sum_{i=1}^{n}\mathbb{E}[d_i^2(y)|\mathcal{F}_{i-1}]-1\right\|_{3/2}=O[(nb_n)^{-1/2}+b_n^{2/3}]. \tag{4.36}$$

Proof. We drop the argument "y" and write $\xi_i=\xi_i(y)$. By property (C7), we can show

$$\begin{aligned}\mathbb{E}\{[\mathcal{P}_i(\xi_i+\xi_{i+1})]^2|\mathcal{F}_{i-1}\} &= 2\mathbb{E}(\xi_i\xi_{i+1}|\mathcal{F}_{i-1})-2\mathbb{E}(\xi_i|\mathcal{F}_{i-1})\mathbb{E}(\xi_{i+1}|\mathcal{F}_{i-1})-[\mathbb{E}(\xi_{i+1}|\mathcal{F}_{i-1})]^2\\ &\quad+\mathbb{E}(\xi_i^2|\mathcal{F}_{i-1})-[\mathbb{E}(\xi_i|\mathcal{F}_{i-1})]^2+\mathbb{E}\{[\mathbb{E}(\xi_{i+1}|\mathcal{F}_i)]^2|\mathcal{F}_{i-1}\}\\ &:= 2A_{i,1}-2A_{i,2}-A_{i,3}+A_{i,4}-A_{i,5}+A_{i,6}. \end{aligned}\tag{4.37}$$

Below we consider each of these six terms separately. For convenience, sometimes we give bounds for $\|\cdot\|_2$, which dominates $\|\cdot\|_{3/2}$.

(*$A_{i,5}$ and $A_{i,6}$ terms:*) Let $\nu_i=[\mathbb{E}(\xi_{i+1}|\mathcal{F}_i)]^2$. Note that $A_{i+1,5}-A_{i,6}=\mathcal{P}_i\nu_i$. By consecutively using properties (C5), (C6) and (C3) in Section 4.4.1, we have

$$\left\|\sum_{i=1}^{n}(A_{i+1,5}-A_{i,6})\right\|_2^2=\sum_{i=1}^{n}\|\mathcal{P}_i\nu_i\|_2^2\le\sum_{i=1}^{n}\|\nu_i\|_2^2\le n\mathbb{E}(\xi_1^4)=O(nb_n). \tag{4.38}$$

Therefore, by the triangle inequality,

$$\left\|\sum_{i=1}^{n}(A_{i,5}-A_{i,6})\right\|_2=\left\|\nu_0-\nu_n+\sum_{i=1}^{n}(A_{i+1,5}-A_{i,6})\right\|_2=O(\sqrt{nb_n}). \tag{4.39}$$

(*$A_{i,3}$ terms:*) Recall $\xi_{i,1}(y)$ and $\xi_{i,2}(y)$ in (4.27). By $\xi_i=\xi_{i,1}(y)-\xi_{i,2}(y)$, (4.28) and (4.29), and the boundedness of $K(\cdot)$, $q_\varepsilon(\cdot)$ and $\mathcal{G}_g(y), y\in\mathcal{Y}_\epsilon$, there exists a constant c_1 such that

$$|\mathbb{E}(\xi_{i+1}|\mathcal{F}_{i-1})|\le c_1b_n\{1+\mathbb{E}[|g(Y_{i+1})||X_i]\},\quad\text{for all } y. \tag{4.40}$$

Using $\{1+\mathbb{E}[|g(Y_{i+1})||X_i]\}^2\le 2+2\{\mathbb{E}[|g(Y_{i+1})||X_i]\}^2\le 2+2\mathbb{E}[g^2(Y_{i+1})|X_i]$, we obtain

$$\left\|\sum_{i=1}^{n}A_{i,3}\right\|_2 \le 2c_1^2b_n^2\left\|\sum_{i=1}^{n}\{1+\mathbb{E}[g^2(Y_{i+1})|X_i]\}\right\|_2$$

$$\leq \quad 2c_1^2 b_n^2 \sum_{i=1}^{n} \|1 + \mathbb{E}[g^2(Y_{i+1})|X_i]\|_2 = O(nb_n^2). \tag{4.41}$$

Here, the last equality follows from $\|\mathbb{E}[g^2(Y_{i+1})|X_i]\|_2 \leq \|g^2(Y_{i+1})\|_2 = \|g(Y_1)\|_4^2 < \infty$.

($A_{i,2}$ terms:) By the bounded support of $K(\cdot)$, it suffices to consider $|Y_{i-1}-y| = O(b_n)$ in ξ_i so that $|\mathcal{G}_g(Y_{i-1})| \leq c_2$ for some constant c_2. Note that $\mathbb{E}[|g(Y_i)||\mathcal{F}_{i-1}] = \mathbb{E}[|g(Y_i)||X_i]$. Thus, $|\mathbb{E}(\xi_i|\mathcal{F}_{i-1})| \leq |K_{b_n}(y - Y_{i-1})|\{c_2 + \mathbb{E}[|g(Y_i)||X_i]\}$. Combining this with (4.40) gives

$$|A_{i,2}| \leq c_1 b_n |K_{b_n}(y - Y_{i-1})| Z_i, \quad Z_i = \{c_2 + \mathbb{E}[|g(Y_i)||X_i]\}\{1 + \mathbb{E}[|g(Y_{i+1})||X_i]\}. \tag{4.42}$$

By the same argument in (4.29), $\mathbb{E}[|K_{b_n}(y-Y_{i-1})|^{3/2}|X_{i-1}, X_i] \leq c_3 b_n$ for some constant c_3. Also, Z_i is a function of X_i and independent of ε_{i-1}. Thus, $\mathbb{E}[|K_{b_n}(y - Y_{i-1})Z_i|^{3/2}] = \mathbb{E}\{\mathbb{E}[|K_{b_n}(y - Y_{i-1})Z_i|^{3/2}|X_{i-1}, X_i]\} \leq c_3 b_n \mathbb{E}(|Z_i|^{3/2})$. Now, by (4.42), we have

$$\left\| \sum_{i=1}^{n} A_{i,2} \right\|_{3/2} \leq \sum_{i=1}^{n} \|A_{i,2}\|_{3/2} \leq c_1 b_n \sum_{i=1}^{n} \|K_{b_n}(y - Y_{i-1})Z_i\|_{3/2} = O(nb_n^{5/3}). \tag{4.43}$$

Here, we have used $\|Z_i\|_{3/2} \leq \|Z_1\|_2 < \infty$ by the condition $g(Y_i) \in \mathcal{L}^\delta$ with $\delta \geq 4$.

($A_{i,1}$ terms:) In $\xi_i\xi_{i+1}$, thanks to the terms $K_{b_n}(y - Y_{i-1})$ and $K_{b_n}(y - Y_i)$, it suffices to consider $|Y_{i-1}-y| = O(b_n)$ and $|Y_i-y| = O(b_n)$. Thus, $|g(Y_i)|+|\mathcal{G}_g(Y_i)|+|\mathcal{G}_g(Y_{i-1})| \leq c_4$ for some constant c_4, and consequently $|\xi_i| \leq c_4|K_{b_n}(y - Y_{i-1})|$. Also, conditioning on $\mathcal{F}_{i-1}$, the ε_i term in $Y_i = X_i + \varepsilon_i$ is independent of everything else, and thus the same argument in (4.29) shows that the term $K_{b_n}(y - Y_i)$ will result in an $O(b_n)$ factor. Thus,

$$|A_{i,1}| \leq O(b_n)|K_{b_n}(y - Y_{i-1})|\{c_4 + \mathbb{E}[|g(Y_{i+1})||\mathcal{F}_{i-1}]\}.$$

By the independence between $\{\varepsilon_i\}$ and $\{X_i\}$ as well as the Markovian assumption on $\{X_i\}$, $\mathbb{E}[|g(Y_{i+1})||\mathcal{F}_{i-1}] = \mathbb{E}[|g(Y_{i+1})||X_i]$. The same argument in (4.43) then

gives

$$\left\| \sum_{i=1}^{n} A_{i,1} \right\|_{3/2} \leq O(b_n) \sum_{i=1}^{n} \|K_{b_n}(y - Y_{i-1})\{c_4 + \mathbb{E}[|g(Y_{i+1})||X_i]\}\|_{3/2} = O(nb_n^{5/3}). \quad (4.44)$$

($A_{i,4}$ terms:) Write $A_{i,4} - \mathbb{E}(\xi_i^2) = \mathcal{P}_{i-1}\xi_i^2 + [\mathbb{E}(\xi_i^2|\mathcal{F}_{i-2}) - \mathbb{E}(\xi_i^2)]$. Since $\{\mathcal{P}_{i-1}\xi_i^2\}_{i\in\mathbb{Z}}$ are martingale differences with respect to $\{\mathcal{F}_{i-1}\}_{i\in\mathbb{Z}}$, by the same argument in (4.38),

$$\left\| \sum_{i=1}^{n} \mathcal{P}_{i-1}\xi_i^2 \right\|_2 = O(\sqrt{nb_n}). \quad (4.45)$$

By the same argument in (4.28) and (4.29), it can be shown that

$$\mathbb{E}(\xi_i^2|\mathcal{F}_{i-2}) \;=\; b_n \int_{-\omega}^{\omega} K^2(u)\mathbb{E}\{[g(Y_i) - \mathcal{G}_g(y - ub_n)]^2 q_\varepsilon(y - X_{i-1} - ub_n)|X_{i-1}\}.$$

Thus, as in the proof of Lemma 3, an application of Lemma 2 gives $\| \sum_{i=1}^{n} [\mathbb{E}(\xi_i^2|\mathcal{F}_{i-2}) - \mathbb{E}(\xi_i^2)] \|_2 = O(b_n\sqrt{n})$. The latter bound along with (4.45) gives

$$\left\| \sum_{i=1}^{n} [A_{i,4} - \mathbb{E}(\xi_i^2)] \right\|_2 \leq \left\| \sum_{i=1}^{n} \mathcal{P}_{i-1}\xi_i^2 \right\|_2 + \left\| \sum_{i=1}^{n} [\mathbb{E}(\xi_i^2|\mathcal{F}_{i-2}) - \mathbb{E}(\xi_i^2)] \right\|_2 = O(\sqrt{nb_n}). \quad (4.46)$$

Elementary calculation shows that $\mathbb{E}(\xi_i^2) = b_n\varphi_K f_Y(y)\sigma_g^2(y) + O(b_n^2)$. Finally, (4.36) then follows from (4.37), (4.39), (4.41), (4.43), (4.44) and (4.46) via the triangle inequality. ♢

Lemma 7. *Recall $S_n(y)$ in (4.26). Under the conditions and notations in Theorem 9 on page 57,*

$$\lim_{n\to\infty} \mathbb{P}\left\{ \sup_{y\in\mathcal{Y}_n} \frac{|S_n(y)|}{\sigma_g(y)\sqrt{nb_n\varphi_K f_Y(y)}} \leq B_{m_n}(z) \right\} = e^{-2e^{-z}}, \quad z \in \mathbb{R}. \quad (4.47)$$

Proof. Let $d_i(y)$ be defined in (4.35). Write

$$\tilde{S}_n(y) := \frac{|S_n(y)|}{\sigma_g(y)\sqrt{nb_n\varphi_K f_Y(y)}} = \sum_{i=1}^{n} d_i(y).$$

Write $\mathcal{Y}_n = \{y_1 < \cdots < y_{m_n}\}$. For fixed $k \in \mathbb{N}$ distinct integers $1 \le j_1, j_2, \cdots, j_k \le m_n$, define the k-dimensional column vectors

$$D_i = [d_i(y_{j_1}), \cdots, d_i(y_{j_k})]^T \quad \text{and} \quad S_{n,k} = \sum_{i=1}^{n} D_i = [\tilde{S}_n(y_{j_1}), \cdots, \tilde{S}_n(y_{j_k})]^T.$$

Then $\{D_i\}_{i\in\mathbb{Z}}$ are k-dimensional vectors of martingale differences with respect to $\{\mathcal{F}_i\}_{i\in\mathbb{Z}}$. Denote by Q_n the quadratic characteristic matrix of the martingale $S_{n,k}$, i.e.,

$$Q_n = \sum_{i=1}^{n} \mathbb{E}(D_i D_i^T | \mathcal{F}_{i-1}) := (q_{rs})_{1\le r,s\le k}.$$

Let $\tau_{rs} = \varphi_K \sigma_g(y_{j_r})\sigma_g(y_{j_s})\sqrt{f_Y(y_{j_r})f_Y(y_{j_s})}$. Then we can write q_{rs} as

$$\begin{aligned} q_{rs} &= \sum_{i=1}^{n} \mathbb{E}[d_i(y_{j_r})d_i(y_{j_s})|\mathcal{F}_{i-1}] \\ &= \frac{1}{nb_n\tau_{rs}} \sum_{i=1}^{n} \Big\{ \mathbb{E}[\mathcal{P}_i\xi_i(y_{j_r})\mathcal{P}_i\xi_i(y_{j_s})|\mathcal{F}_{i-1}] + \mathbb{E}[\mathcal{P}_i\xi_{i+1}(y_{j_r})\mathcal{P}_i\xi_{i+1}(y_{j_s})|\mathcal{F}_{i-1}] \\ &+ \mathbb{E}[\mathcal{P}_i\xi_i(y_{j_r})\mathcal{P}_i\xi_{i+1}(y_{j_s})|\mathcal{F}_{i-1}] + \mathbb{E}[\mathcal{P}_i\xi_{i+1}(y_{j_r})\mathcal{P}_i\xi_i(y_{j_s})|\mathcal{F}_{i-1}] \Big\}. \end{aligned} \tag{4.48}$$

For $r = s$, by Lemma 6, $\|q_{rr} - 1\|_{3/2} = O[(nb_n)^{-1/2} + b_n^{2/3}]$. For $r \neq s$, by the definition of $\mathcal{Y}_n$, since $|y_{j_r} - y_{j_s}| \ge \tau_n$, $b_n = o(\tau_n)$, and the kernel function $K(\cdot)$ has bounded support, we have $K_{b_n}(y_{j_r} - Y_{i-1})K_{b_n}(y_{j_s} - Y_{i-1}) = 0$ for large enough n. Thus, $\mathcal{P}_i\xi_i(y_{j_r})\mathcal{P}_i\xi_i(y_{j_s}) = 0$. For the other three terms on the right hand side of (4.48), their expansions of the form (4.37) involve terms of the form $A_{i,1}, A_{i,2}, A_{i,3}, A_{i,5}, A_{i,6}$ (the term $A_{i,4}$ vanishes, thanks to the choice of $\mathcal{Y}_n$ and the bounded support of $K(\cdot)$). Thus, we can use the same argument in Lemma 6 to show that their $\|\cdot\|_{3/2}$ norm can be bounded by $O[(nb_n)^{-1/2} + b_n^{2/3}]$. In summary, let I_{rs} be the (r,s)-element of the $k \times k$ identity matrix, then $\|q_{rs} -$

$I_{rs}\|_{3/2} = O[(nb_n)^{-1/2} + b_n^{2/3}]$ uniformly over $1 \le r, s \le k$. It is easily seen that $\sum_{i=1}^n \mathbb{E}|d_i(y_{j_r})|^3 = O[(nb_n)^{-1/2}]$ uniformly over $1 \le r \le k$. Thus $\sum_{i=1}^n \mathbb{E}|d_i(y_{j_r})|^3 + \mathbb{E}(|q_{rs} - I_{rs}|^{3/2}) = O(\Omega_n)$ uniformly, where $\Omega_n = (nb_n)^{-1/2} + b_n$.

For $j = 1, \ldots, m_n$, define events $A_j = \{|\tilde{S}_n(y_j)| \ge B_{m_n}(z)\}$ and $E_{m_n} = \{\sup_{y\in\mathcal{Y}_n} |\tilde{S}_n(y)| \ge B_{m_n}(z)\} = \cup_{j=1}^{m_n} A_j$. Let $N_1, N_2, \ldots$ be i.i.d. $N(0,1)$ variables. The imposed condition $(\log n)^3[(nb_n)^{-1} + b_n^2]m_n^2 \to 0$ implies $[1+B_{m_n}(z)]^4 \exp[B_{m_n}^2(z)/2]\Omega_n \to 0$ for fixed z. Then, by Theorem 1 in Grama and Haeusler (2006),

$$\mathbb{P}[\cap_{r=1}^k A_{j_r}] = \mathbb{P}[\cap_{r=1}^k \{|N_r| > B_{m_n}(z)\}][1 + o(1)] = \Big(\frac{2e^{-z}}{m_n}\Big)^k [1 + o(1)]. \quad (4.49)$$

Here the second equality follows from $\mathbb{P}(N_1 > x) = [1 + o(1)]\phi(x)/x$ as $x \to \infty$, where ϕ is the standard normal density. Therefore, by (4.49) and the inclusion-exclusion inequality, we have, for any fixed k and large enough n,

$$\begin{aligned}
\mathbb{P}(E_{m_n}) &\ge \sum_{j=1}^{m_n} \mathbb{P}(A_j) - \sum_{1\le j_1<j_2\le m_n} \mathbb{P}(A_{j_1} \cap A_{j_2}) + \cdots - \sum_{1\le j_1<\cdots<j_{2k}\le m_n} \mathbb{P}(\cap_{r=1}^{2k} A_{j_r}) \\
&= \sum_{r=1}^{2k} (-1)^{r-1} \binom{m_n}{r} \Big(\frac{2e^{-z}}{m_n}\Big)^r [1 + o(1)] \\
&= -\sum_{r=1}^{2k} \frac{(-2e^{-z})^r}{r!} [1 + o(1)].
\end{aligned}$$

Thus, $\liminf_{n\to\infty} \mathbb{P}(E_{m_n}) \ge -\sum_{r=1}^{2k} (-2e^{-z})^r/r!$. Similarly, we can use the inclusion-exclusion inequality to include $2k-1$ terms to obtain an upper bound, which gives $\limsup_{n\to\infty} \mathbb{P}(E_{m_n}) \le -\sum_{r=1}^{2k-1} (-2e^{-z})^r/r!$. Finally, letting $k \to \infty$, (4.47) follows since $\lim_{k\to\infty} \sum_{r=1}^{2k} (-2e^{-z})^r/r! = e^{-2e^{-z}} - 1$. ♢

4.4.4 Proof of Theorem 10

Lemma 8. *The uniform consistency holds:* $\sup_{y\in\mathcal{Y}} |\hat{\mathcal{G}}_g(y) - \mathcal{G}_g(y)| = O_p[b_n^2 + (b_n\sqrt{n})^{-1} \log n]$.

Proof. By (4.25), (4.26), Lemmas 3–4, it suffices to prove $\sup_{y\in\mathcal{Y}} |S_n(y)| = O_p(\sqrt{n} \log n)$, where $S_n(y) = \sum_{i=1}^n \mathcal{P}_i[\xi_i(y) + \xi_{i+1}(y)]$ is defined in (4.26). Again, we adopt the chain argument in Lemma 4 to establish the uniform

bound for the martingale $S_n(y)$.

Let $y_{-N}, \ldots, y_N$ be the grid points defined in the proof of Lemma 4. By the same chain argument in Lemma 4, it suffices to prove $\max_{-N \le j \le N} |S_n(y_j)| = O_p(\sqrt{n} \log n)$. However, since the summands $\mathcal{P}_i[\xi_i(y) + \xi_{i+1}(y)]$ and their conditional variances are no longer bounded, we cannot directly use Freedman's exponential inequality for bounded martingale differences. To solve this issue, we adopt the following argument. Define

$$\begin{aligned} A_1 &= \max_{1 \le i \le n} [\mathbb{E}(\zeta_i | \mathcal{F}_i) + \mathbb{E}(\zeta_i | \mathcal{F}_{i-1})], \quad \zeta_i = |g(Y_i)| + |\mathcal{G}_g(Y_{i-1})| + |g(Y_{i+1})| + |\mathcal{G}_g(Y_i)|, \\ A_2 &= \sum_{i=1}^{n} \mathbb{E}[|g(Y_i)|^2 + |\mathcal{G}_g(Y_{i-1})|^2 + |g(Y_{i+1})|^2 + |\mathcal{G}_g(Y_i)|^2 | \mathcal{F}_{i-1}]. \end{aligned}$$

Let $c_1 = \sup_u |K(u)|$. Then $|\xi_i(y)| + |\xi_{i+1}(y)| \le c_1 \zeta_i$ uniformly in y, and consequently

$$|\mathcal{P}_i[\xi_i(y) + \xi_{i+1}(y)]| \le c_1 A_1, \quad \text{uniformly in } i = 1, \ldots, n, y \in \mathbb{R}. \tag{4.50}$$

Now, consider conditional variance. By (C8) in Section 4.4.1,

$$\mathbb{E}(\{\mathcal{P}_i[\xi_i(y) + \xi_{i+1}(y)]\}^2 | \mathcal{F}_{i-1}) \le \mathbb{E}\{[\xi_i(y) + \xi_{i+1}(y)]^2 | \mathcal{F}_{i-1}\} \le c_1^2 \mathbb{E}(\zeta_i^2 | \mathcal{F}_{i-1}).$$

Thus, by the Cauchy-Schwarz inequality,

$$\sum_{i=1}^{n} \mathbb{E}(\{\mathcal{P}_i[\xi_i(y) + \xi_{i+1}(y)]\}^2 | \mathcal{F}_{i-1}) \le c_1^2 \sum_{i=1}^{n} \mathbb{E}(\zeta_i^2 | \mathcal{F}_{i-1}) \le 4 c_1^2 A_2, \tag{4.51}$$

uniformly in $y \in \mathbb{R}$. By (4.50) and (4.51), on the event $\{A_1 \le n^{1/4} \log n, A_2 \le n \log n\}$, the martingale differences $\mathcal{P}_i[\xi_i(y) + \xi_{i+1}(y)]$ are upper bounded by $c_1 n^{1/4} \log n$ and the sum of conditional variances is upper bounded by $4 c_1^2 n \log n$. Therefore, as in (4.34), for any $j = -N, \ldots, N$ and $c > 0$,

$$\begin{aligned} p_j &:= \mathbb{P}\Big\{ |S_n(y_j)| \ge c\sqrt{n} \log n, A_1 \le n^{1/4} \log n, A_2 \le n \log n \Big\} \\ &\le 2 \exp\left\{ - \frac{c^2 n (\log n)^2}{2[(c_1 n^{1/4} \log n)(c\sqrt{n} \log n) + 4 c_1^2 n \log n]} \right\} \\ &= 2 \exp(-\lambda \log n), \quad \lambda = \frac{c^2}{2[c c_1 n^{-1/4} (\log n)^2 + 4 c_1^2]}. \end{aligned} \tag{4.52}$$

For large enough c and n, we have $\lambda > 3$. Thus,

$$\begin{aligned} & \mathbb{P}\Big\{ \max_{-N\le j\le N} |S_n(y_j)| \ge c\sqrt{n}\log n\Big\} \\ \le\ & \mathbb{P}\Big\{ \max_{-N\le j\le N} |S_n(y_j)| \ge c\sqrt{n}\log n, A_1 \le n^{1/4}\log n, A_2 \le n\log n\Big\} \\ & +\mathbb{P}\{A_1 > n^{1/4}\log n\} + \mathbb{P}\{A_2 > n\log n\} \\ \le\ & \sum_{j=-N}^{N} p_j + \mathbb{P}\{A_1 > n^{1/4}\log n\} + \mathbb{P}\{A_2 > n\log n\}. \end{aligned} \tag{4.53}$$

By (4.52) and $N = n^2$, $\sum_{j=-N}^{N} p_j = O(1/n) \to 0$. Note that

$$\begin{aligned} \mathbb{E}(A_1^4) &\le \sum_{i=1}^{n} \mathbb{E}\{[\mathbb{E}(\zeta_i|\mathcal{F}_i) + \mathbb{E}(\zeta_i|\mathcal{F}_{i-1})]^4\} \\ &\le 16\sum_{i=1}^{n} \mathbb{E}\{[\mathbb{E}(\zeta_i|\mathcal{F}_i)]^4 + [\mathbb{E}(\zeta_i|\mathcal{F}_{i-1})]^4\} \ \le\ 32n\mathbb{E}(\zeta_1^4). \end{aligned}$$

Here the second "$\le$" follows from $(u+v)^4 \le 16(u^4+v^4)$ and the third "$\le$" follows from property (C3) in Section 4.4.1. Thus, by Markov's inequality, $\mathbb{P}\{A_1 > n^{1/4}\log n\} \le \mathbb{E}(A_1^4)/[n^{1/4}(\log n)]^4 = O[(\log n)^{-4}]$. Another application of Markov's inequality gives $\mathbb{P}\{A_2 > n\log n\} \le \mathbb{E}(A_2)/(n\log n) = O[(\log n)^{-1}]$. Thus, the right hand side of (4.53) goes to zero, and we conclude $\max_{-N\le j\le N} |S_n(y_j)| = O_p(\sqrt{n}\log n)$, completing the proof.

♢

Proof of Theorem 10. (i) By the same argument in Lemma 4, $\sup_{y\in\mathcal{Y}} |\hat{f}_Y(y) - f_Y(y)| = O_p[l_n^2 + (nl_n/\log n)^{-1/2}] = o_p[(\log n)^{-1/2}]$.

(ii) Recall the definition of $\mathcal{Y}_\epsilon$ in Condition 2. Write $\Delta_1 = \sup_{y\in\mathcal{Y}_{\epsilon/2}} |\hat{\mathcal{G}}_g(y) - \mathcal{G}_g(y)|$. Clearly, the uniform bound in Lemma 8 also holds on $\mathcal{Y}_{\epsilon/2}$, i.e., $\Delta_1 = O_p[b_n^2 + (b_n\sqrt{n})^{-1}\log n]$. Define

$$\bar{\sigma}_g^2(y) = \frac{\sum_{i=1}^{n}[g(Y_i) - \mathcal{G}_g(Y_{i-1})]^2 K_{h_n}(y - Y_{i-1})}{\sum_{i=1}^{n} K_{h_n}(y - Y_{i-1})}.$$

By the bounded support of $K(\cdot)$, it suffices to consider Y_{i-1} in a neighborhood of $y \in \mathcal{Y}$ or consider Y_{i-1} in the neighborhood $\mathcal{Y}_{\epsilon/2}$ of $\mathcal{Y}$ so that $\max_{1\le i\le n} |\hat{\mathcal{G}}_g(Y_{i-1}) -$

$\mathcal{G}_g(Y_{i-1})| \leq \Delta_1$. Applying the inequality $|a^2 - b^2| \leq x(x + 2|b|)$ for all $|a - b| \leq x$, we have

$$|\hat{\sigma}_g^2(y) - \bar{\sigma}_g^2(y)| \leq \Delta_1 \frac{\sum_{i=1}^n [\Delta_1 + 2|g(Y_i) - \mathcal{G}_g(Y_{i-1})|]|K_{h_n}(y - Y_{i-1})|}{|\sum_{i=1}^n K_{h_n}(y - Y_{i-1})|}. \quad (4.54)$$

Let $\Delta_2 = \max_{1 \leq i \leq n} |g(Y_i) - \mathcal{G}_g(Y_{i-1})|$. Then $\mathbb{E}(\Delta_2^4) \leq \sum_{i=1}^n \mathbb{E}[|g(Y_i) - \mathcal{G}_g(Y_{i-1})|^4] = O(n)$, which implies $\Delta_2 = O_p(n^{1/4})$. Thus, by (4.54) and Lemma 4, $\hat{\sigma}_g^2(y) - \bar{\sigma}_g^2(y) = O_p(\Delta_1^2 + \Delta_1\Delta_2) = o_p[(\log n)^{-1/2}]$, uniformly in $y \in \mathcal{Y}$. Finally, by the same argument in Lemma 8 (i.e., use a similar decomposition as in (4.26) along with Lemma 3, Lemma 4 and the argument in Lemma 8), we can show that $\sup_{y \in \mathcal{Y}} |\bar{\sigma}_g^2(y) - \sigma_g^2(y)| = O_p[h_n^2 + (h_n\sqrt{n})^{-1} \log n] = o_p[(\log n)^{-1/2}]$. This completes the proof. ♢

4.4.5 Proof of Theorem 11

Proof of Theorem 11. Using (4.6) and the independence between $\varepsilon \sim q_\varepsilon(\cdot)$ and (X_{i-1}, X_i),

$$\begin{aligned}
\mathcal{G}_g(y) &= \int_{\mathbb{R}} g(y')q_Y(y'|y)dy' \\
&= \frac{1}{\mathbb{E}[q_\varepsilon(y - X_{i-1})]} \mathbb{E}\Big[q_\varepsilon(y - X_{i-1}) \int_{\mathbb{R}} g(y')q_\varepsilon(y' - X_i)dy'\Big] \\
&= \frac{1}{\mathbb{E}[q_\varepsilon(y - X_{i-1})]} \mathbb{E}\Big[q_\varepsilon(y - X_{i-1}) \int_{\mathbb{R}} g(X_i + z)q_\varepsilon(z)dz\Big] \\
&= \frac{\mathbb{E}\{q_\varepsilon(y - X_{i-1})\mathbb{E}[g(X_i + \varepsilon)|X_{i-1}, X_i]\}}{\mathbb{E}[q_\varepsilon(y - X_{i-1})]} \\
&= \frac{\mathbb{E}[g(X_i + \varepsilon)q_\varepsilon(y - X_{i-1})]}{\mathbb{E}[q_\varepsilon(y - X_{i-1})]}.
\end{aligned}$$

Here the third equality follows by taking the transformation $z = y' - X_i$. ♢

Chapter 5

Quantile regression for locally stationary process

5.1 Introduction

Quantile regression (QR) has recently received considerable attention in time series analysis to address important problems in many disciplines such as finance, economics and environment. Notable changes in climate variability can make a devastating impact on the environment, and thus the climate extremes are more informative than the mean. For example, in modeling the sea level along with the global temperature, it is practically important to study extreme quantiles close to 0 or 1 and address related inferences. Especially, researchers would like to investigate how much the global temperature affects the sea level when the sea level is unusually high. Recently, there is a substantial literature on quantile regression under various time series settings: quantile autoregression model (Koenker and Xiao, 2006); varying coefficient quantile regression model (Cai and Xu, 2008); quantile regression dynamic panel model with fixed effects (Galvao, 2011); partially varying coefficient quantile regression model using a semiparametric method (Cai and Xiao, 2012).

Another currently attracting research area is non-stationary time series analysis. In the classical time series analysis including the above literature on quantile regression, stationarity plays an indispensable role in addressing model estimation

and inference. While being reasonable in some applications, the stationary assumption is often too strict that it is hardly fulfilled in many applications. It is easy to see that the underlying model structure of the sea level series such as the mean function and variance function has changed over time. Especially, in this case the time-varying non-stationary pattern is the main objective of the study. In response to these up-to-data needs in time series analysis, we develop statistical inference for quantile regression of a special class of non-stationary time series model in this paper.

The time series model we want to study in this article is the autoregressive (AR) model with exogenous inputs. The model has been an active research area due to the fact that many of the AR processes are affected by factors outside of the process. It is well known that the global temperature series have a very strong impact on the sea level series so that it is necessary to include the global temperature measurement as an exogenous input in the model. To involve outside factors and time-varying structure, we consider the time-varying coefficient AR(p) model $\{X_i\}$ with time series exogenous inputs $\{Z_i\}$ of order q denoted by TV-ARX(p,q):

$$X_i = \alpha(i/n) + \sum_{j=1}^{p} \phi_j(i/n) X_{i-j} + \sum_{r=0}^{q} \beta_r(i/n) Z_{i-r} + \sigma(i/n)\varepsilon_i, \quad i = 1, \ldots, n, \quad (5.1)$$

where $\alpha(t), \phi_1(t), \ldots, \phi_p(t), \beta_0(t), \ldots, \beta_q(t), \sigma(t)$ are unknown functions in $t \in [0, 1]$. By contrast with stationary ARX models, variant coefficients over time enable (5.1) to model time-varying non-stationary pattern. Recently, Xioaye and Zhao (2013) studied nonparametric estimation and hypothesis testing for the coefficient functions.

Reflected by the above discussions, our primary goal of this paper is to establish asymptotic properties and hypothesis testing of time-varying QR estimation of the non-stationary TV-ARX model. Since the time-varying quantile coefficients are allowed to vary over quantiles $\tau \in [0, 1]$, the model allows one to study how lagged observations of the main series $\{X_i\}$ and exogenous inputs $\{Z_i\}$ make an influence on the response at a specific quantile over time. Furthermore, the model enjoys the flexibility of nonparametric modeling and the interpretability of linear models. Recently, there is a theoretical literature on quantile estimation for non-

stationary time series. Draghicescu et al. (2009) and Zhou and Wu (2010) studied estimation of time-varying quantile curves for non-stationary processes, and Zhou (2010) proposed specification testing of time-varying quantile curves. However, to the best of our knowledge, quantile regression for non-stationary time series has not been studied in either statistics or econometrics literature.

For general non-stationary process, it is not trivial to derive asymptotic properties because the probabilistic structure of a non-stationary process at present may not provide any information for future observations of the process. To alleviate the difficulty from non-stationarity, we use an idea of locally stationary process by imposing suitable smoothness conditions on parameters. The conditions guarantee that the underlying model dynamics is nearly sustained within a small time window, and hence it is locally approximated by a stationary process. Since Dahlhaus (1997) introduced this idea, there has been growing interest in studying locally stationary models: piecewise stationary AR models (Davis et al., 2006); locally stationary ARCH models (Dahlhaus and Subba Rao, 2006); localy stationary wavelet processes (Van Bellegem and Von Sachs, 2008). I also refer to Dahlhaus (2012) and reference therein for detailed discussion.

5.2 Methodology and future works

Define the vectors

$$\begin{aligned} U_i &= (1, X_{i-1}, \ldots, X_{i-p}, Z_i, \ldots, Z_{i-q})^T, \\ \theta(\cdot) &= (\alpha(\cdot), \phi_1(\cdot), \ldots, \phi_p(\cdot), \beta_1(\cdot), \ldots, \beta_q(\cdot))^T, \\ \theta(\cdot|\tau) &= (\alpha(\cdot|\tau), \phi_1(\cdot|\tau), \ldots, \phi_p(\cdot|\tau), \beta_1(\cdot|\tau), \ldots, \beta_q(\cdot|\tau))^T, \end{aligned}$$

where $\alpha(\cdot|\tau),\ \phi_1(\cdot|\tau), \ldots, \phi_p(\cdot|\tau),\ \beta_0(\cdot|\tau), \ldots, \beta_q(\cdot|\tau)$ are the τ-th quantile time-varying parameters of interest.

Given (5.1), denote by $Q_{X_i}(\tau|U_i)$ the conditional τ-th quantile of X_i given U_i. We consider τ-th quantile regression for locally stationary TV-ARX:

$$Q_{X_i}(\tau|U_i) = \alpha(i/n|\tau) + \sum_{j=1}^{p} \phi_j(i/n|\tau) X_{i-j} + \sum_{r=0}^{q} \beta_r(i/n|\tau) Z_{i+1-r}, \quad i = 1, \ldots, n,$$

Write $K_i(t) = K\{(i/n - t)/b_n\}$. To estimate the quantile regression curves, we adopt the local linear quantile regression

$$\Big\{\hat{\theta}(t|\tau), \hat{\theta}'(t|\tau)\Big\} = \underset{\theta,\theta^*}{\operatorname{argmin}} \sum_{i=1}^{n} \rho_\tau\Big\{X_i - U_i^T\theta - (i/n - t)U_i^T\theta^*\Big\}K_i(t),$$

where $\rho_\tau(z) = z(\tau - \mathbf{1}_{z\leq 0})$ is the quantile loss function at a quantile $\tau \in (0,1)$, $K(\cdot)$ is a kernel function, and $b_n > 0$ is a bandwidth. Because the probabilistic structure of X_i changes over time, it is infeasible to estimate $\hat{\theta}(t|\tau), \hat{\theta}'(t|\tau)$ with non-stationary process X_i without some smoothness conditions.

By the idea of locally stationarity (Dahlhaus, 2012), for the neighborhood of a fixed point $t \in [0,1]$, the process $\{X_i\}$ can be approximated by a stationary process $\{X_i(t)\}$ which is defined by

$$X_i(t) = \alpha(t) + \sum_{j=1}^{p} \phi_j(t)X_{i-j}(t) + \sum_{r=0}^{q} \beta_r(t)Z_{i+1-r}(t) + \sigma(t)\varepsilon_i, \quad i = 1, \ldots, n, \quad (5.2)$$

where $\{Z_i(t)\}$ is also a stationary process which approximates a locally stationary process $\{Z_i\}$. Furthermore, as shown in Xiaoye and Zhao (2013), under suitable smoothness conditions on parameters we have the following degree of approximation:

$$\begin{aligned}
X_i &= X_i(t) + O_p(|i/n - t| + 1/n) \\
X_i &= X_i(t) + (i/n - t)X_i'(t) + O_p(|i/n - t|^2 + 1/n), \\
Z_i &= Z_i(t) + O_p(|i/n - t| + 1/n) \\
Z_i &= Z_i(t) + (i/n - t)Z_i'(t) + O_p(|i/n - t|^2 + 1/n).
\end{aligned}$$

As $n \to \infty$ and $i/n \approx t$, $X_i \approx X_i(t)$ and $Z_i \approx Z_i(t)$. Thus, in the derivation of asymptotic properties of this quantile regression estimator we adopt the stationary process $U_i(t)$ as a local approximation of U_i:

$$U_i(t) = (1, X_{i-1}(t), \ldots, X_{i-p}(t), Z_i(t), \ldots, Z_{i-q}(t)). \quad (5.3)$$

Based on the above discussions, for given $t \in [0,1]$, the asymptotic normality of $\hat{\theta}(t|\tau)$ is derived in Theorem 12.

Theorem 12. *Let* $\mathcal{I}(\tau) = \tau(1-\tau)/f_\varepsilon^2(Q_\varepsilon(\tau))$. *As* $nb_n^2 \to \infty$ *and* $nb_n^9 \to 0$, *under suitable regularity conditions*

$$V_n(t) := \sqrt{nb_n}\Big\{\hat{\theta}(t|\tau) - \theta(t|\tau) - \frac{\theta''(t|\tau)\mu_K}{2}b_n^2 + o(b_n^2)\Big\} \Rightarrow N\Big(0, \sigma^2(t)\varphi_K\Gamma_U^{-1}(t)\mathcal{I}(\tau)\Big),$$

where $\Gamma_U(t) = E[U_1(t)U_1(t)^T]$, $\mu_K = \int_{\mathbb{R}} u^2K(u)du$, *and* $\varphi_K = \int_R K^2(u)du$. *In addition,* $V_n(t)$ *and* $V_n(t')$ *are asymptotically independent for* $t \neq t'$.

In addition, I would like to develop hypothesis testing for a fixed quantile $\tau \in (0,1)$ whether the quantile curves are truly time varying

$$H_0 : \beta_r(t|\tau) = c_r \quad v.s \quad H_a : \beta_r(t|\tau) \neq c_r \quad \text{for} \quad \forall r \in \{1, 2, \ldots, q\},$$

whether some variables are significant

$$H_0 : \phi_j(t|\tau) = 0 \quad v.s \quad H_a : \phi_j(t|\tau) \neq 0,$$

and for a fixed time t whether the quantile curves are varying with quantiles

$$H_0 : \beta_r(t|\tau) = c_r \quad v.s \quad H_a : \beta_r(t|\tau) \neq c_r \quad \text{for} \quad \forall r \in \{1, 2, \ldots, q\} \text{ and } \tau \in [0,1].$$

If the quantile curves are not varying with quantiles, then we can apply the idea of combining multiple quantile regressions in Chapter 4 to obtain efficient estimation for the time varying coefficients.

I also plan to extend this idea of locally stationary process to longitudinal data.

$$X_{ij} = \alpha(j/n) + \sum_{k=1}^{p}\phi_k(j/n)X_{ij-k} + Z_{ij}^T\beta(j/n) + \sigma(j/n)\varepsilon_{ij}, \quad i = 1, \ldots, m, j = 1, \ldots, n$$

where $\alpha(\cdot)$ is a time-varying trend intercept, $\beta(\cdot) = (\beta_1(\cdot), \ldots, \beta_d(\cdot))^T$ is a d-dimensional vector of coefficient functions, $\phi_1(\cdot), \ldots, \phi_p(\cdot)$ are AR coefficient functions, $\sigma(\cdot)$ is an unknown deterministic time-dependent variance, and $\{\varepsilon_{ij}\}$ are i.i.d. noises. The main goals are to develop robust and efficient estimations for $\phi_k(\cdot)$ and $\beta(\cdot)$ by combining multiple quantile regressions and also to address the hypothesis testing problem for $H_0 : \beta_r(t) = c$, where c is a constant.

Bibliography

[1] Aït-Sahalia, Y. (1996) Testing continuous-time models of the spot interest rate. *Review of Financial Studies*, **9** 385–426.

[2] Aït-Sahalia, Y., Fan, J. and Peng, H. (2009) Nonparametric transition-based tests for jump-diffusions. *Journal of the American Statistical Association*, **104** 1102–1116.

[3] 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** 351–416.

[4] Azzalini, A. and Bowman, A. (1993) On the use of nonparametric regression for checking linear relationships. *Journal of the Royal Statistical Society: Series B*, **55** 549–557.

[5] Bickel, P.J. and Rosenblatt, M. (1973) On some global measures of the deviations of density function estimates. *Annals of Statistics*, **1** 1071–1095.

[6] Bradic, J., Fan, J. and Wang, W. (2011) Penalized composite quasi-likelihood for ultra-high dimensional variable selection. *Journal of the Royal Statistical Society, Series B*, **73** 325–349.

[7] Cai Z. and Xiao, Z (2012) Semiparametric quantile regression estimation in dynamic models with partially varying coefficients. *Journal of Econometrics*, **167** 413–425.

[8] Cai, Z. and Xu, X. (2008) Nonparametric quantile estimations for dynamic smooth coefficient models. *Journal of the American Statistical Association*, **103** 1595–1608.

[9] Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC Press.

[10] Chen, K. and Jin. Z.H. (2005) Local polynomial regression analysis of cluster data. *Biometrika*, **92** 59–74.

[11] Chen, X., Hong, H. and Nekipelov, D. (2011) Nonlinear models of measurement errors. *Journal of Economic Literature*, **49** 901–937.

[12] Dahlhaus, R. (1997) Fitting time series models to nonstationary processes. *Annals of Statistics*, **25** 1–37.

[13] Dahlhaus, R. (2012) Locally stationary processes. In *Handbook of Statistics, Time Series Analysis: Methods and Applications*, edited by Subba Rao, T., Subba Rao, S. and Rao, C.R., **30** 351–413.

[14] Dahlhaus, R. and Subba Rao, S. (2006) Statistical inference for time-varying ARCH processes. *Annals of Statistics*, **34** 1075–1114.

[15] Davis, R.A., Lee, T. and Rodriguez-Yam, G. (2006) Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101** 223–239.

[16] De la Peña, V.H., Lai, T.L. and Shao, Q.M. (2009). *Self-Normalized Processes.* New York: Springer.

[17] Dockery, D.W., Berkey, C.S., Ware, J.H., Speizer, F.E, and Ferris Jr, B.G. (1983) Distribution of forced vital capacity and forced expiratory volume in one second in children 6 to 11 years of age. *The American Review of Respiratory Disease*, **128** 405–412.

[18] Draghicescu, D., Guillas, S. and Wu, W.B. (2009). Quantile curve estimation and visualization for non-stationary time series. *Journal of Computational and Graphical Statistics*, **18** 1–20.

[19] Eubank, R.L. and Speckman, P.L. (1993) Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, **88** 1287–1301.

[20] Fan, J. (1991) On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, **19** 1257–1272

[21] Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

[22] Fan, J. and Yao, Q. (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods.* Springer, New York.

[23] Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, **29** 153–193.

[24] Fan, J. and Zhang, J.T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B*, **62** 303–322.

[25] Fan, J. and Zhang, W. (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics*, **27** 715–731.

[26] Fan, J. and Zhang, W. (2008) Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1** 179–195.

[27] Fan, Y. and Li, Q. (1996) Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica*, **64** 865–890.

[28] Fitzmaurice, G.M., Laird, N.M., and Ware, J.M. (2004) *Applied Longitudinal Analysis*. Wiley, Hoboken, NJ.

[29] Freedman, D.A. (1975) On tail probabilities for martingales. *Annals of Probability*, **3** 100–118.

[30] Fuller, W. (1987) *Measurement Error Models*. New York: John Wiley & Sons.

[31] Galvao, A.F. (2011) Quantile regression for dynamic panel data with fixed effects. *Journal of Econometrics*, **164** 142–157.

[32] Gao, J. and King, M. (2004) Adaptive testing in continuous-time diffusion models. *Econometric Theory*, **20** 844–882.

[33] Gould, S.J. (1966) Allometry and size in ontogeny and phylogeny. *Biological Reviews of the Cambridge Philosophical Society*, **41** 587–640.

[34] Grama, I.G. and Haeusler, E. (2006) An asymptotic expansion for probabilities of moderate deviations for multivariate martingales. *Journal of Theoretical Probability*, **19** 1–44.

[35] Hall, P., Müller, H.G. and Wang, J.L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Annals of Statistics*, **34** 1493–1517.

[36] Härdle, W. and Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, **21** 1926–1947.

[37] He, X., Fu, B. and Fung, W.K. (2003) Median regression for longitudinal data. *Statistics in Medicine*, **22** 3655–3669.

[38] He, X., Zhu, Z.Y. and Fung, W.K. (2002) Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, **89** 579–590.

[39] Honda, T. (2004) Quantile regression in varying coefficient models. *Journal of Statistical Planning and Inference*, **121** 113–125.

[40] Hong, Y. (1999) Hypothesis testing in time series via the empirical characteristic function: a generalized spectral density approach. *Journal of the American Statistical Association*, **94** 1201–1220.

[41] Hong, Y. and Li, H. (2005) Nonparametric specification testing for continuous-time models with applications to term structure of interest rates. *Review of Financial Studies*, **18** 37–84.

[42] Hong, Y. and White, H. (1995) Consistent specification testing via nonparametric series regression. *Econometrica*, **63** 1133–1159.

[43] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.P. (1998) Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85** 809–822.

[44] Kai, B., Li, R. and Zou, H. (2010) Local composite quantile regression smoothing: an efficient and safe alternative to local polynomial regression. *Journal of the Royal Statistical Society: Series B*, **72** 49–69.

[45] Kiefer, N.M. and Vogelsang, T.J. (2005). A new asymptotic theory for heteroskedasticity autocorrelation robust tests. *Econometric Theory* **21**, 1130–1164.

[46] Kim, M.O. (2007) Quantile regression with varying coefficients. *Annals of Statistics*, **35** 92–108.

[47] Knafl, G., Sacks, J. and Ylvisaker, D. (1985) Confidence bands for regression functions. *Journal of the American Statistical Association*, **80** 683–691.

[48] Koenker, R. (1984) A note on L-estimators for linear models. *Statistics and Probability Letters*, **2** 323–325.

[49] Koenker, R. (2004) Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, **91** 74–89.

[50] Koenker, R. (2005) *Quantile Regression*. Cambridge University Press, New York.

[51] Koenker, R. and Xiao, Z. (2006) Quantile autoregression. *Journal of the American Statistical Association*, **101** 980–990.

[52] Li, Q. and Racine, J. (2007) *Nonparametric Econometrics*. Princeton University Press, Princeton, New Jersey.

[53] Li, R. and Li, Y. (2009) Local linear regression for data with AR errors. *Acta Mathematicae Applicatae Sinica*, **25** 427–444.

[54] Li, X. and Zhao, Z. (2013) Inference for time-varying exogenous autoregressive Models. *Manuscript*.

[55] Li,Y. and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, **38** 3321–3351.

[56] Lobato, I.N. (2001). Testing that a dependent process is uncorrelated. *Journal of the American Statistical Association*, **96** 1066–1076.

[57] Ma, S., Yang, L. and Carroll, R.J. (2012). A simultaneous confidence band for sparse longitudinal regression. *Statistical Sinica*, **22** 95–122.

[58] Müller, H.G. (2005). Functional modeling and classification of longitudinal data. *Scandinavian Journal of Statistics*, **32** 223–240.

[59] Pinkse, J. (1998) A consistent nonparametric test for serial independence. *Journal of Econometrics*, **84** 205–231.

[60] Portnoy, S. and Koenker, R. (1989) Adaptive L-estimation of linear models. *Annals of Statistics*, **17** 362–381.

[61] Ramsay, J.O. and Silverman, B.W. (2005). *Functional Data Analysis*. New York: Springer.

[62] Rice, J.A. and Silverman, B.W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, **53** 233–243.

[63] Ruppert, D., Sheather, S.J. and Wand, M.P. (1995) An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, **90** 1257–1270.

[64] Shao, X. (2010). A self-normalized approach to confidence interval construction in time series. *Journal of the Royal Statistical Society, Series B*, **72** 343–366.

[65] Silverman, B.W. (1986) *Density Estimation*. Chapman and Hall, London.

[66] Stam, A.J. (1959) Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, **2** 101–112.

[67] Van Bellegem, S. and Von Sachs, R. (2008) Locally adaptive estimation of evolutionary wavelet spectra. *Annals of Statistics*, **36** 1879–1924.

[68] Wang, H.J. and Fygenson, M. (2009) Inference for censored quantile regression models in longitudinal studies. *Annals of Statistics*, **37** 756–781.

[69] Wang, H.J., Zhu, Z. and Zhou, J. (2009) Quantile regression in partially linear varying coefficient models. *Annals of Statistics*, **37** 3841–3866.

[70] Wang, X., Dockery, D.W., Wypij, D., Fay, M.E. and Ferris Jr, B.G. (1993) Pulmonary function between 6 and 18 years of age. *Pediatric Pulmonology*, **15** 75–88.

[71] Wei, Y. and He, X. (2006) Conditional growth charts. *Annals of Statistics*, **34** 2069–2097.

[72] Wu, C.O., Chiang, C.T. and Hoover, D.R (1998) Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American Statistical Association*, **93** 1388–1402.

[73] Wu, C.O. and Yu, K.F. (2002) Nonparametric varying-coefficient models for the analysis of longitudinal data. *International Statistical Review*, **70** 373–393.

[74] Wu, H. and Zhang, J.T. (2002). Local polynomial mixed-effects for longitudinal data. *Journal of the American Statistical Association*, **97** 883–897.

[75] Wu, H. and Zhang, J.T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches.* New Jersey: Wiley.

[76] Wu, W.B. (2005) Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences USA*, **102** 14150–14154.

[77] Xiao, Z., Linton, O.B., Carroll, R.J. and Mammen, E. (2003) More efficient local polynomial estimation in nonparametric regression with autocorrelated errors. *Journal of the American Statistical Association*, **98** 980–992.

[78] Yao, F. (2007). Asymptotic distributions of nonparametric regression estimators for longitudinal or functional data. *Journal of Multivariate Analysis*, **98** 40–56.

[79] Yao, F., Müller, H.G., and Wang, J.L. (2005a). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, **33** 2873–2903.

[80] Yao, F., Müller, H.G., and Wang, J.L. (2005b). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, **100** 577–590.

[81] Yu, K. and Jones, M.C. (1998) Local linear quantile regression. *Journal of the American Statistical Association*, **93** 228–237.

[82] Zhang, J.T. and Chen, J. (2007). Statistical inferences for functional data. *Annals of Statistics*, **35** 1052–1079.

[83] Zhang, L., Mykland, P.A. and Aït-Sahalia, Y. (2005) A tale of two time scales: determining integrated volatility with noisy high-frequency data. *Journal of the American Statistical Association*, **472** 1394–1411.

[84] Zeger, S.L. and Diggle, P.J. (1994) Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50** 689–699.

[85] Zhao, Z. (2008) Parametric and nonparametric models and methods in financial econometrics. *Statistics Surveys*, **2** 1–42.

[86] Zhao, Z. (2011) Nonparametric model validations for hidden Markov models with applications in financial econometrics. *Journal of Econometrics*, **162** 225–239.

[87] Zhao, Z. and Wu, W.B. (2008) Confidence bands in nonparametric time series regression. *Annals of Statistics*, **36** 1854–1878.

[88] Zhou, Z. (2010). Nonparametric inference of quantile curves for nonstationary time series. *Annals of Statistics*, **38** 2187–2217.

[89] Zhou, Z. and Wu, W.B. (2009). Local linear quantile estimation for nonstationary time series. *Annals of Statistics*, **37** 2696-2729.

[90] Zhu, Z., Fung, W.K. and He, X. (2008) On the asymptotics of marginal regression splines with longitudinal data. *Biometrika*, **95** 907–917.

[91] Zou, H. and Yuan, M. (2008) Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, **36** 1108–1126.

Vita

Seonjin Kim

Education

Ph.D., Statistics, The Pennsylvania State University, University Park, PA.

B.S., Applied Mathematics, Korea Advanced Institute of Science and Technology (KAIST), Deajeon, Republic of Korea, February 2006.

Honors and Awards

Nonparametric Statistics Student Paper Award, American Statistical Association, JSM 2012.

Department Draft Scholarship, KAIST, 2003–2006.

Publications

Kim, S. and Zhao, Z. (2013) Unified inference for sparse and dense longitudinal models. *Biometrika*, **100**, 203-212.

Kim, S., Zhao, Z and Xiao, Z. (2013) Efficient estimation for time-varying coefficient longitudinal models. *Manuscript*.

Kim, S. and Zhao, Z. (2013) Specification test for Markov models with measurement errors. *Manuscript*.

Zhao, Z., Kim, S. and Shao, X. and (2013) Nonparametric functional central limit theorem for time series with application to self-normalized confidence interval. *Manuscript*.

Teaching Experience

STAT463: Applied Time Series Analysis, Spring 2012, Spring 2013.

STAT401: Experimental Methods, Summer 2012.

STAT418: Introduction to Probability and Stochastic Processes for Engineering, Fall 2012.

Consulting Experience

Graduate Student Consultant of Statistical Consulting Center, August 2010–May 2011.